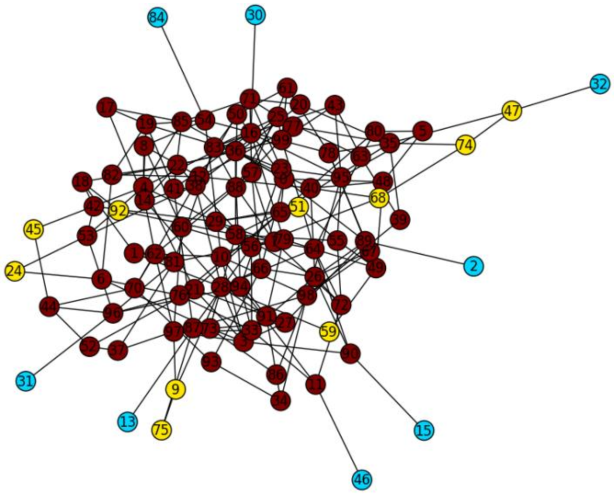
Topological Analysis (1)



Hiroki Sayama
sayama@binghamton.edu

Network data import & export

- read_gml

- read_adjlist
- read_edgelist
  - Creates undirected graphs by default; use “create_using=NX.DiGraph()” option to generate directed graphs

Exercise

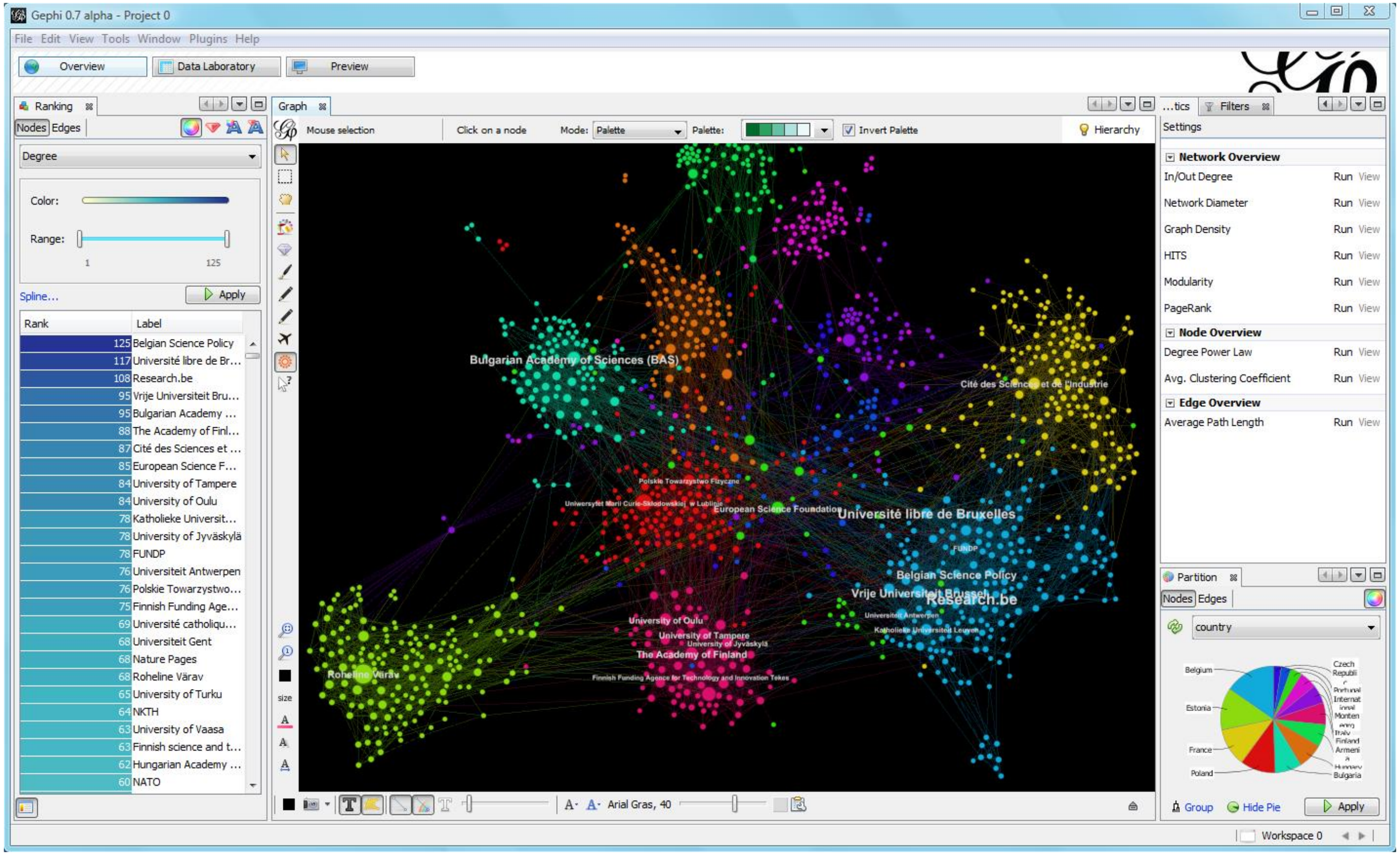
- Import Supreme Court Citation Network Data into NetworkX (http://jhfowler.ucsd.edu/judicial.htm)
  - Import as an undirected graph
  - Import as a directed graph

