Network Analysis

Maneesh Agrawala

CS 448B: Visualization
Fall 2020

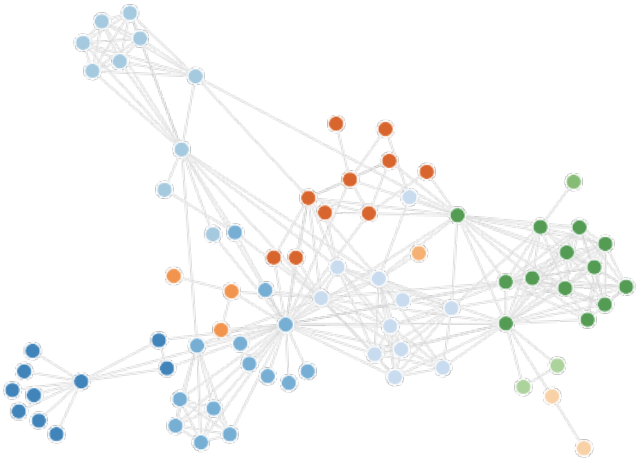
1

Last Time: Network Layout

2

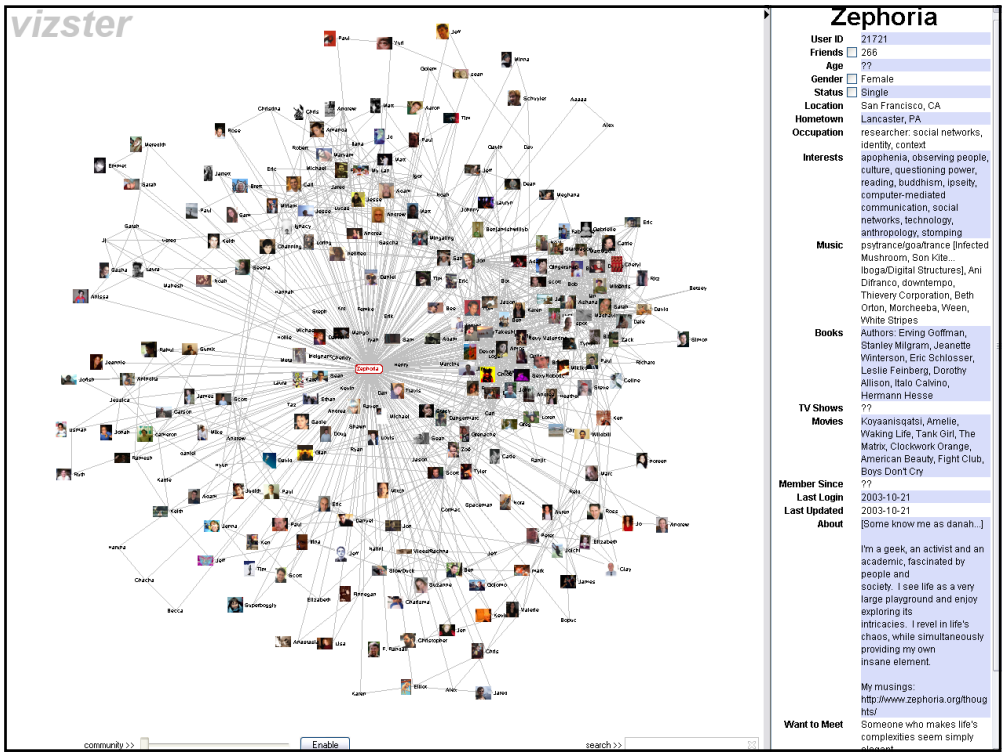
Force-Directed Layout

3

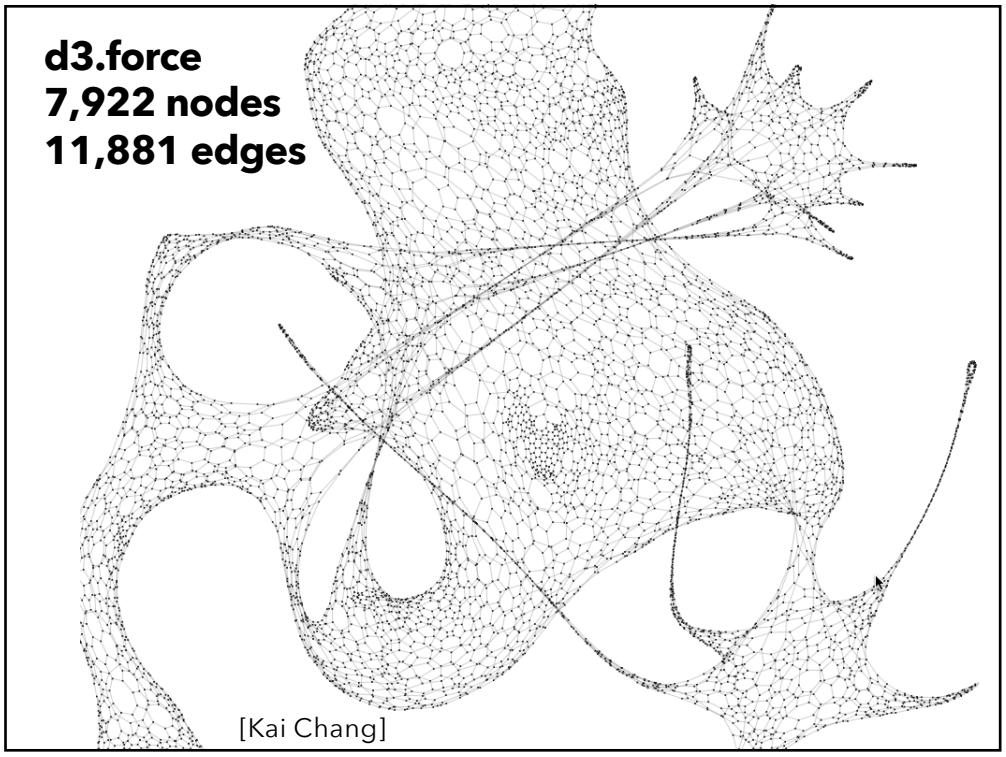


Interactive Example: Configurable Force Layout

4


vizster
Zephoria
User ID 21721
Friends 266
Age ??
Gender Female
Status Single
Location San Francisco, CA
Hometown Lancaster, PA
Occupation researcher: social networks, identity, context
Interests apophenia, observing people, culture, questioning power, reading, buddhism, ipseity, computer-mediated communication, social networks, technology, anthropology, stomping
Music psytrance/goa/trance [Infected Mushroom, Son Kite... Iboga/Digital Structures], Ani Difranco, downtempo, Thievery Corporation, Beth Orton, Morcheeba, Ween, White Stripes
Books Authors: Erving Goffman, Stanley Milgram, Jeanette Winterson, Eric Schlosser, Leslie Feinberg, Dorothy Allison, Italo Calvino, Hermann Hesse
TV Shows ??
Movies Koyaanisqatsi, Amelie, Waking Life, Tank Girl, The Matrix, Clockwork Orange, American Beauty, Fight Club, Boys Don't Cry
Member Since ??
Last Login 2003-10-21
Last Updated 2003-10-21
About [Some know me as danah...]
I'm a geek, an activist and an academic, fascinated by people and society. I see life as a very large playground and enjoy exploring its intricacies. I revel in life's chaos, while simultaneously providing my own insane element.
My musings: http://www.zephoria.org/thoughts/
Want to Meet Someone who makes life's complexities seem simply
community >>
Enable
search >>

5


d3.force
7,922 nodes
11,881 edges
[Kai Chang]

6

Use the Force!

http://mbostock.github.io/d3/talk/20110921/

7

Force-Directed Layout

Nodes = charged particles $F = q_i * q_j / d_{ij}^2$
with air resistance $F = -b * v_i$
Edges = springs $F = k * (L - d_{ij})$

D3's force layout uses velocity Verlet integration

Assume uniform mass *m* and timestep Δt:

$F = ma \rightarrow F = a \rightarrow F = \Delta v / \Delta t \rightarrow F = \Delta v$

Forces simplify to velocity offsets!

