Network Analysis

Dr. Shaina Race
Institute for Advanced Analytics

Spring 2018

Course Overview

Day 1: Visualization/Terminology basics

Day 2: Descriptive Statistics

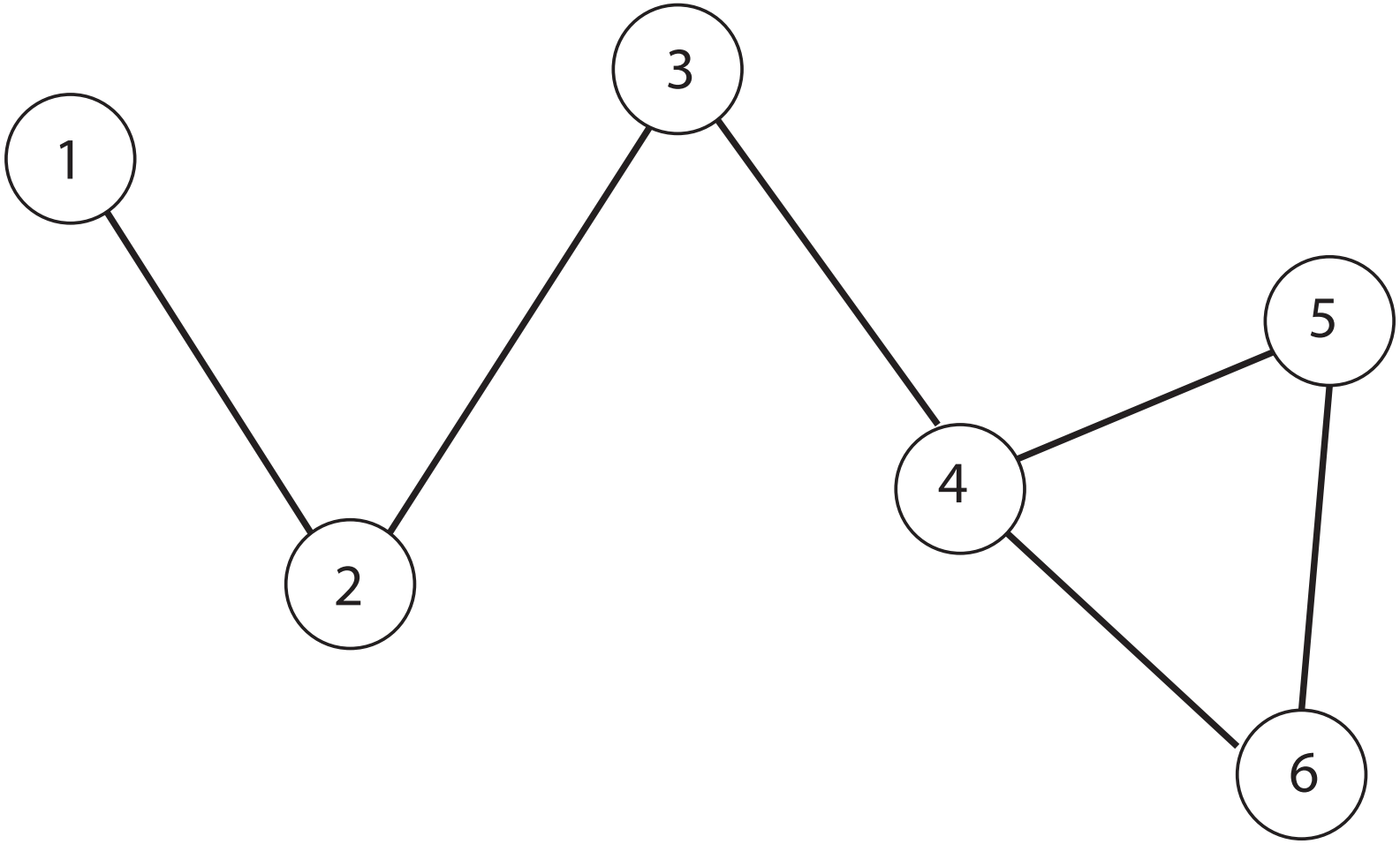
Day 3: Measures of Influence/Importance in a network

Day 4: Community Detection (clustering)

Day 5: Hypothesis Testing on networks

Day 6: Test or Project – up to you!

