Network Analysis

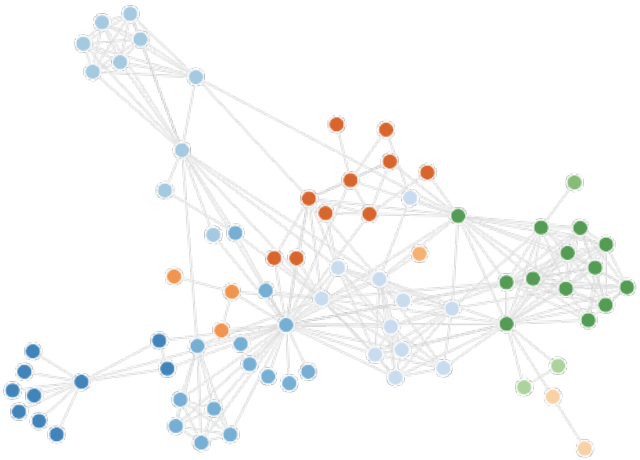
Maneesh Agrawala

CS 448B: Visualization
Winter 2020

1

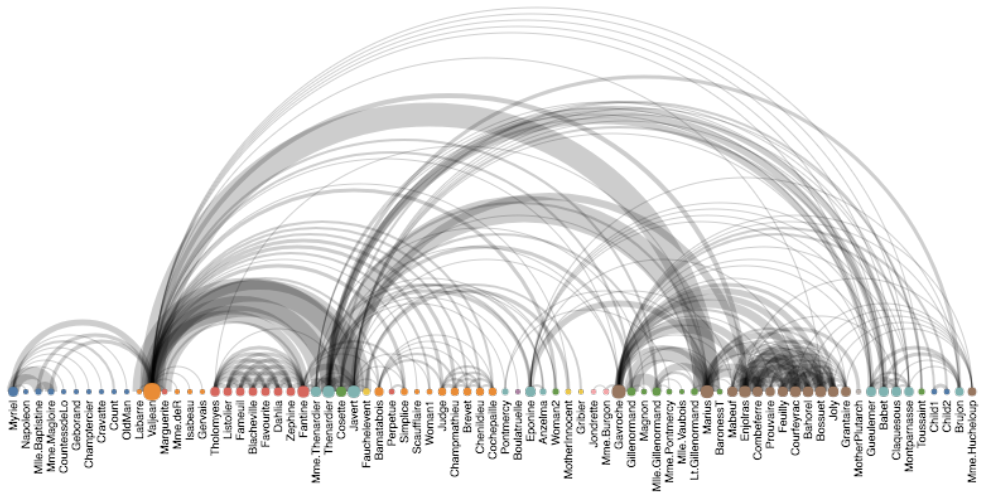
Last Time: Network Layout

2



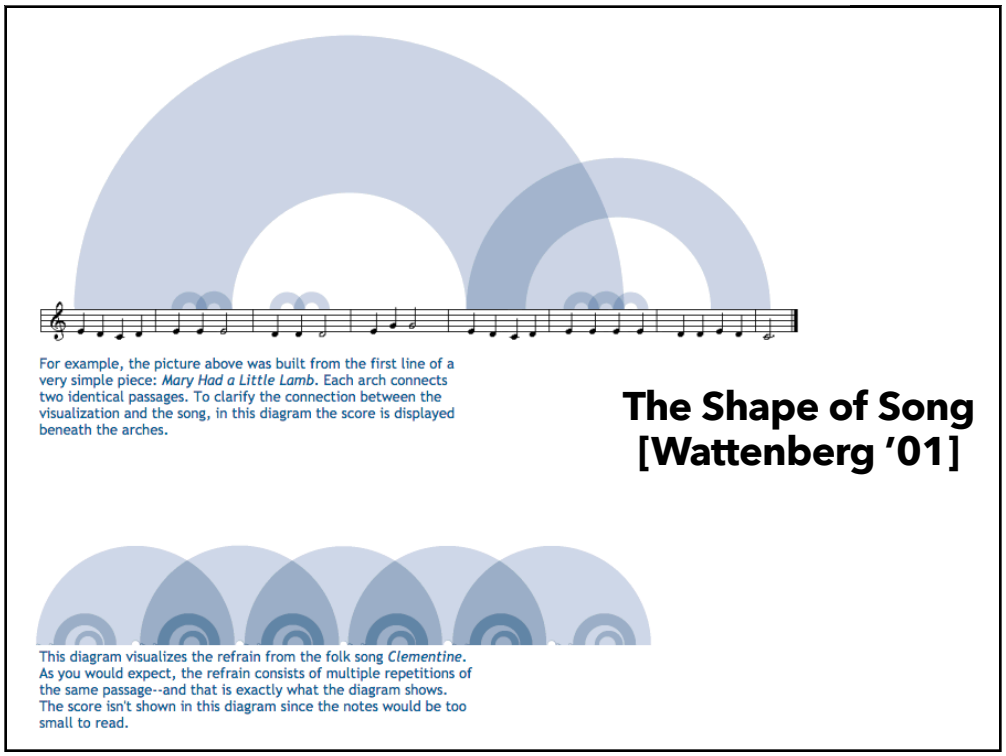
Interactive Example: Configurable Force Layout

3

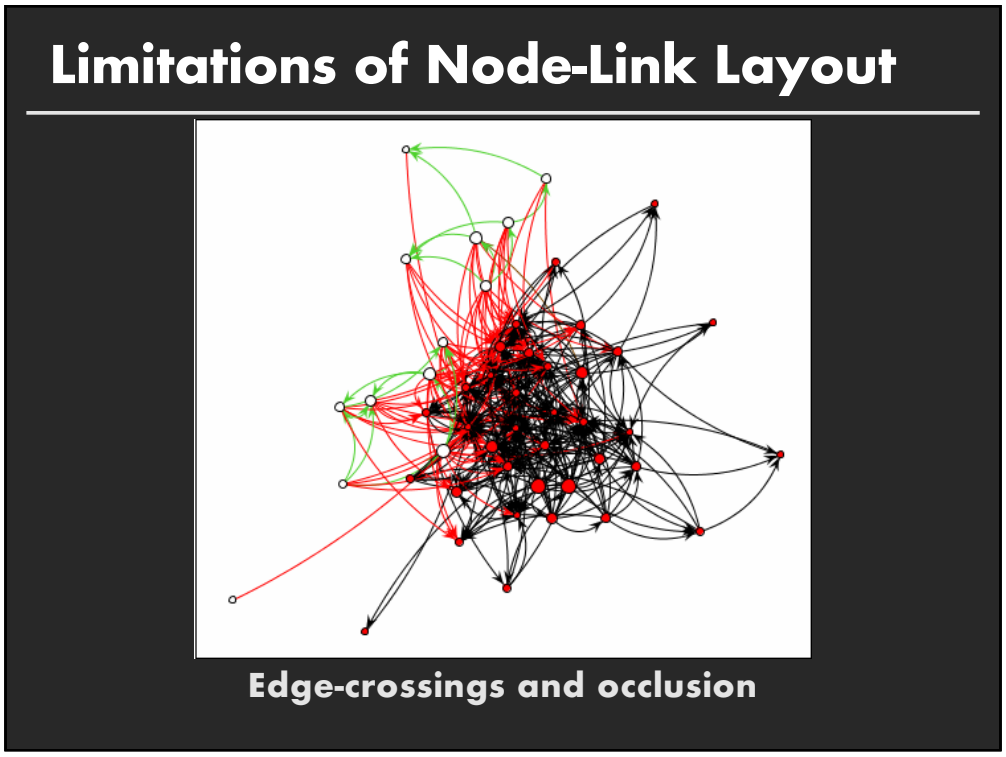


Linear node layout, circular arcs show connections. Layout quality sensitive to node ordering!