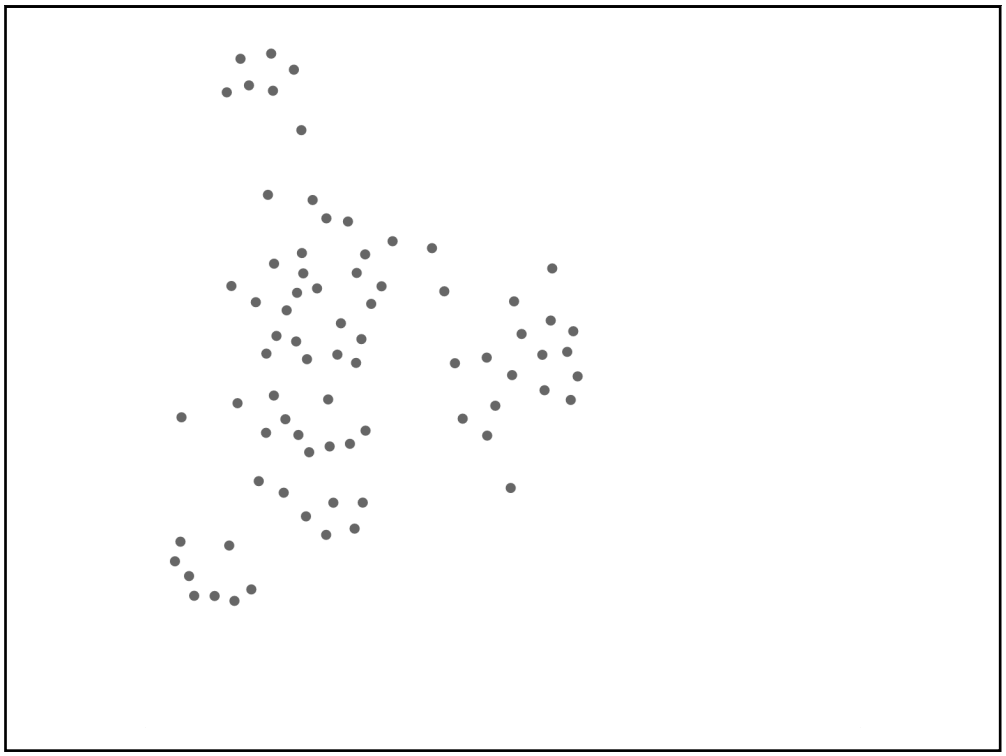
Repeatedly calculate forces, update node positions

Naïve approach $O(N^2)$

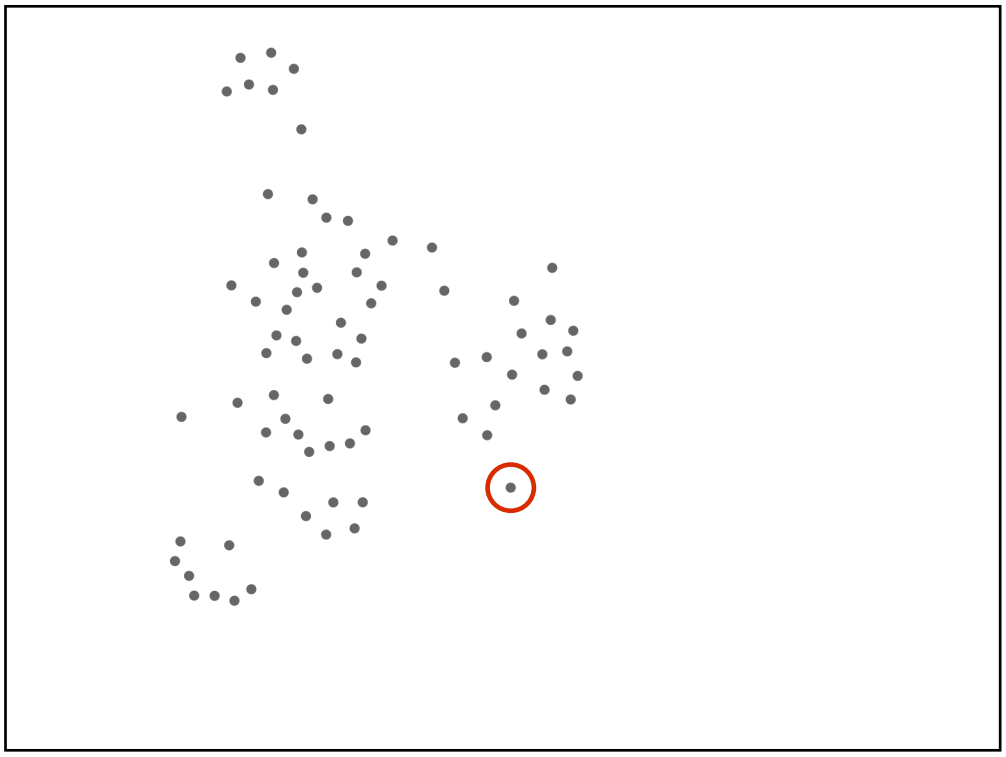
Speed up to O(N log N) using quadtree or k-d tree

Numerical integration of forces at each time step

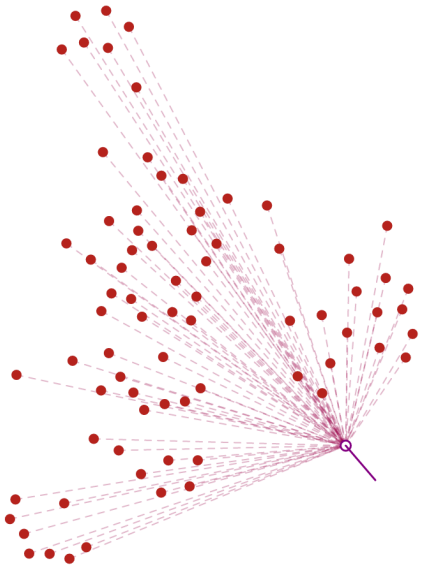
8



9

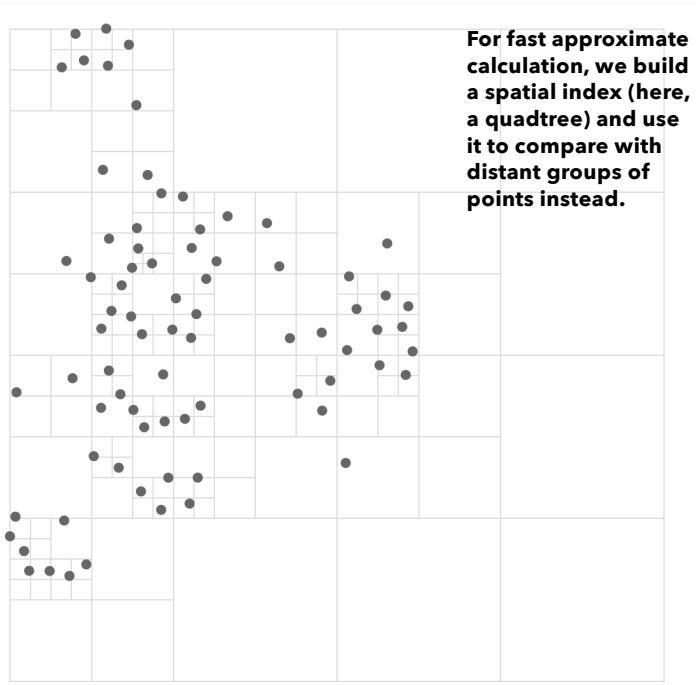


10



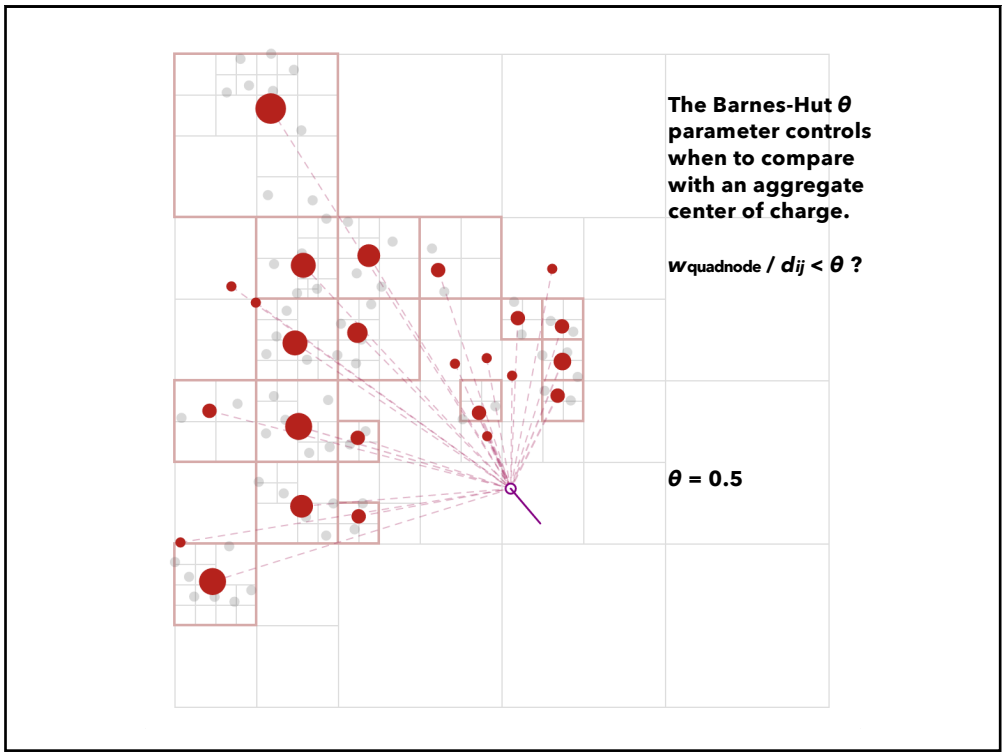
Naive calculation of forces at a point uses sum of forces from all other n-1 points.