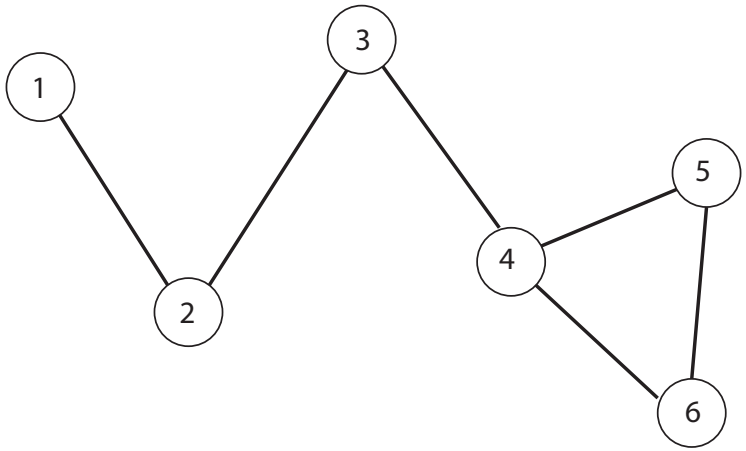
What is a graph/network?



What is a graph/network?

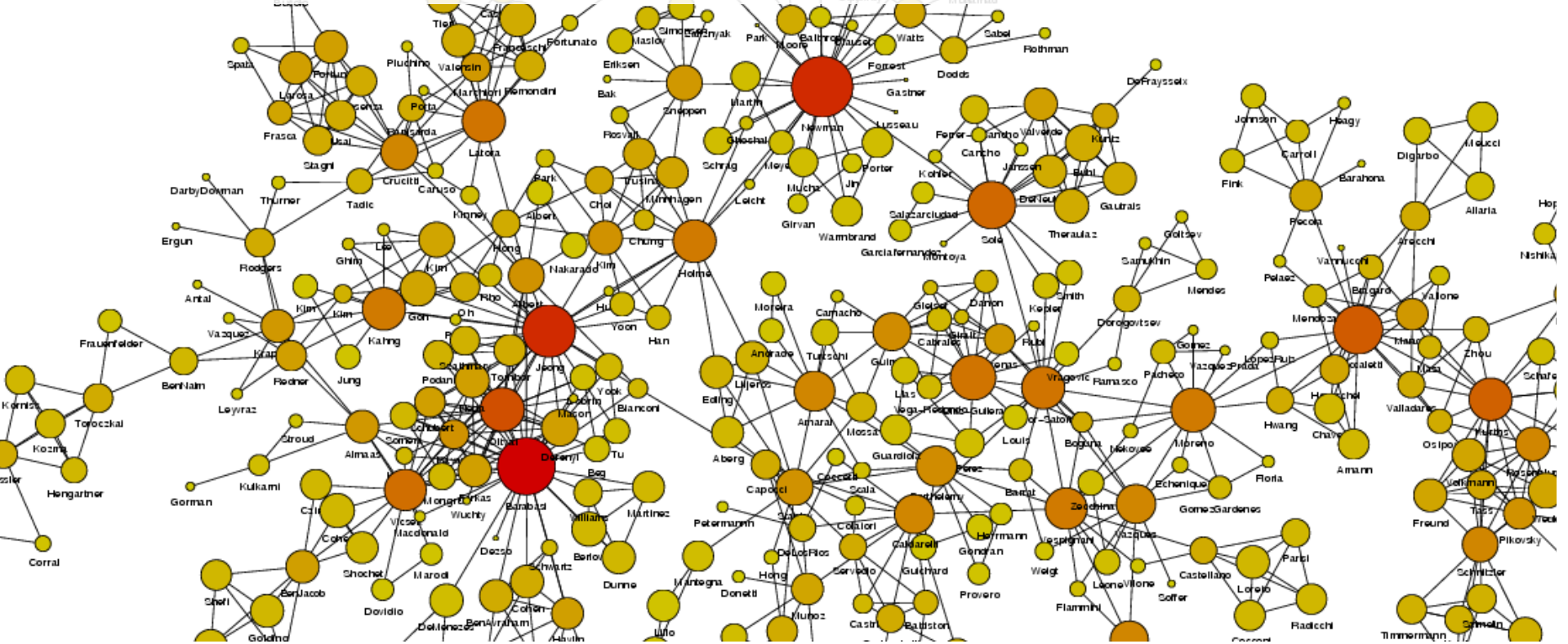
- Collection of **Entities**
  - i.e. **Nodes/Vertices**/Actors/Sites
  - could be individuals, organizations, places, objects