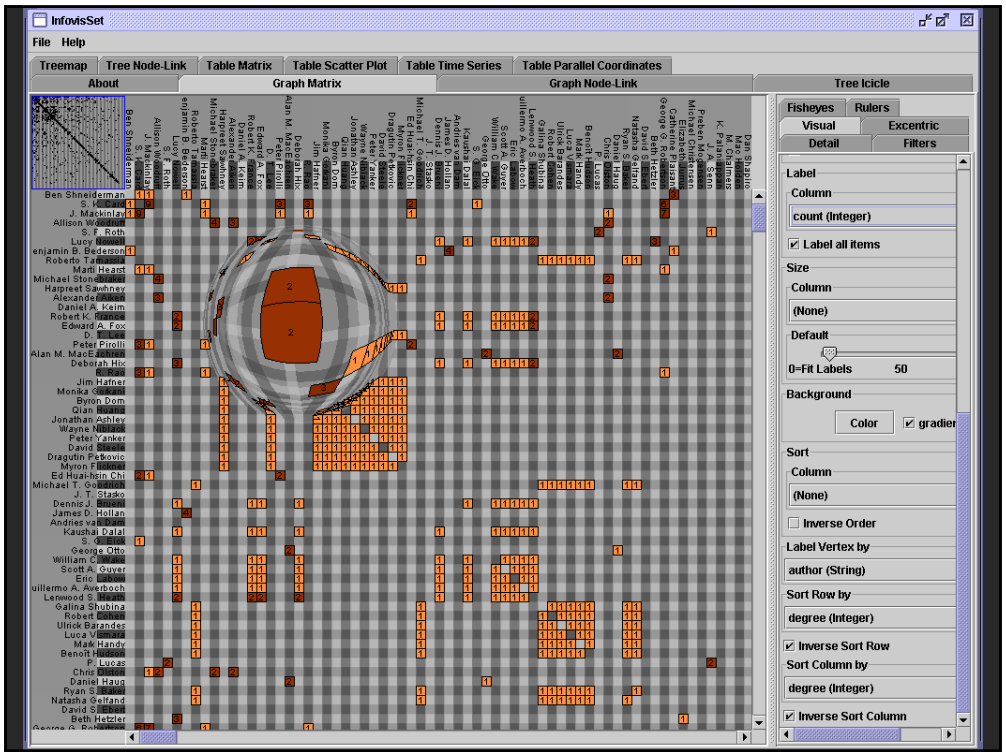
4


For example, the picture above was built from the first line of a very simple piece: *Mary Had a Little Lamb*. Each arch connects two identical passages. To clarify the connection between the visualization and the song, in this diagram the score is displayed beneath the arches.
The Shape of Song [Wattenberg '01]
This diagram visualizes the refrain from the folk song *Clementine*. As you would expect, the refrain consists of multiple repetitions of the same passage--and that is exactly what the diagram shows. The score isn't shown in this diagram since the notes would be too small to read.

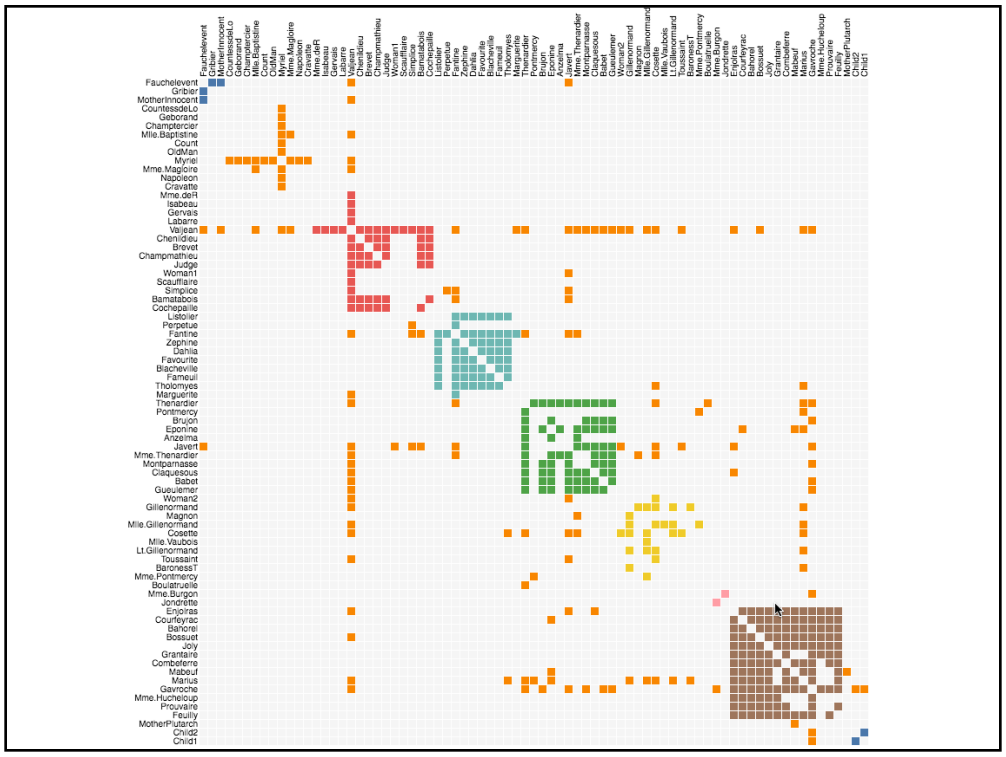
5


Limitations of Node-Link Layout
Edge-crossings and occlusion

6


InfovisSet
File Help
Treemap
Tree Node-Link
Table Matrix
Table Scatter Plot
Table Time Series
Table Parallel Coordinates
About
Graph Matrix
Graph Node-Link
Tree Icicle
Fisheyes
Rulers
Visual
Excentric
Detail
Filters
Label
Column
count (Integer)
Label all items
Size
Column
(None)
Default
0=Fit Labels
50
Background
Color
gradien
Sort
Column
(None)
Inverse Order
Label Vertex by
author (String)
Sort Row by
degree (Integer)
Inverse Sort Row
Sort Column by
degree (Integer)
Inverse Sort Column

8



9

Seriation/Ordination Permutation

Goal: Ensure similar items placed near each other.
E.g., minimize sum of distances of adjacent items.

Requires combinatorial optimization: NP-Hard!

Instead, approximate / heuristic approaches used:
Perform hierarchical clustering, sort cluster tree
Apply approximate traveling salesperson solver

Seriation initially used in archaeology for relative dating of artifacts based on observed properties

11

Attribute-Driven Layout

Large node-link diagrams get messy!
Is there additional structure we can exploit?