11

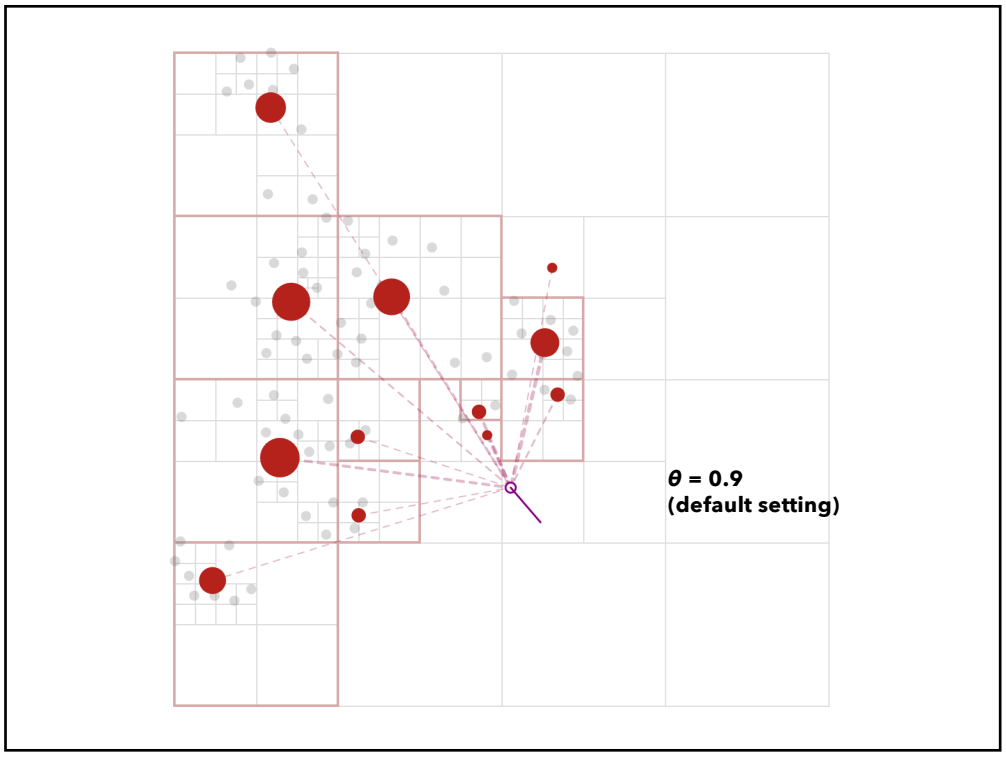


For fast approximate calculation, we build a spatial index (here, a quadtree) and use it to compare with distant groups of points instead.

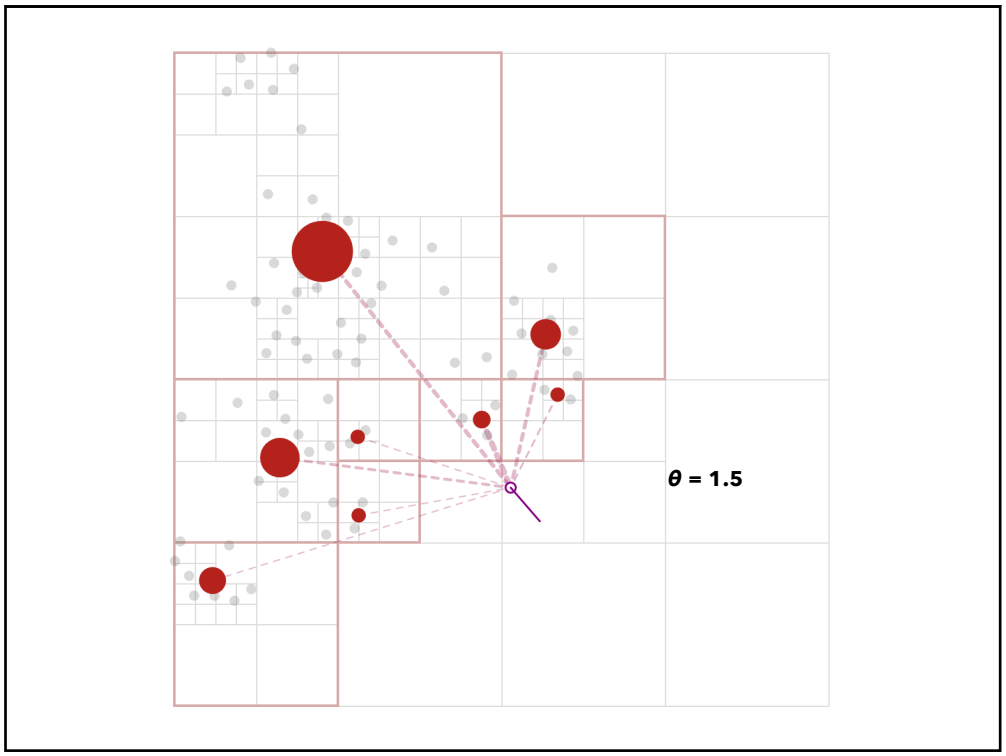
12


The Barnes-Hut θ parameter controls when to compare with an aggregate center of charge.
$w_{quadnode} / d_{ij} < \theta$?
θ = 0.5

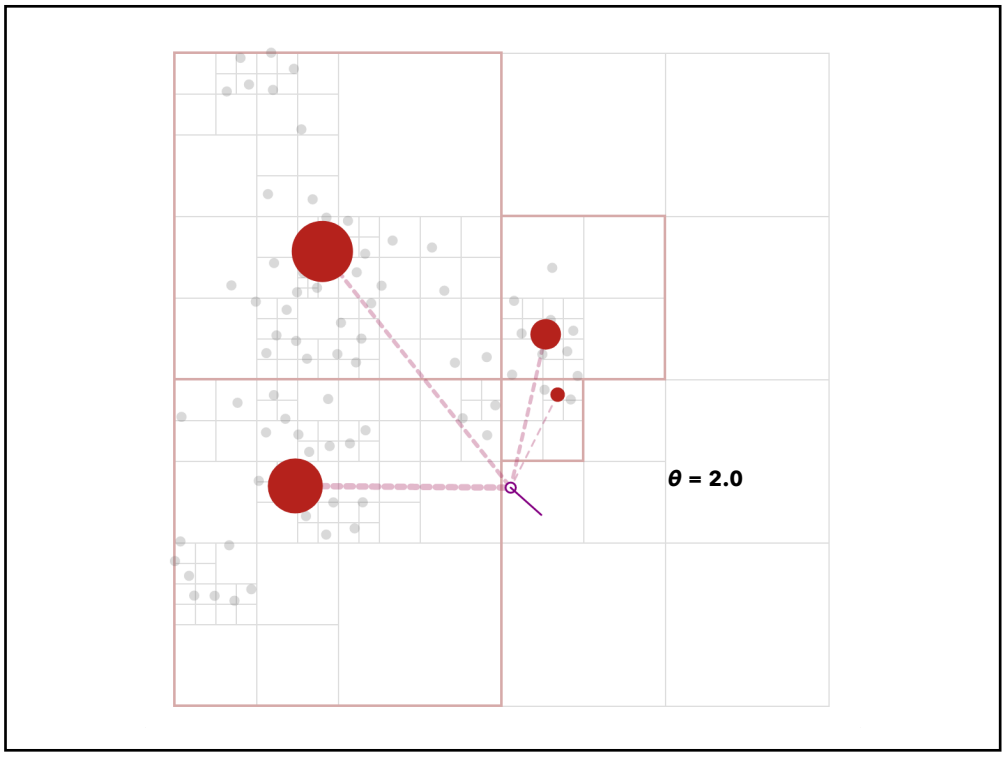
13


θ = 0.9
(default setting)