Network visualization

- “nx.draw”
- Various layout functions
  - Spring, circular, random, spectral, etc.
- For visualization of large-scale networks, use “Gephi”

Gephi

- **Network visualization & analysis tool**



Degree

Mode: Palette

Palette: 🟩

☑ Invert Palette

Rank	Label
125	Belgian Science Policy
117	Université libre de Br...
108	Research.be
95	Vrije Universiteit Bru...
95	Bulgarian Academy ...
88	The Academy of Finl...
87	Cité des Sciences et ...
85	European Science F...
84	University of Tampere
84	University of Oulu
78	Katholieke Universit...
78	University of Jyväskylä
78	FUNDP
76	Universiteit Antwerpen
76	Polskie Towarzystwo...
75	Finnish Funding Age...
69	Université catholiqu...
68	Universiteit Gent
68	Nature Pages
68	Roheline Värav
65	University of Turku
64	NKTH
63	University of Vaasa
63	Finnish science and t...
62	Hungarian Academy ...
60	NATO

Settings

Network Overview

In/Out Degree	Run	View
Network Diameter	Run	View
Graph Density	Run	View
HITS	Run	View
Modularity	Run	View
PageRank	Run	View

Node Overview

Degree Power Law	Run	View
Avg. Clustering Coefficient	Run	View

Edge Overview

Average Path Length	Run	View

country

Basic Properties of Networks

Basic properties of networks

- Number of nodes
- Number of links
- Network density

- Connected components

Network density

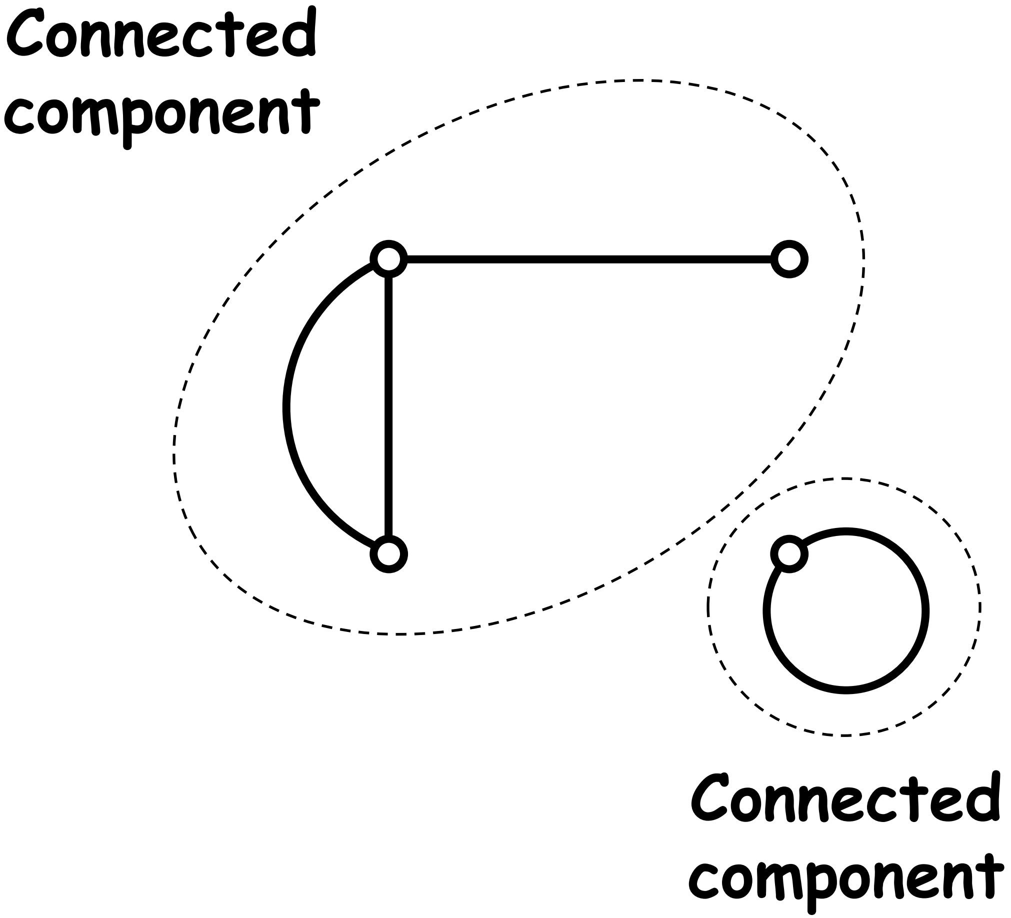
- The ratio of # of actual links and # of possible links
  - For an undirected graph:

    $d = |E| / (|V| (|V| - 1) / 2)$

  - For a directed graph:

    $d = |E| / (|V| (|V| - 1))$

Connected components



Number of connected components = 2

Exercise

- Measure the following for the (undirected) Supreme Court Citation Network
  - Number of nodes, links
  - Network density
  - Number of connected components
  - Size of the largest connected component
  - Distribution of the sizes of connected components

Shortest path lengths, etc.

- shortest_path
- shortest_path_length
- eccentricity
  - Max shortest path length from each node
- diameter
  - Max eccentricity in the network
- radius
  - Min eccentricity in the network

Exercise

- Draw the Karate Club network with its nodes painted with different colors according to their eccentricity

Characteristic path length

- Average shortest path length over all pairs of nodes
- Characterizes how large the world represented by the network is
  - A small length implies that the network is well connected globally

Clustering coefficient

- For each node:
  - Let n be the number of its neighbor nodes
  - Let m be the number of links among the k neighbors
  - Calculate c = m / (n choose 2)

  Then C = <c> (the average of c)
- C indicates the average probability for two of one's friends to be friends too
  - A large C implies that the network is well connected locally to form a cluster

Exercise

- Measure the average clustering coefficients of the following network:
  - Karate Club graph
  - Krackhardt Kite graph
  - Supreme Court Citation network
  - Any other network of your choice
- Compare them and discuss
  - Can you tell anything meaningful?