Idea: Use data attributes to perform layout

- e.g., scatter plot based on node values

Dynamic queries and/or brushing can be used to explore connectivity

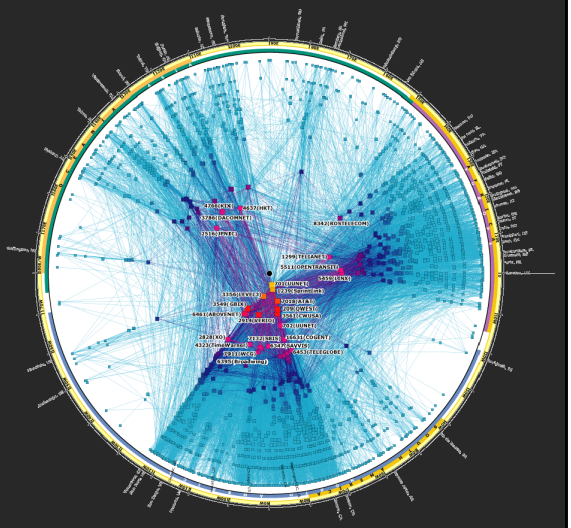
12

Attribute-Driven Layout

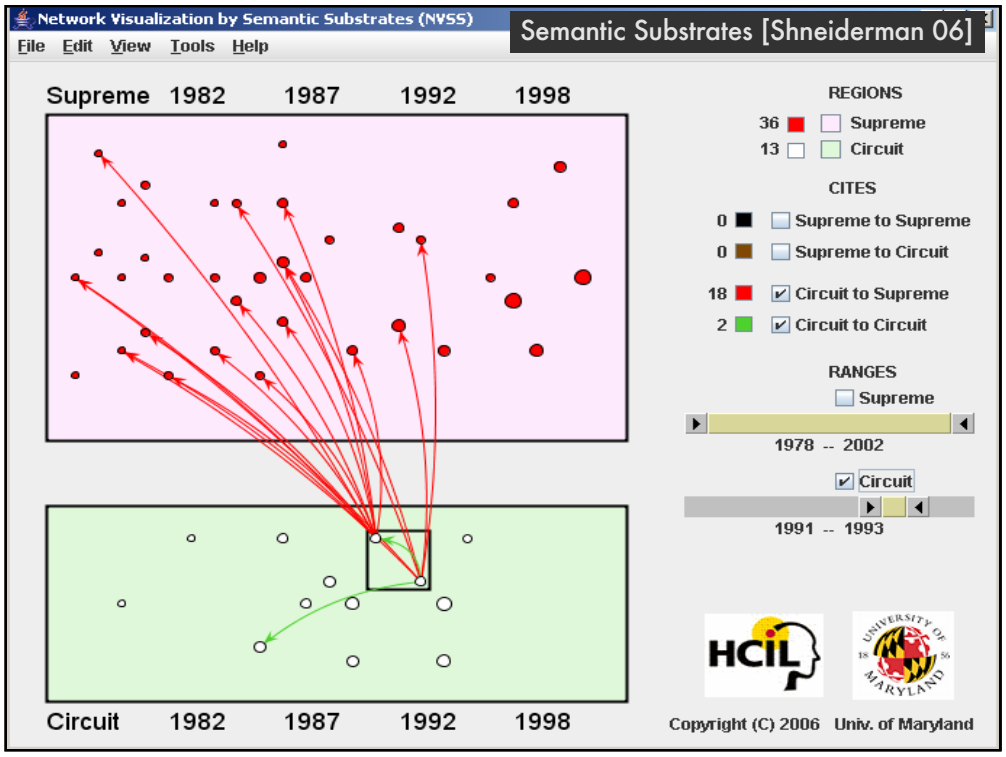
The "Skitter" Layout
- Internet Connectivity
- Radial Scatterplot

Angle = Longitude
- Geography

Radius = Degree
- # of connections
- (a statistic of the nodes)



13



14

Summary

Tree Layout

Indented / Node-Link / Enclosure / Layers

How to address issues of scale?

- Filtering and Focus + Context techniques

Graph Layout

Tree layout over spanning tree

Hierarchical "Sugiyama" Layout

Optimization (Force-Directed Layout)

Attribute-Driven Layout

15

Announcements

16

Final project

New visualization research or data analysis project

- **Research**: Pose problem, Implement creative solution
- **Data analysis**: Analyze dataset in depth & make a visual explainer

Deliverables

- **Research**: Implementation of solution
- **Data analysis/explainer**: Article with multiple interactive visualizations
- 6-8 page paper

Schedule

- Project proposal: Wed 2/19
- Design review and feedback: 3/9 and 3/11
- Final presentation: 3/16 (7-9pm) Location: TBD
- Final code and writeup: 3/18 11:59pm