14



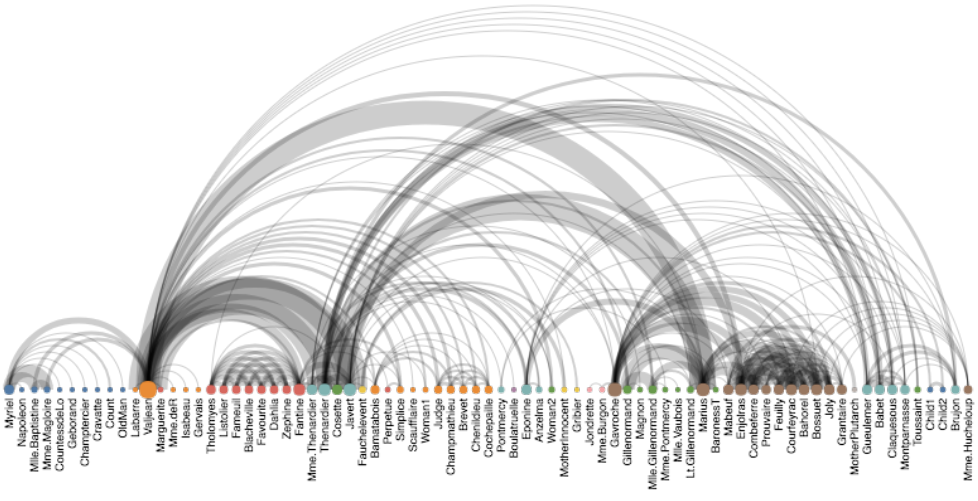
15



16

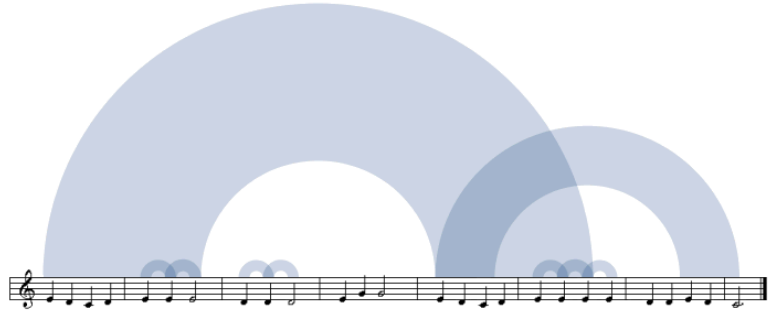
Alternative Layouts

18



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

19



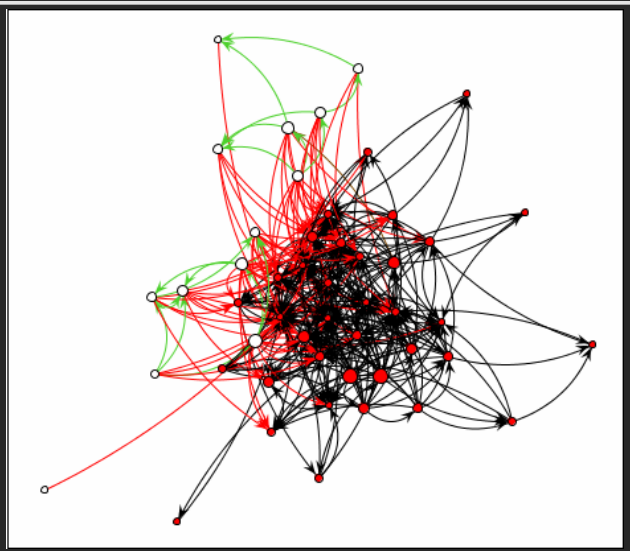
For example, the picture above was built from the first line of a very simple piece: *Mary Had a Little Lamb*. Each arch connects two identical passages. To clarify the connection between the visualization and the song, in this diagram the score is displayed beneath the arches.

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.

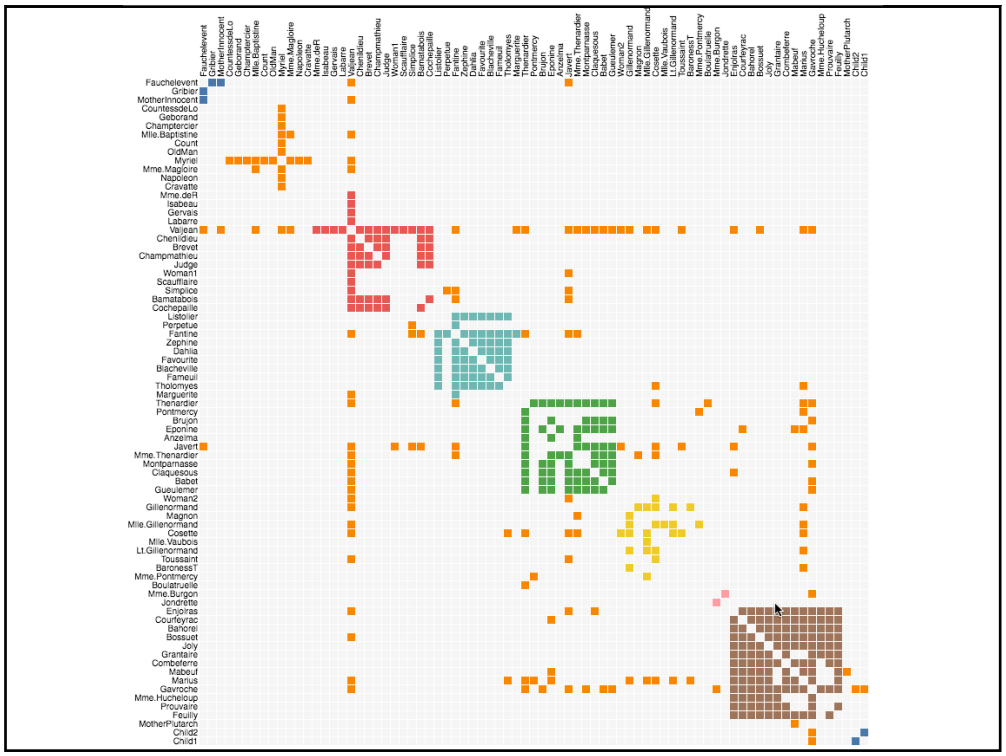
20

Limitations of Node-Link Layout



Edge-crossings and occlusion

21



22

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

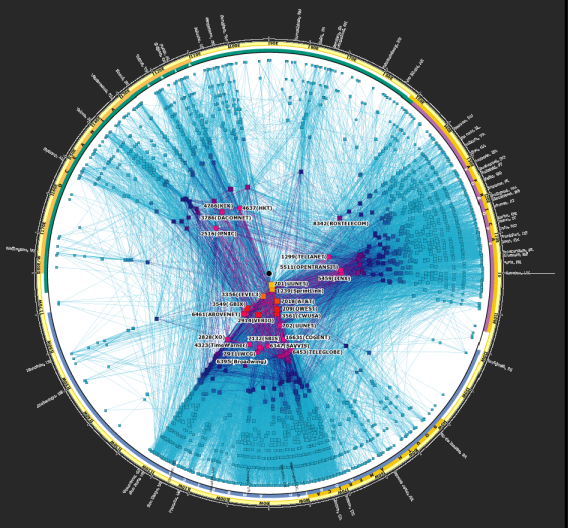
23

Attribute-Driven Layout

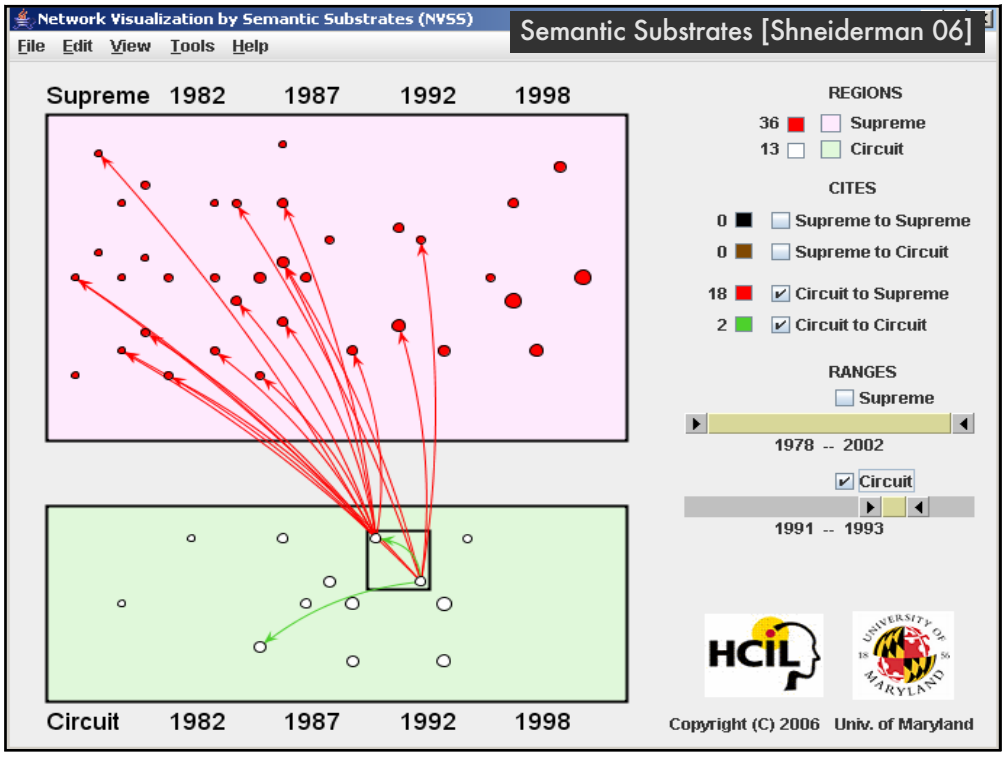
The "Skitter" Layout
- Internet Connectivity
- Radial Scatterplot

Angle = Longitude
- Geography

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

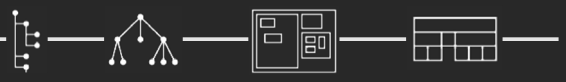


24



25

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

26

Announcements

27

Final project

Data analysis/explainer or conduct research

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

Deliverables

- **Data analysis/explainer**: Article with multiple interactive visualizations
- **Research**: Implementation of solution and web-based demo if possible
- **Short video (2 min)** demoing and explaining the project

Schedule

- Project proposal: Thu 10/29
- Design Review and Feedback: Tue 11/17 & Thu 11/19
- Final code and video: Sat 11/21 11:59pm