- A **structural variable** that measures some relationship between **dyads**/pairs
  - i.e. Edges/Arcs/Links/Ties/Relations
  - Friendship, distance, similarity, trade, management



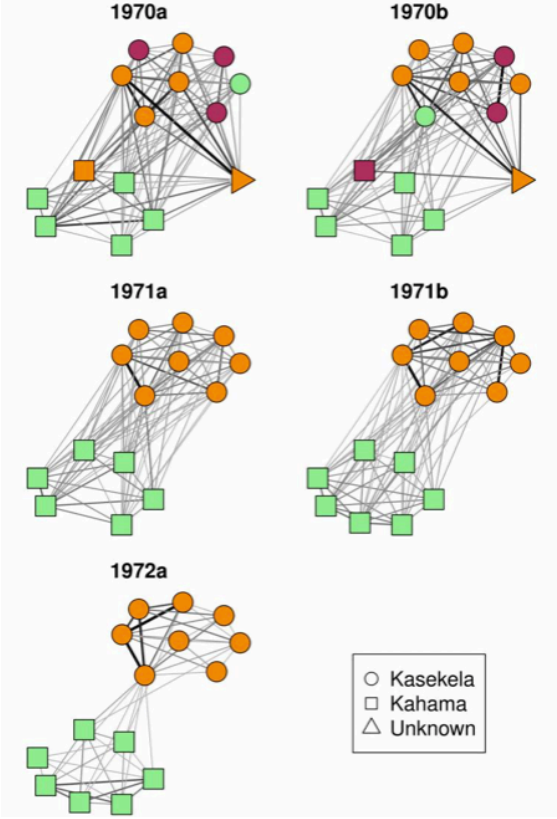
Questions this course might answer

Who is the **most important** or **influential** person in this network?