Grading

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

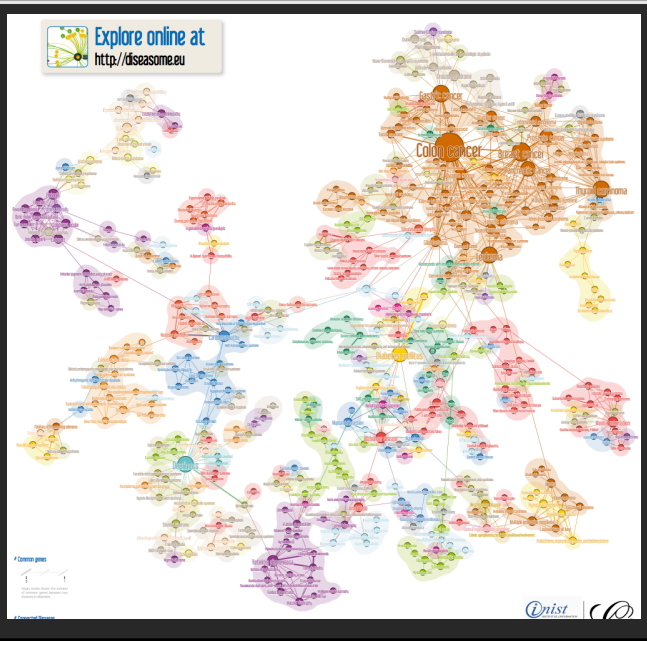
17

Network Analysis

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

18

Diseases

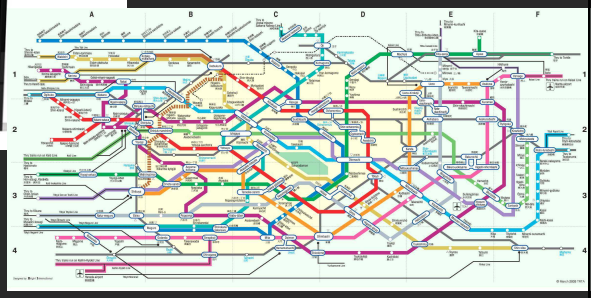


http://diseasome.eu/

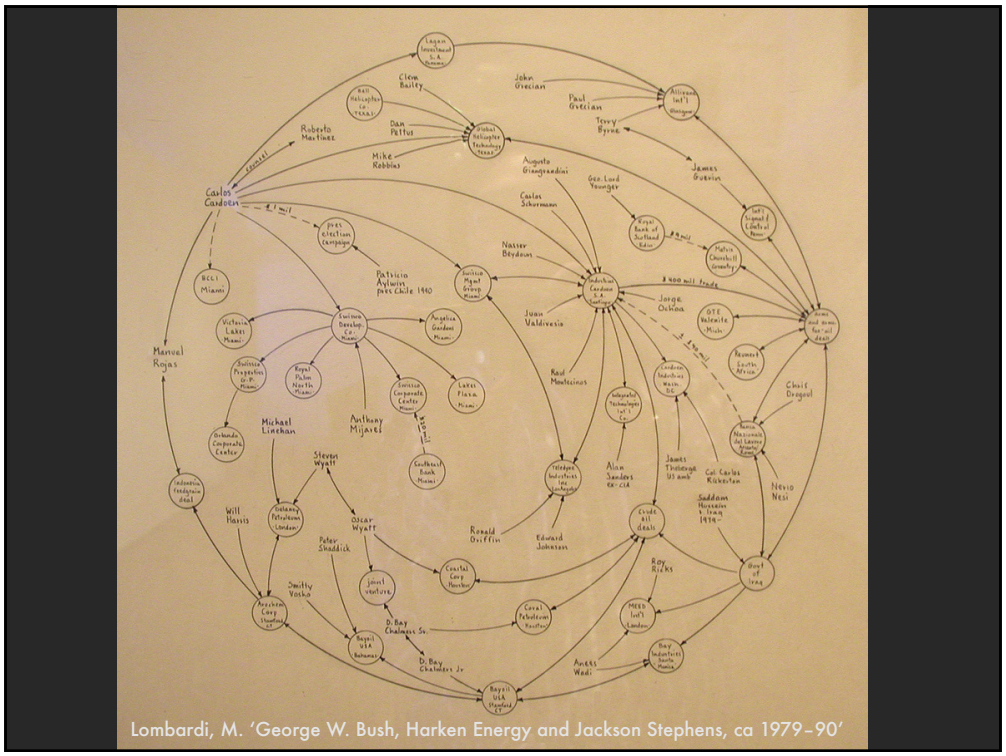
20

Transportation

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



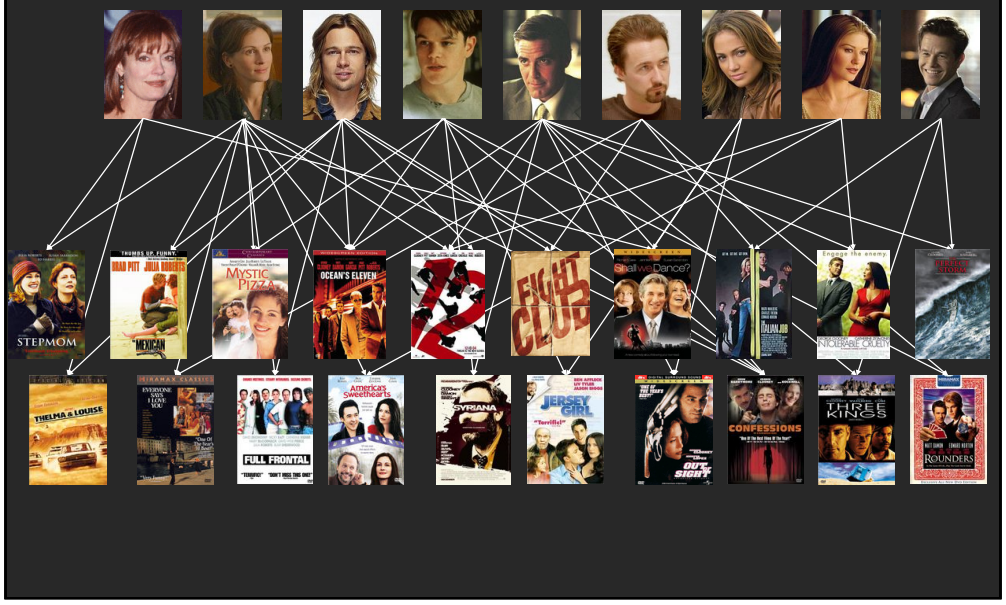
21



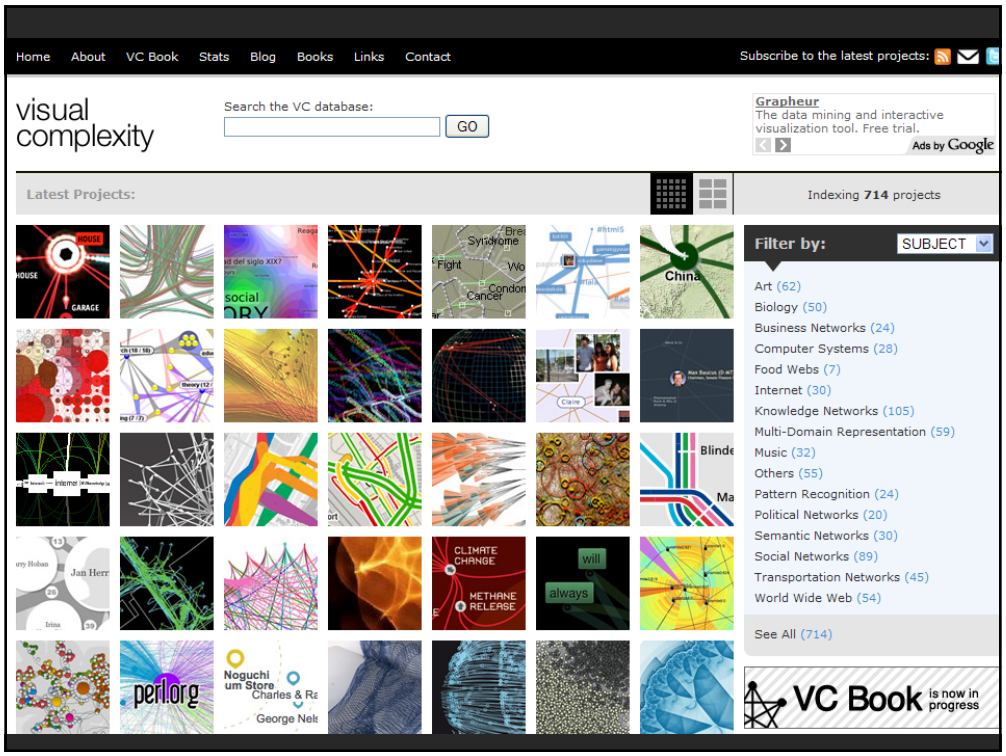
Lombardi, M. 'George W. Bush, Harken Energy and Jackson Stephens, ca 1979–90'

23

Actors and movies (bipartite)