Grading

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

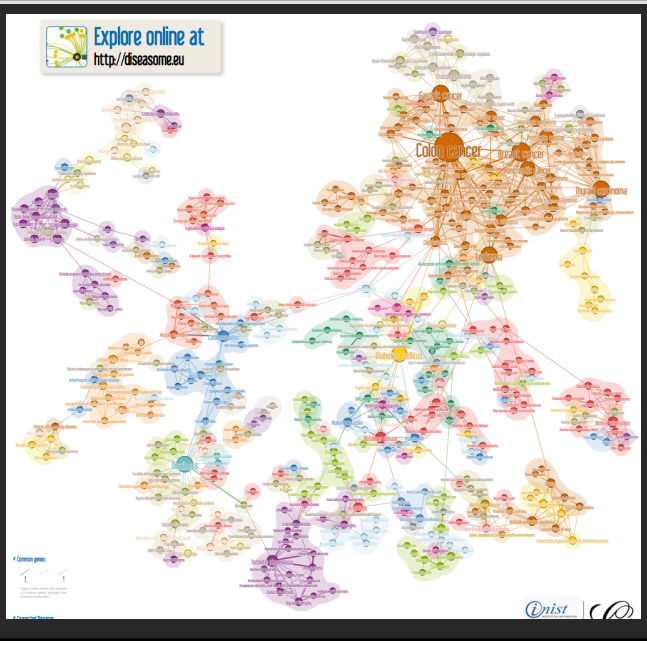
28

Network Analysis

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

29

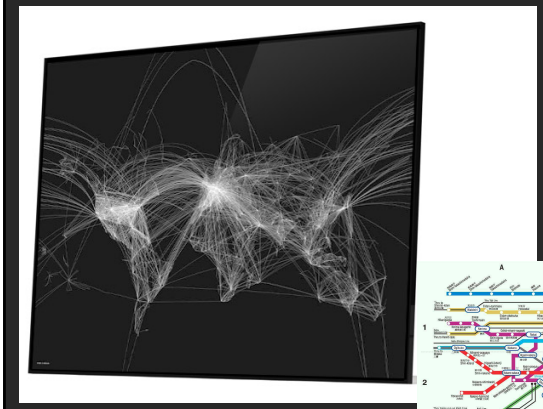
Diseases



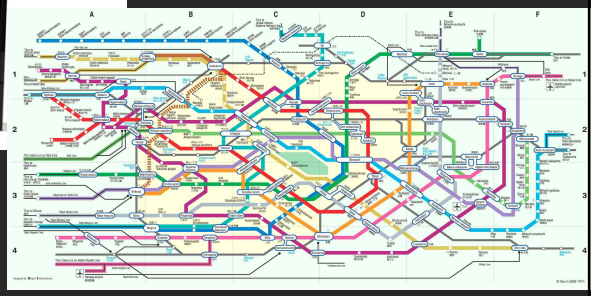
http://diseasome.eu/

31

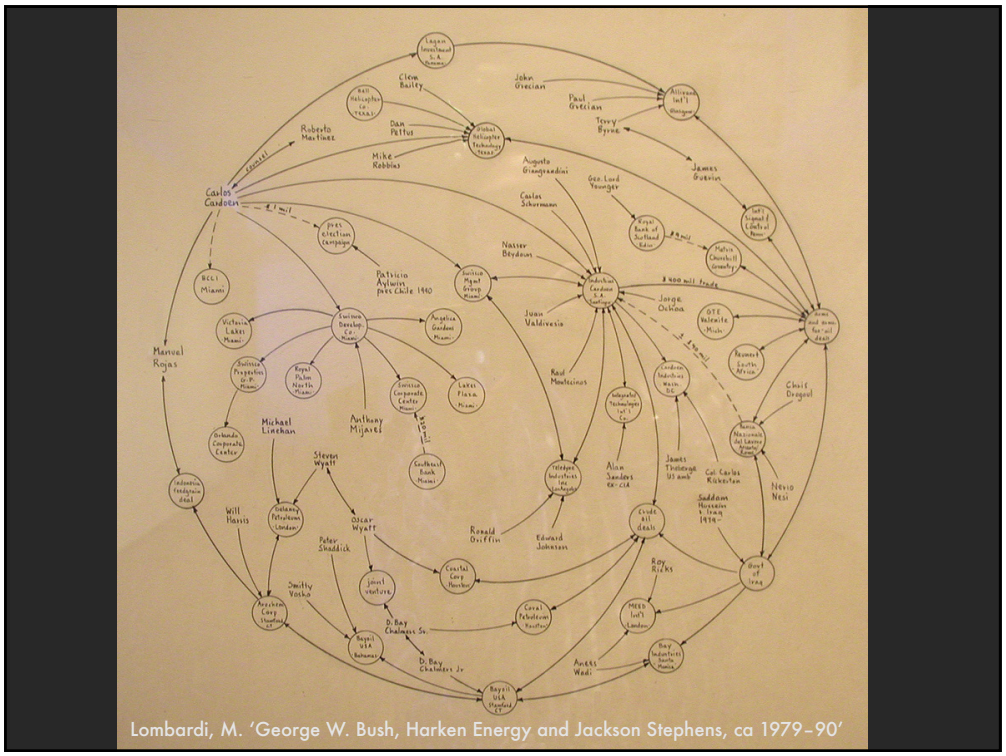
Transportation



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



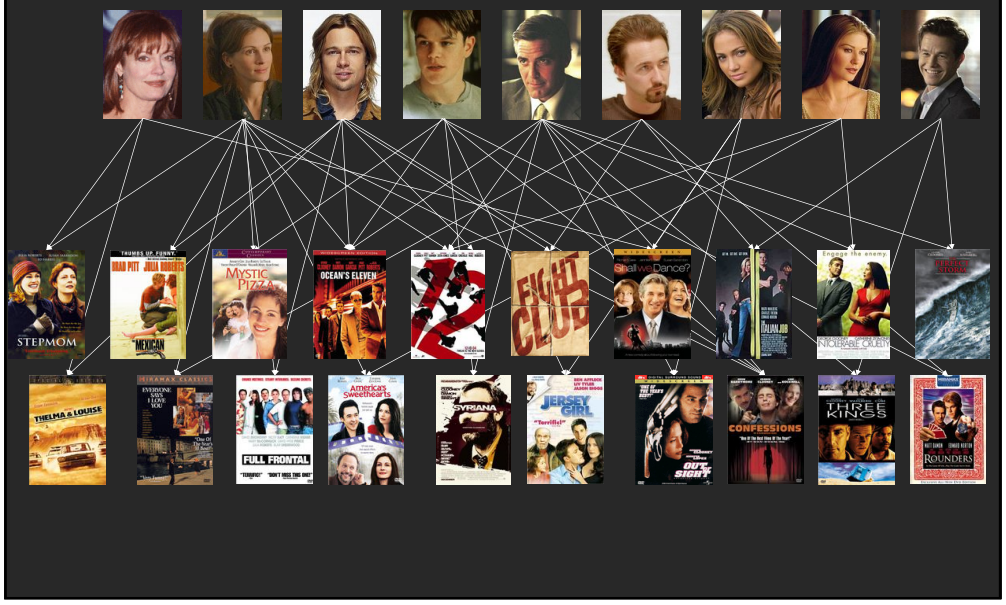
32



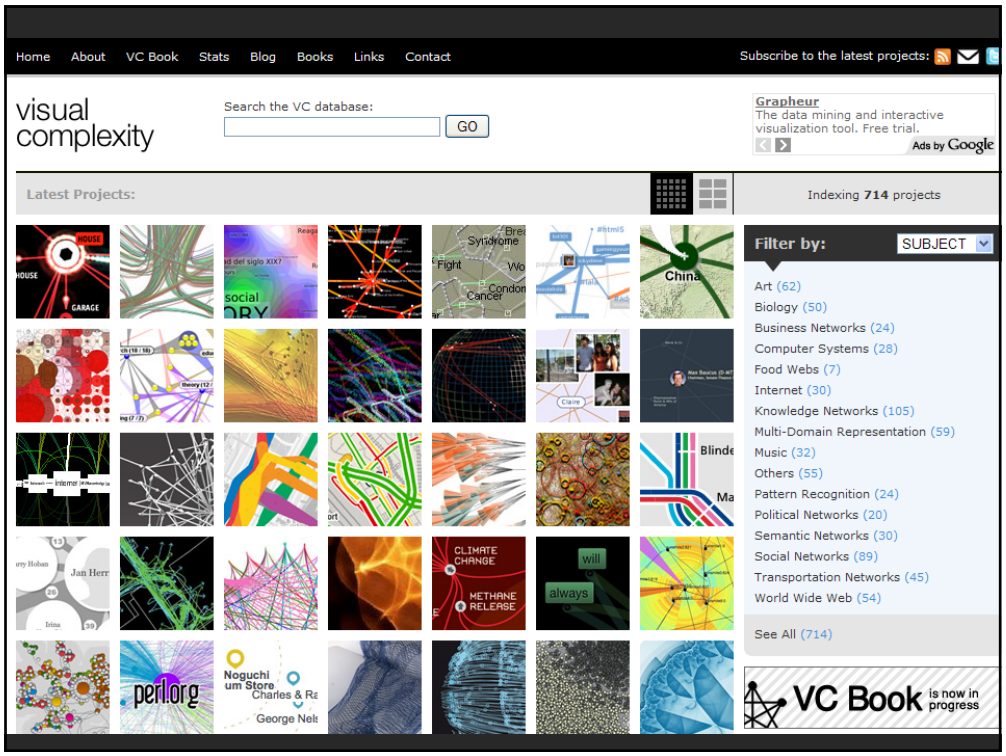
Lombardi, M. 'George W. Bush, Harken Energy and Jackson Stephens, ca 1979–90'

34

Actors and movies (bipartite)