Centralities

Centrality measures (“B,C,D,E”)

- Degree centrality
  - How many connections the node has
- Betweenness centrality
  - How many shortest paths go through the node
- Closeness centrality
  - How close the node is to other nodes
- Eigenvector centrality

Degree centrality

- Simply, # of links attached to a node

$$C_D(v) = deg(v)$$

or sometimes defined as

$$C_D(v) = deg(v) / (N-1)$$

Betweenness centrality

- Prob. for a node to be on shortest paths between two other nodes

$$C_B(v) = \frac{1}{(n-1)(n-2)} \sum_{s \neq v, e \neq v} \frac{\#sp_{(s,e,v)}}{\#sp_{(s,e)}}$$

  - s: start node, e: end node
  - $\#sp_{(s,e,v)}$: # of shortest paths from s to e that go though node v
  - $\#sp_{(s,e)}$: total # of shortest paths from s to e
  - Easily generalizable to "group betweenness"

Closeness centrality

- Inverse of an average distance from a node to all the other nodes

$$C_C(v) = \frac{n-1}{\Sigma_{w \neq v} d(v,w)}$$

- d(v,w): length of the shortest path from v to w
- Its inverse is called "farness"
- Sometimes "Σ" is moved out of the fraction (it works for networks that are not strongly connected)
- NetworkX calculates closeness within each connected component

Eigenvector centrality

- Eigenvector of the largest eigenvalue of the adjacency matrix of a network

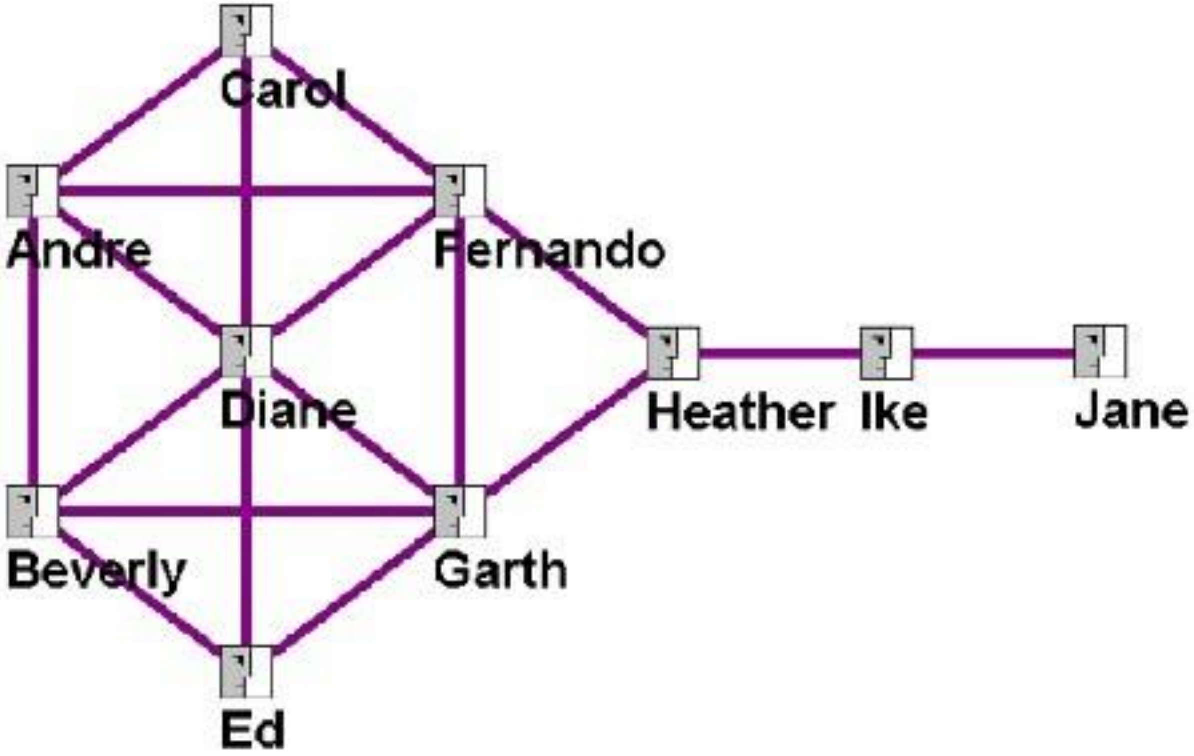
$$C_E(v) = (v\text{-th element of } x)$$

$$Ax = \lambda x$$

- λ: dominant eigenvalue
- x is often normalized ($|x| = 1$)

Exercise

- Who is most central by degree, betweenness, closeness, eigenvector?