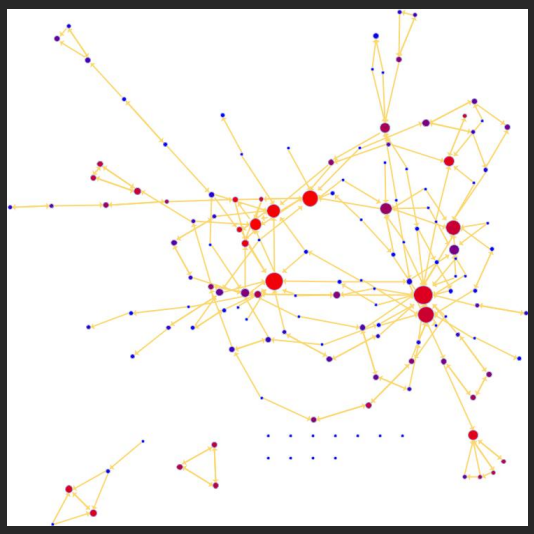
24



25

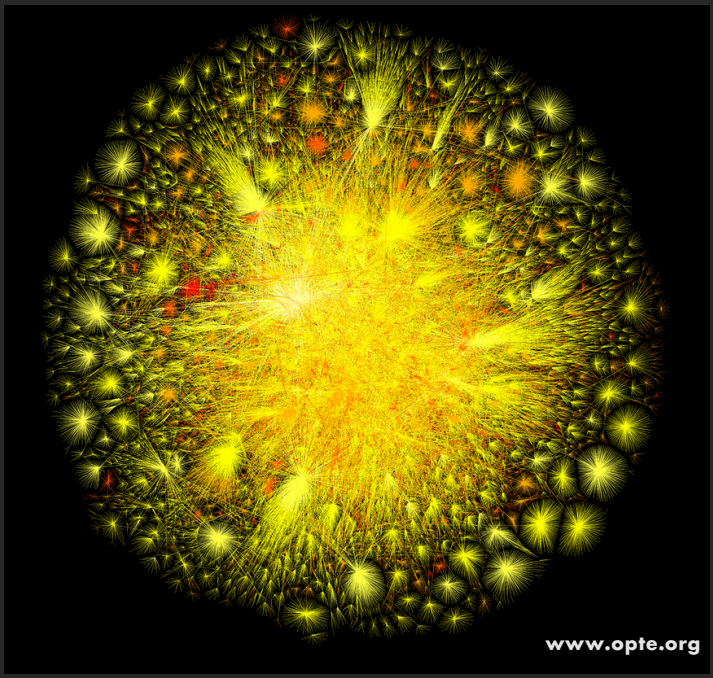
Characterizing networks

What does it look like?



26

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



27

Topics

Network Analysis

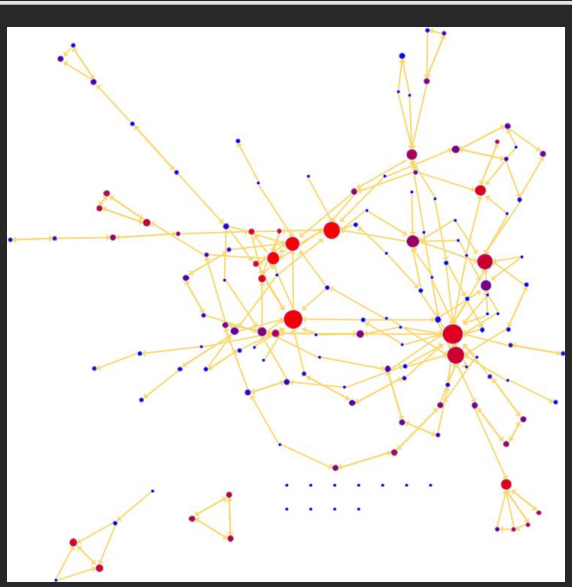
- Centrality / centralization
- Community structure
- Pattern identification
- Models

28

Centrality

29

How far apart are things?



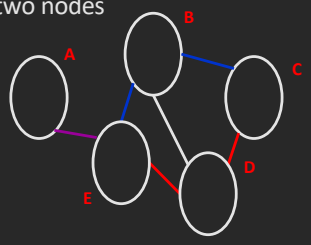
30

Distance: shortest paths

Shortest path (geodesic path)

- The shortest sequence of links connecting two nodes
- Not always unique

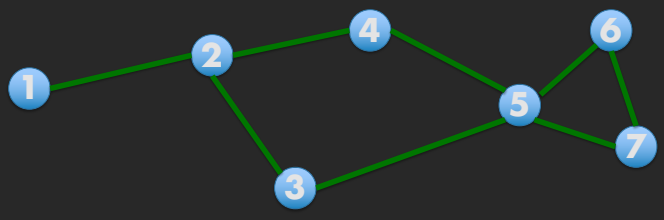
- A and C are connected by 2 shortest paths
  - A – E – B - C
  - A – E – D - C



31

Distance: shortest paths

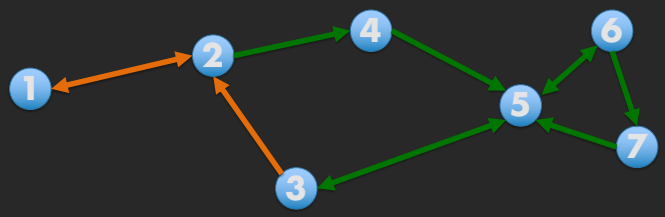
Shortest path from 2 to 3: 1



32

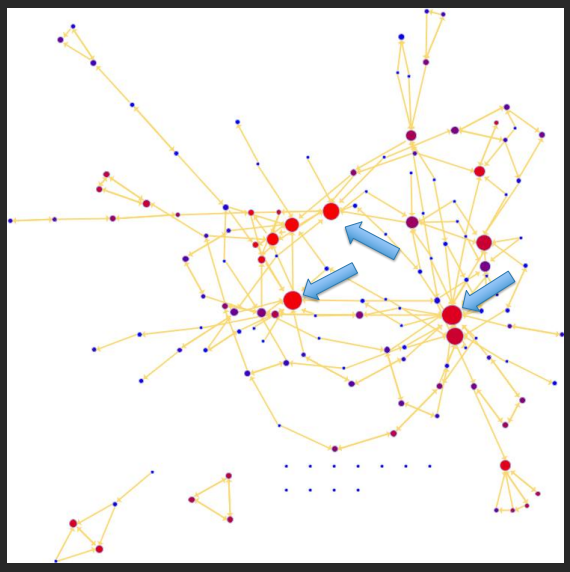
Distance: shortest paths

Shortest path from 2 to 3?



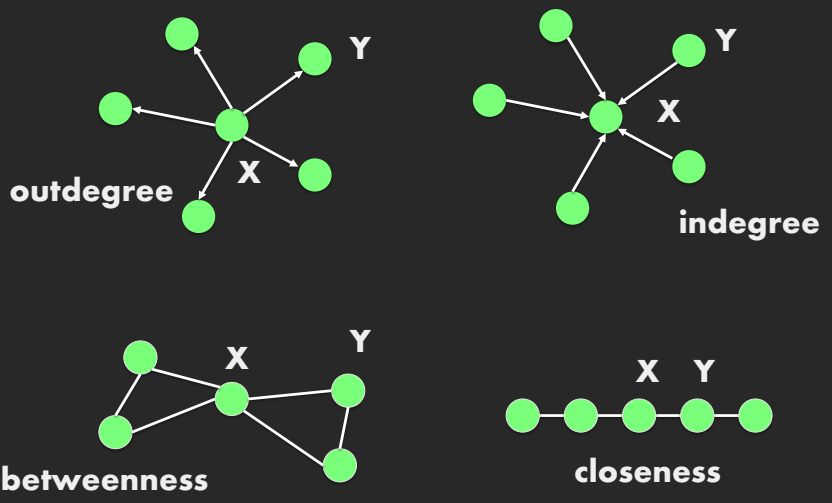
33

Most important node?