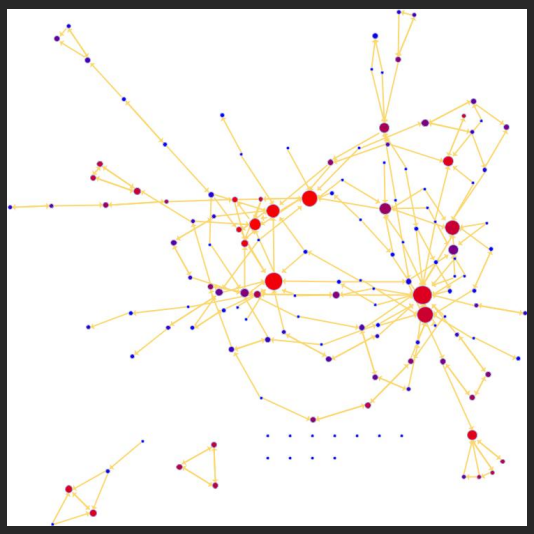
35



36

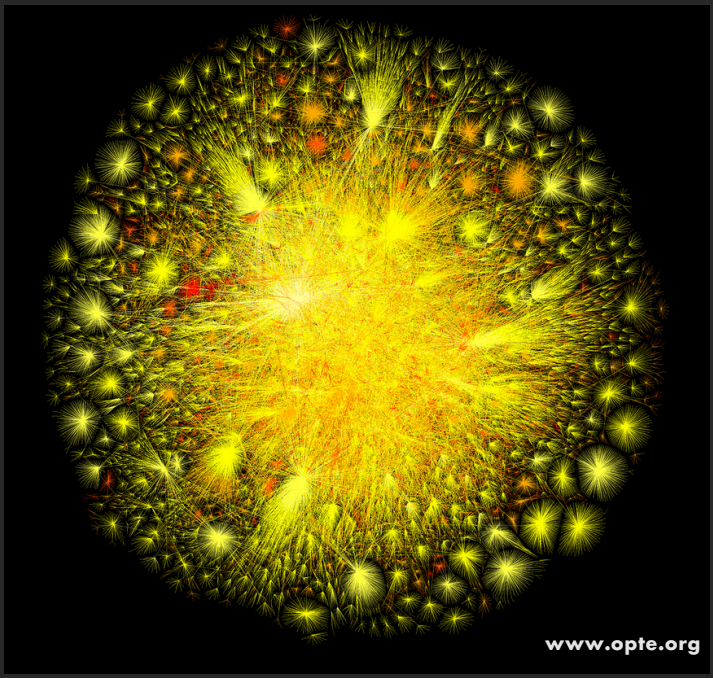
Characterizing networks

What does it look like?



37

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



38

Topics

Network Analysis

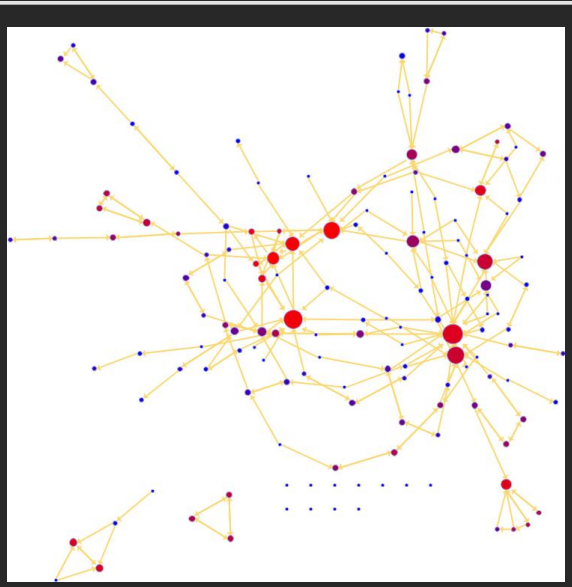
- Centrality / centralization
- Community structure
- Pattern identification
- Models

39

Centrality

40

How far apart are things?



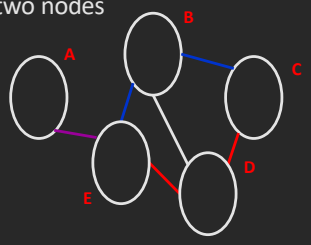
41

Distance: shortest paths

Shortest path (geodesic path)

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

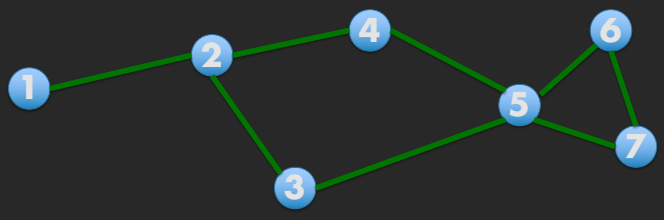
- A and C are connected by 2 shortest paths
  - A – E – B - C
  - A – E – D - C



42

Distance: shortest paths

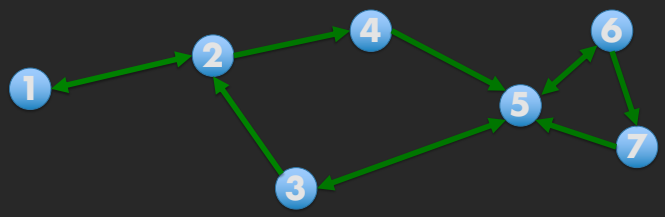
Shortest path from 2 to 3: 1



43

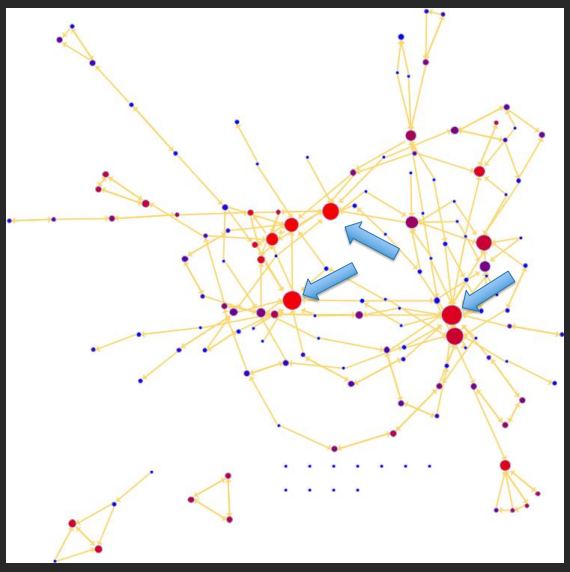
Distance: shortest paths

Shortest path from 2 to 3?



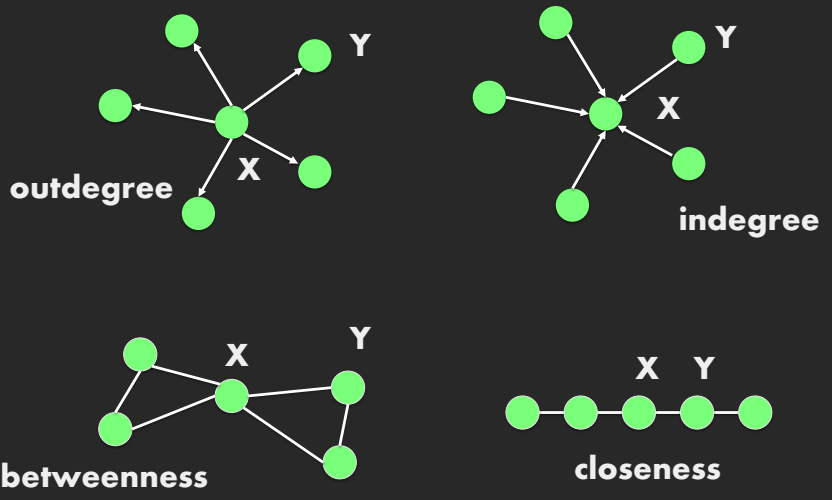
44

Most important node?



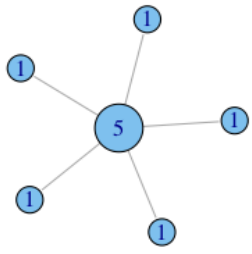
45

Centrality



46

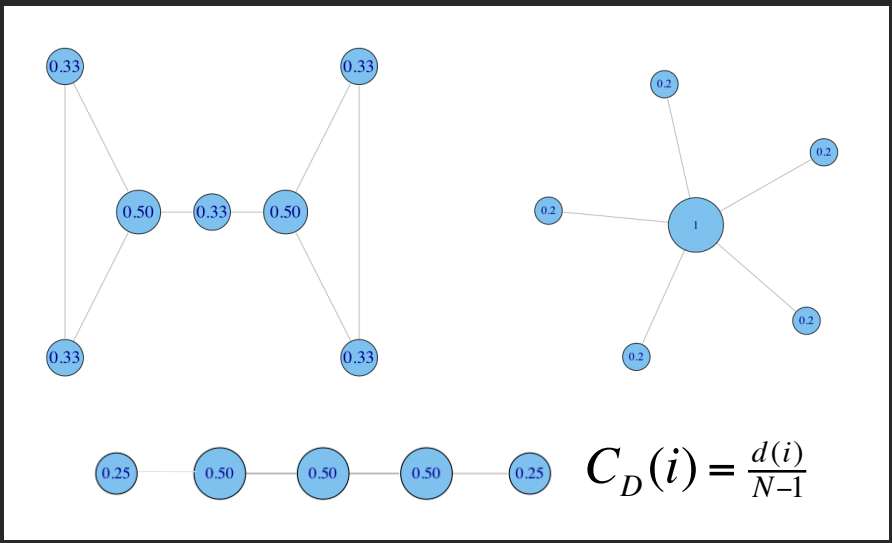
Degree centrality (undirected)



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

47

Normalized degree centrality



48

When is degree not sufficient?