Which centrality to use?

- To find the most popular person
- To find the most efficient person to collect information from the entire organization
- To find the most powerful person to control information flow within an organization
- To find the *most important* person (?)

Exercise

- Measure four different centralities for all nodes in the Karate Club network and visualize the network by coloring nodes with their centralities

Exercise

- Create a directed network of any kind and measure centralities

- Make it undirected and do the same
  - How are the centrality measures affected?

Randomizing Network Topologies

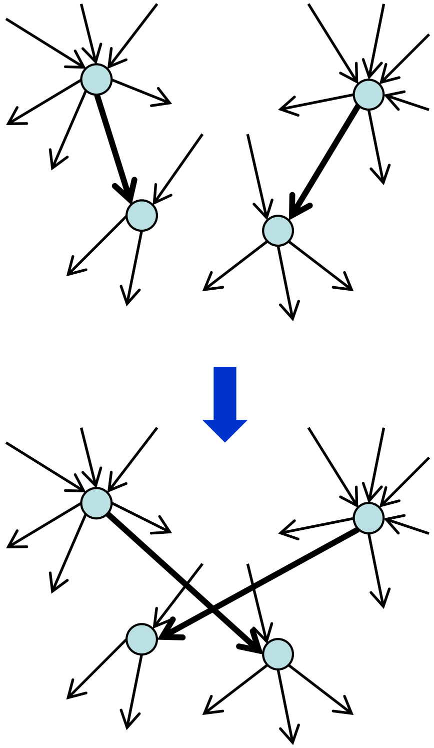
Randomizing networks

- Construct a “null model” network samples to test statistical significance of experimentally observed properties
  - Randomized while some network properties are preserved (e.g., degrees)
  - If the observed properties still remain after randomization, they were simply caused by the preserved properties
  - If not, something else was causing them

Randomlization method (1)

- Double edge swap method

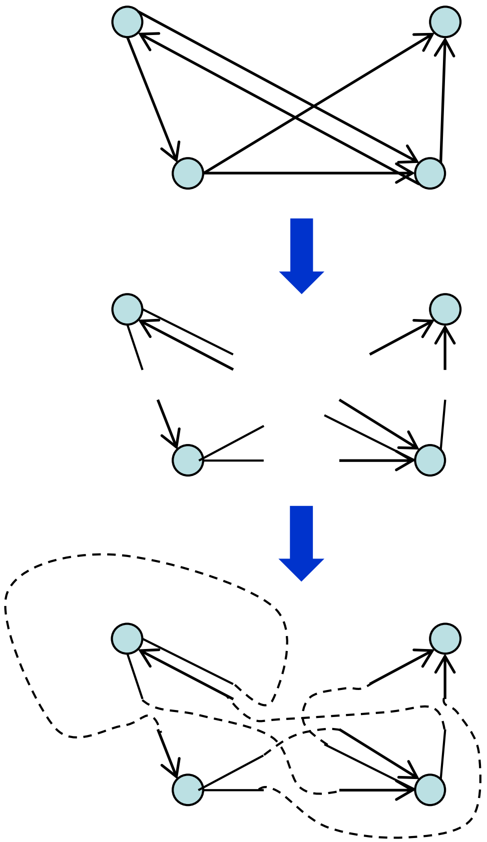
  1. Randomly select two links
  2. Swap its end nodes
     - (If this swap destroys some network property that should be conserved, cancel it)
  3. Repeat above many times



Randomlization method (2)

- Configuration model (Newman 2003)

  1. Cut every link into halves (heads and tails)
  2. Randomly connect head to tail
     - This conserves degree sequences
     - (Could result in multiple links and self-loops)



Other randomization methods

- Keeping only #'s of nodes and edges
- Degree sequence method
- Expected degree sequence method

Exercise

- Randomize connections in the Karate Club graph
- Measure the average clustering coefficient of the randomized network many times
- Test whether the average clustering coefficient of the original network is significantly non-random or not