Questions this course might answer

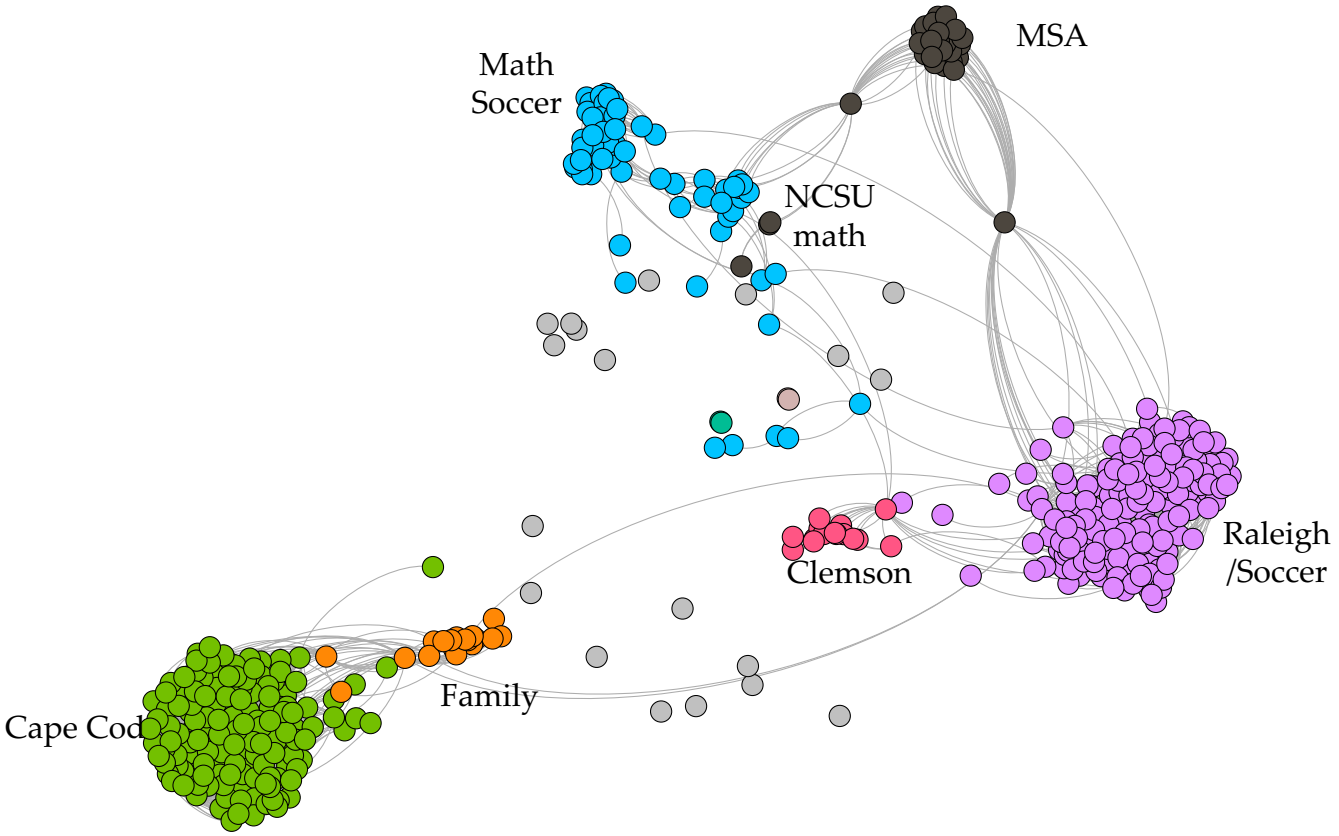
Is there any community structure or clustering apparent? Who is forming social groups? Who is NOT fitting in with the community structure?



Link:
https://today.duke.edu/2018/03/how-infighting-turns-toxic-chimpanzees

Questions this course might answer

What is driving the social atmosphere in our organization? What factors explain the network?