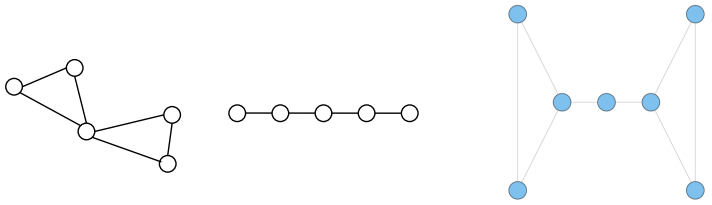
Does not capture

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

49

Betweenness

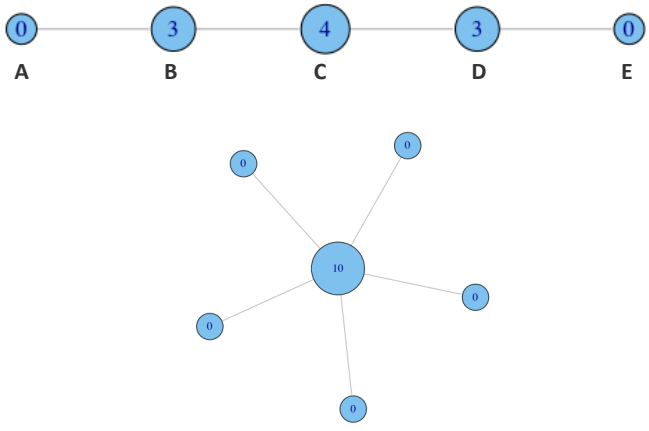
Assuming nodes communicate using the most direct (shortest) route, how many pairs of nodes have to pass information through target node?



50

Betweenness - examples

non-normalized:



51

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

52

When are C_d, C_b not sufficient?

Do not capture

Likelihood that information originating anywhere in the network reaches you

54

Closeness: definition

Being close to the center of the graph

Closeness Centrality:

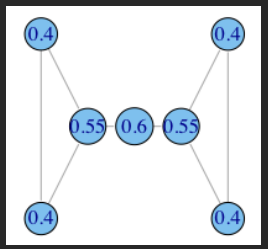
$$C_c(i) = \left[\sum_{j=1, j \neq i}^{N} d(i,j) \right]^{-1}$$

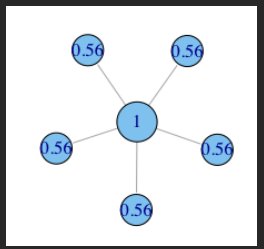
Normalized Closeness Centrality

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

55

Examples - closeness





56

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

57

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.	

58

Centralization – how equal

Variation in the centrality scores among the nodes

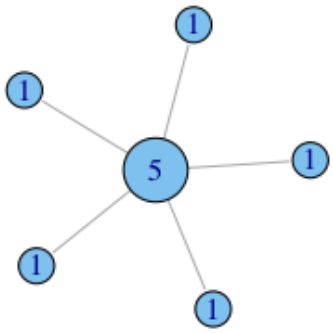
Freeman's general formula for centralization:

maximum value in the network

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

59

Examples

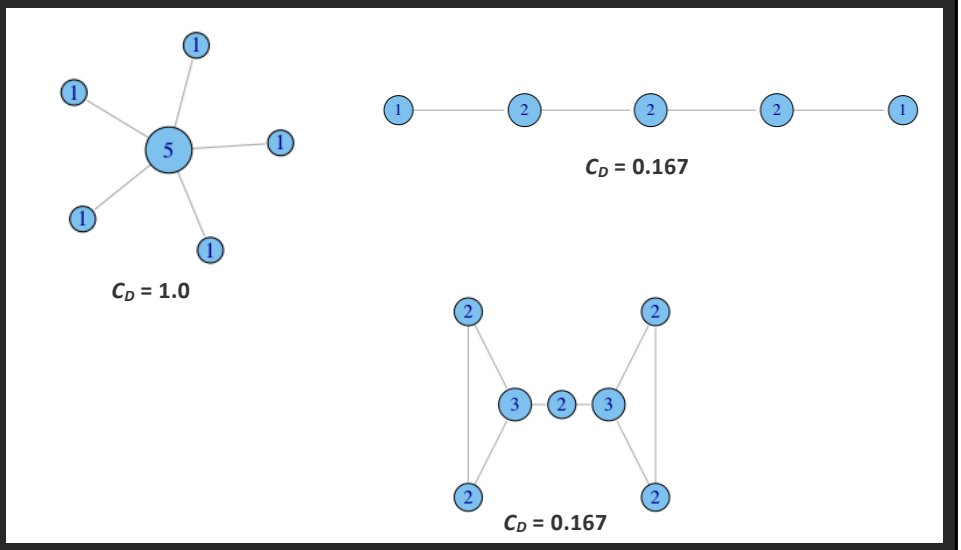


$$C_D = \frac{\sum_{i=1}^{g}\left[C_D(n^*) - C_D(n_i)\right]}{[(N-1)(N-2)]}$$

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

60

Examples

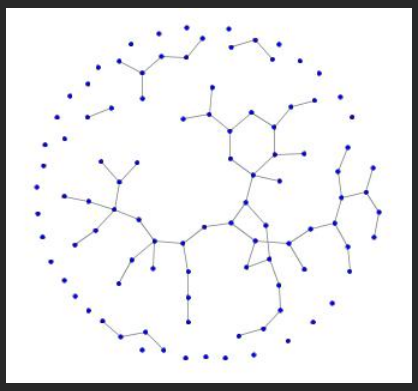


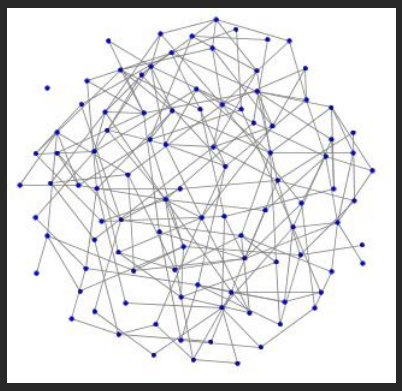
61

Community Structure

66

How dense is it?





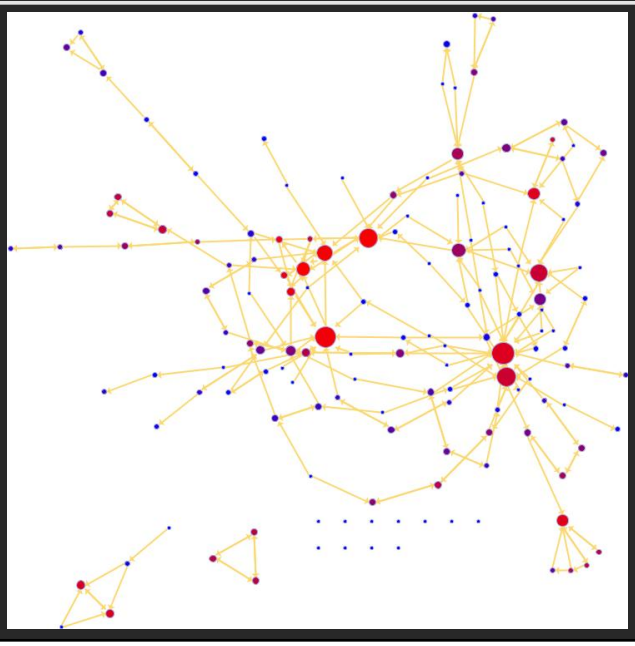
density = e/ e_{max}

Max. possible edges:

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

67

Is everything connected?

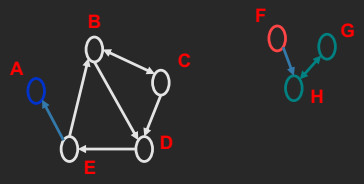


68

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

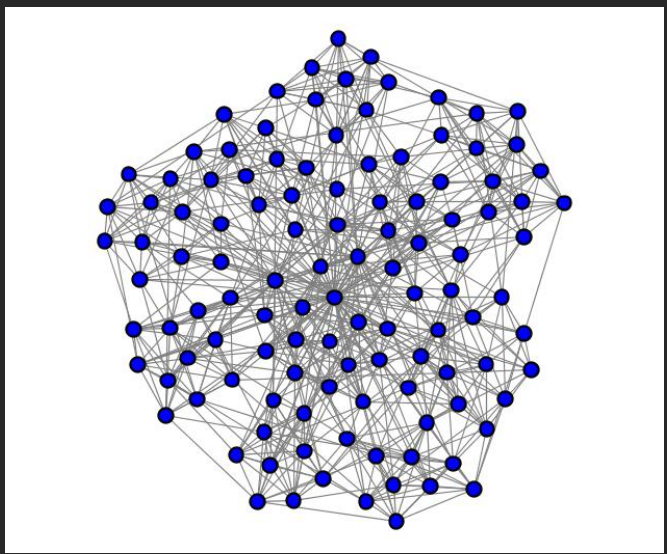


Weakly connected components

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

69

Community finding (clustering)