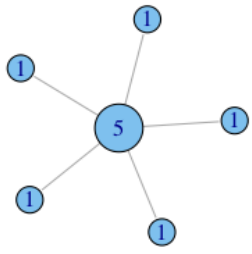
34

Centrality



35

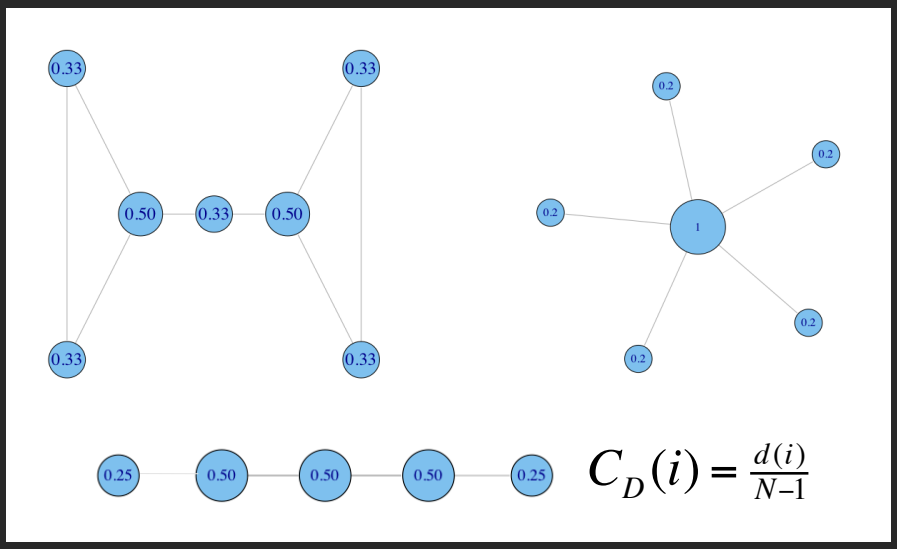
Degree centrality (undirected)



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

36

Normalized degree centrality



37

When is degree not sufficient?

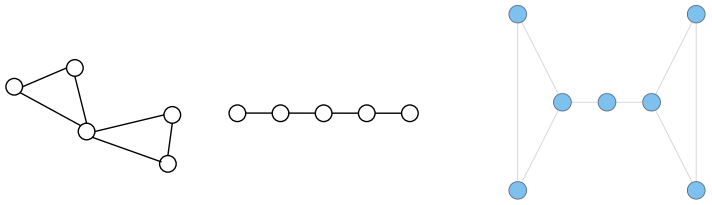
Does not capture

- Ability to broker between groups
- Likelihood that information originating anywhere in the network reaches you

38

Betweenness

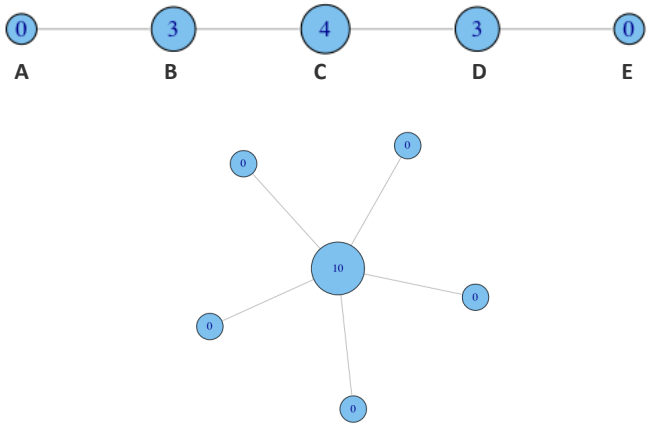
Assuming nodes communicate using the most direct (shortest) route, how many pairs of nodes have to pass information through target node?



39

Betweenness - examples

non-normalized:



40

Betweenness: definition

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

g_{jk} = the number of paths connecting *jk*
$g_{jk}(i)$ = the number that node *i* is on.

Normalization:

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

number of pairs of vertices excluding the vertex itself

41

When are C_d, C_b not sufficient?

Do not capture

Likelihood that information originating anywhere in the network reaches you

43

Closeness: definition

Being close to the center of the graph

Closeness Centrality:

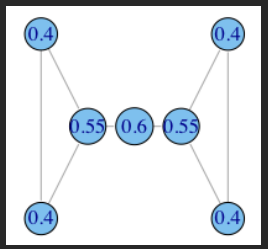
$$C_c(i) = \left[\sum_{j=1, j \neq i}^{N} d(i,j) \right]^{-1}$$

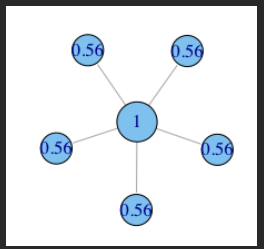
Normalized Closeness Centrality

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

44

Examples - closeness





45

Centrality in directed networks

Prestige ~ indegree centrality

Betweenness ~ consider directed shortest paths

Closeness ~ consider nodes from which target node can be reached

Influence range ~ nodes reachable from target node

Straight-forward modifications to equations for non-directed graphs

46

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.	

47

Centralization – how equal

Variation in the centrality scores among the nodes

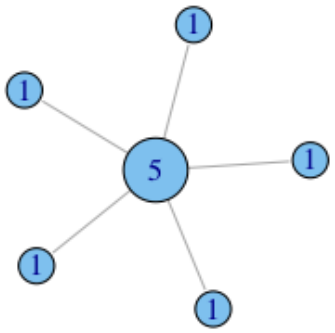
Freeman's general formula for centralization:

maximum value in the network

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

48

Examples

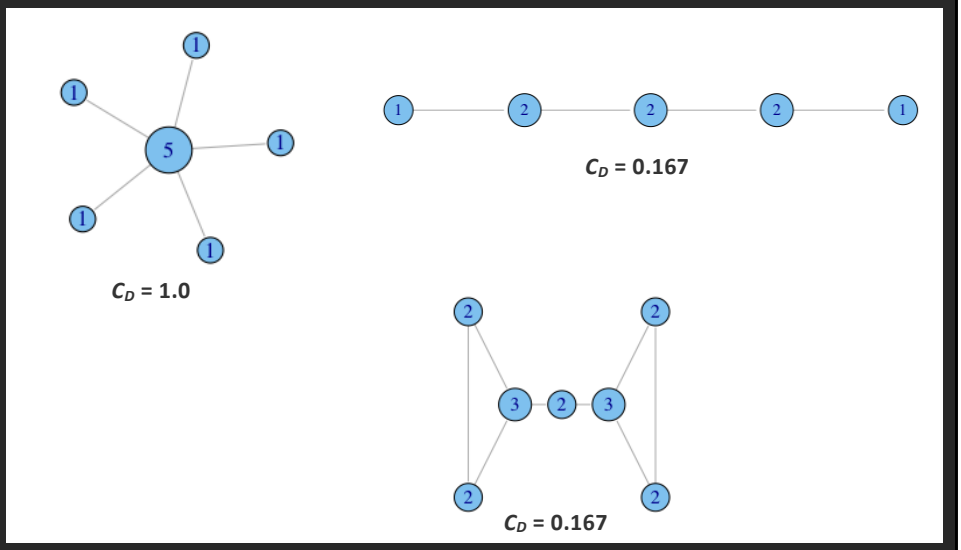


$$C_D = \frac{\sum_{i=1}^{g}\left[C_D(n^*) - C_D(n_i)\right]}{[(N-1)(N-2)]}$$

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

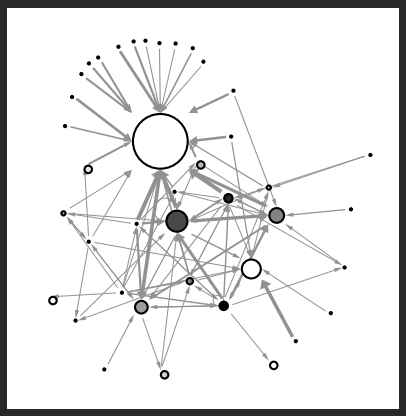
49

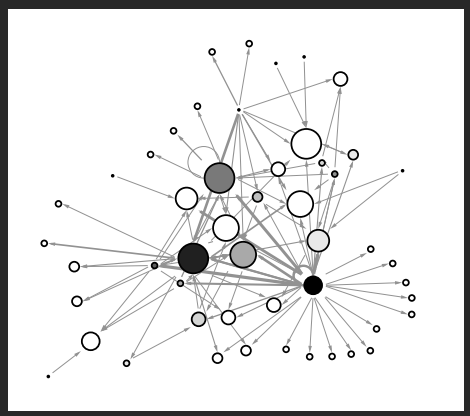
Examples



50

Financial networks



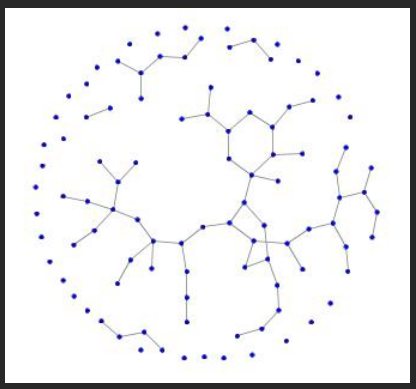


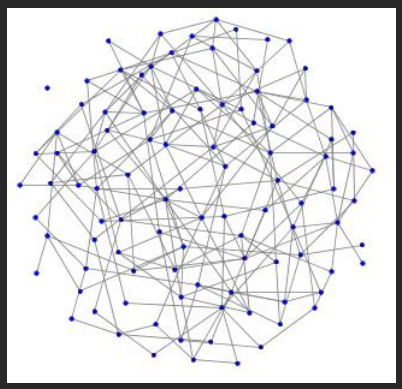
51

Community Structure

55

How dense is it?





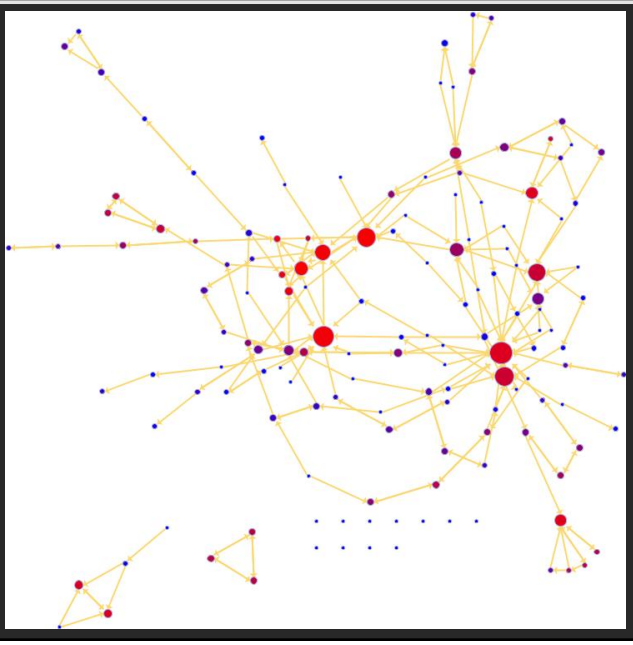
density = e/ e_{max}

Max. possible edges:

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

56

Is everything connected?

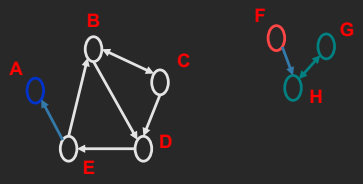


57

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

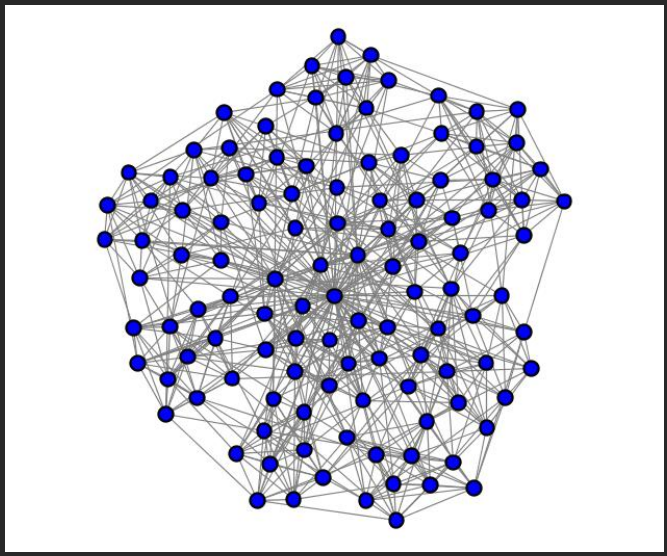


Weakly connected components

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

58

Community finding (clustering)

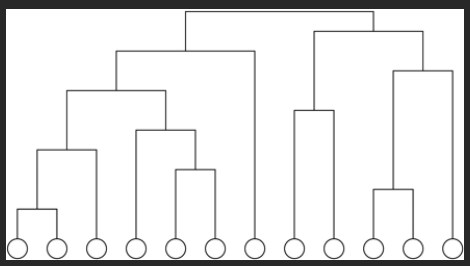


61

Hierarchical clustering

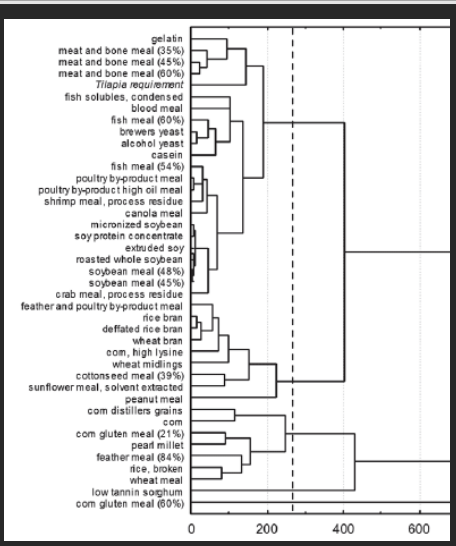
Process:

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



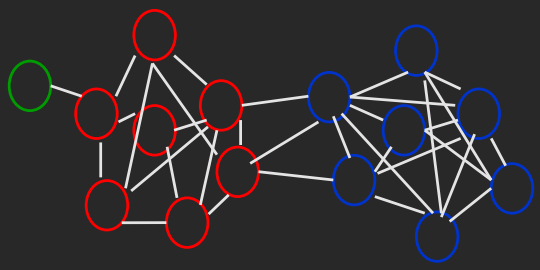
62

Cluster Dendrograms



63

Hierarchical clustering (closeness)

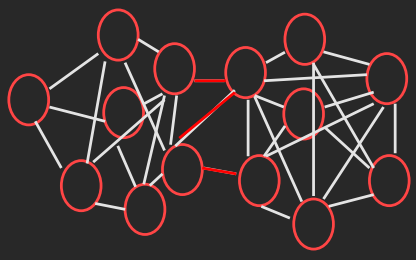


65

Betweenness clustering

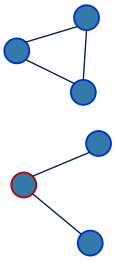
Girvan and Newman 2002 iterative algorithm:

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



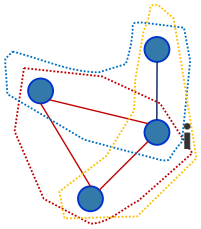
66

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:

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

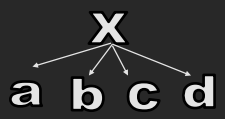
C_i = 1/3 = 0.33

C_G= 3*1/5 = 0.6

67

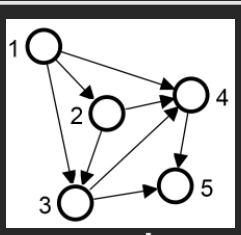
Pattern finding - motifs

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

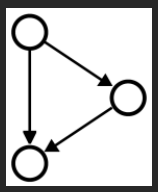


69

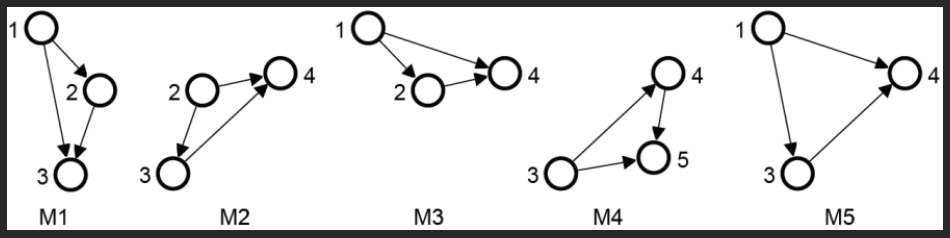
Motifs can overlap in the network



graph



motif to be found

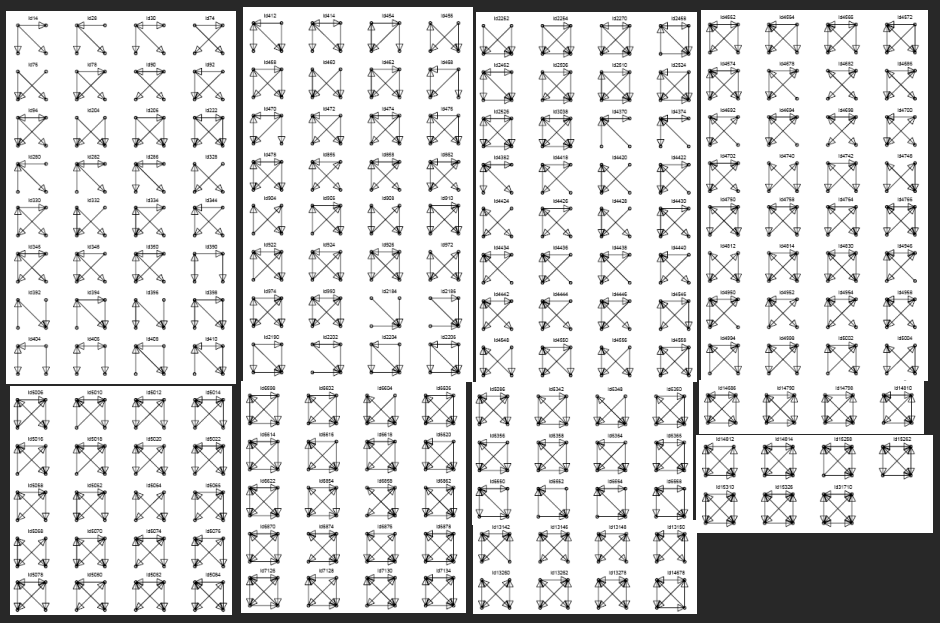


motif matches

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

70

4 node subgraphs



71

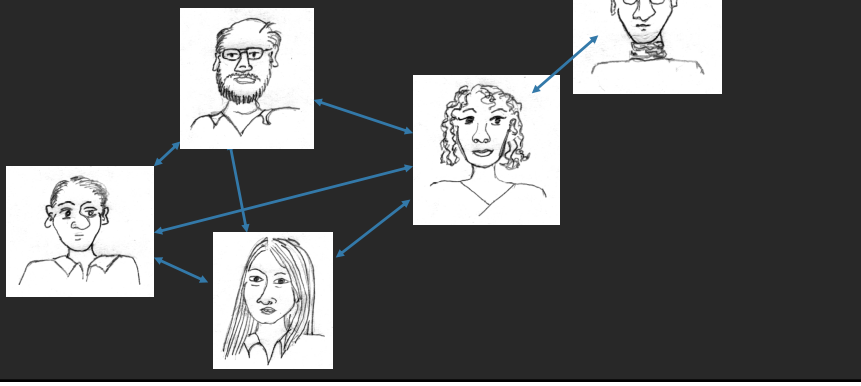
Simulating network models

84

Small world network

Milgram (1967)

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

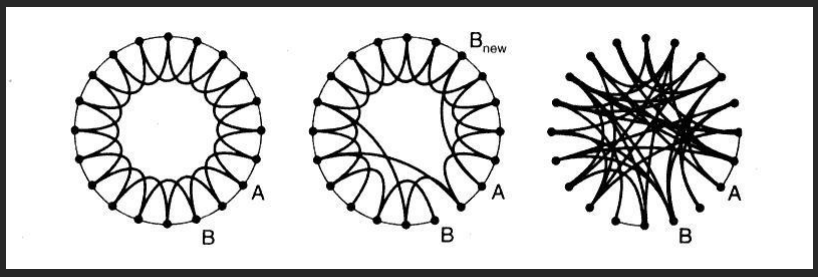


85

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

86

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

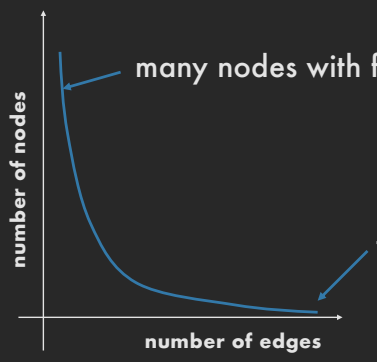
- neural network of C. elegans,
- semantic networks of languages,
- actor collaboration graph
- food webs

87

Power law networks

Many real world networks contain hubs: highly connected nodes

Usually the distribution of edges is extremely skewed



90

Summary

Structural analysis

- Centrality
- Community structure
- Pattern finding

→ Widely applicable across domains

92