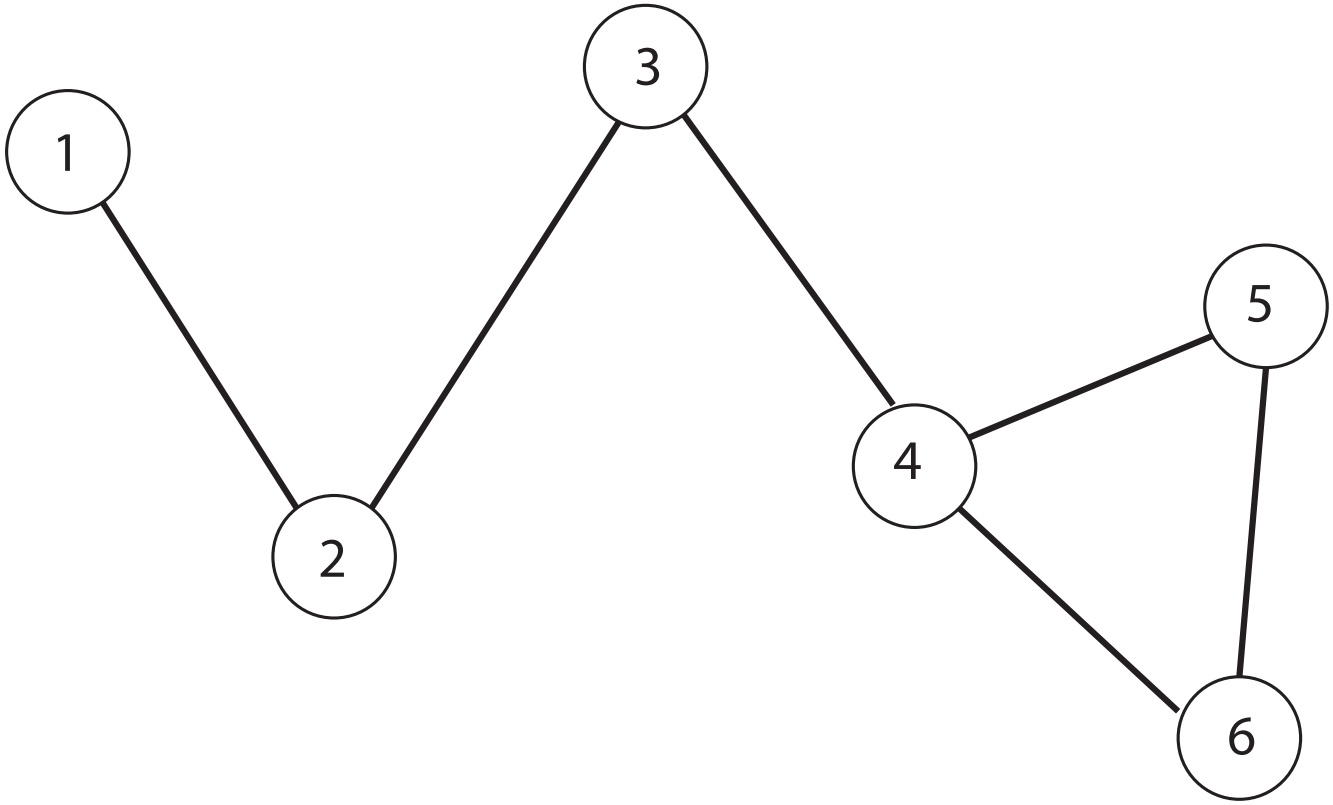
Introduction to Network Data

• • •

Graphs, Nodes, Edges

Graph Diagram

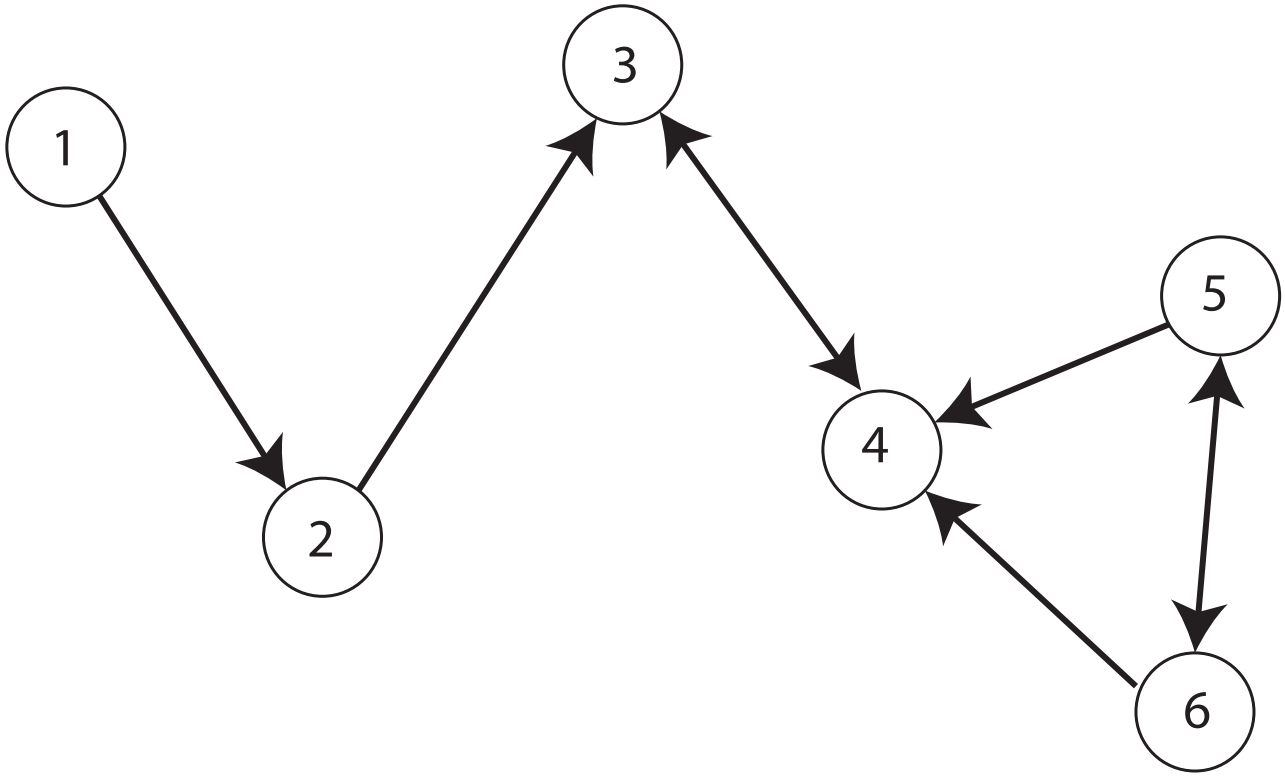
(Binary Graph)