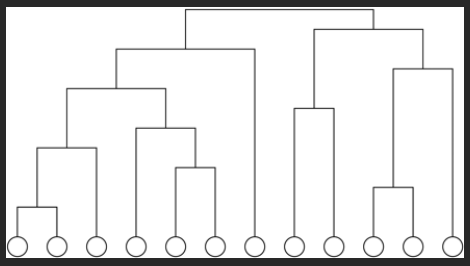


72

Hierarchical clustering

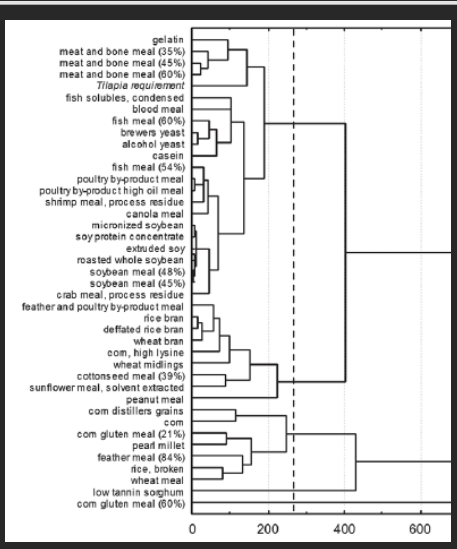
Process:

- 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



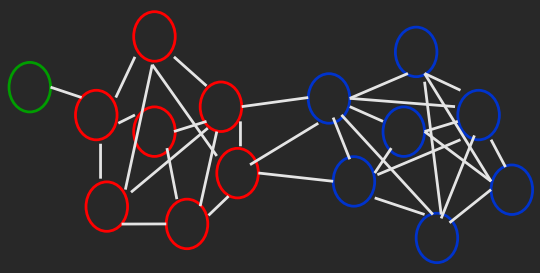
73

Cluster Dendrograms



74

Hierarchical clustering (closeness)

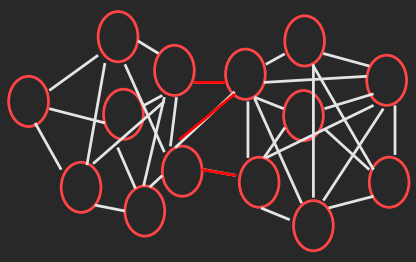


76

Betweenness clustering

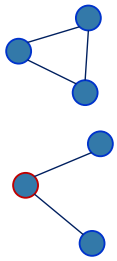
Girvan and Newman 2002 iterative algorithm:

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



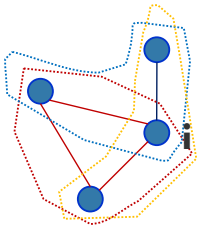
77

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:

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

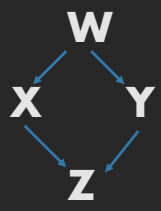
$C_i = 1/3 = 0.33$

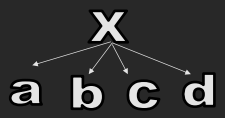
$C_G = 3*1/5 = 0.6$

78

Pattern finding - motifs

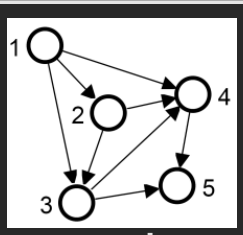
Define / search for a particular structure, e.g. complete triads



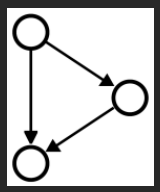


80

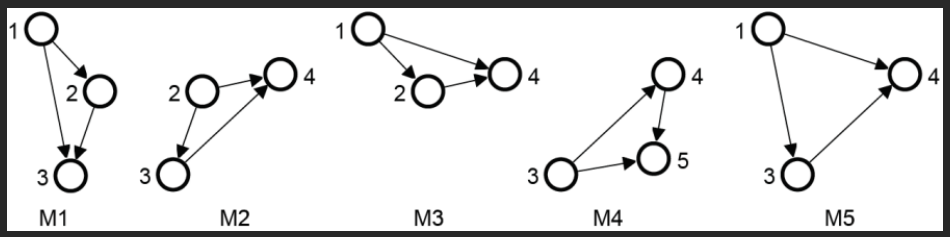
Motifs can overlap in the network



graph



motif to be found

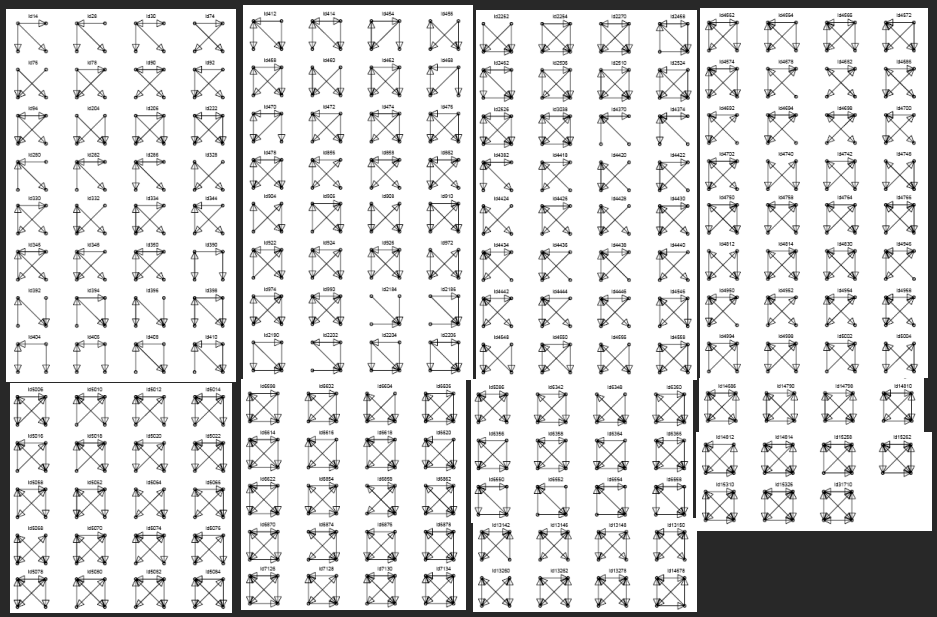


motif matches

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

81

4 node subgraphs



82

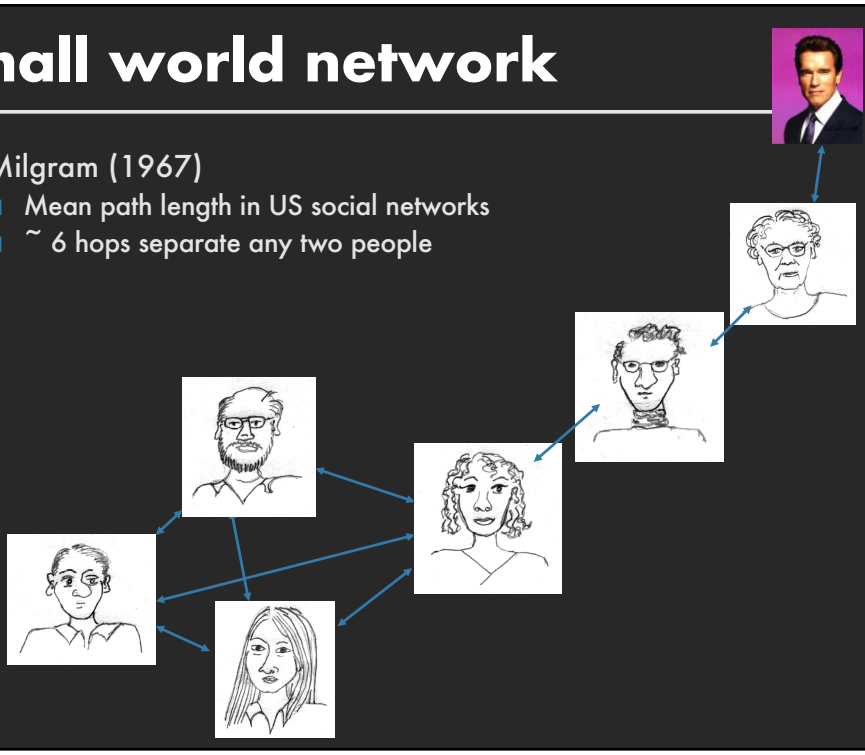
Simulating network models

95

Small world network

Milgram (1967)

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

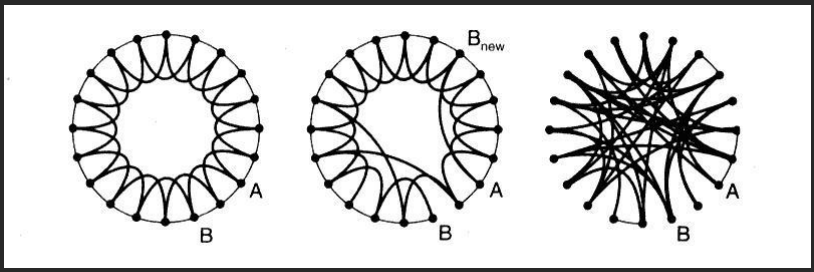


96

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

97

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

98

Summary

Structural analysis

- Centrality
- Community structure
- Pattern finding

Widely applicable across domains

103