Edges are binary variables indicating presence/absence of relationship

Graph Diagram

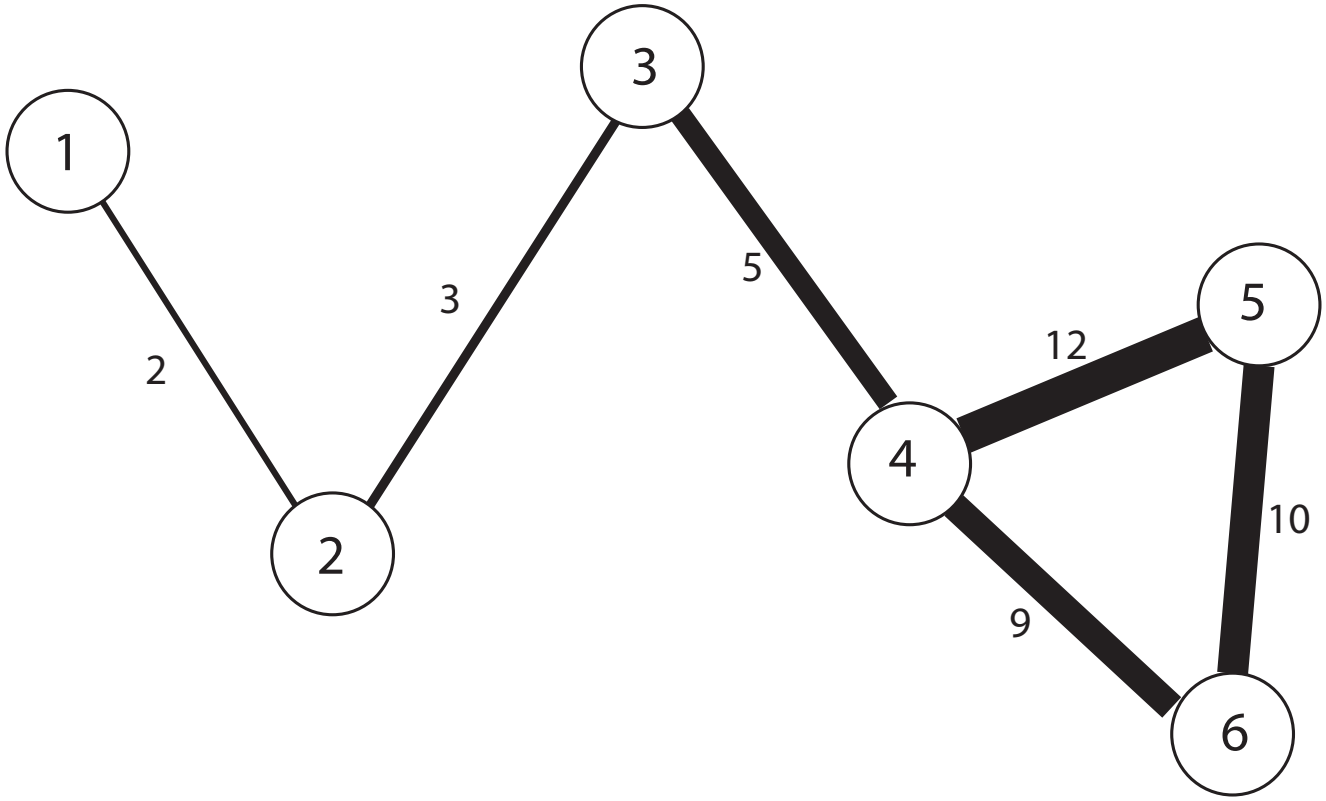
(Directed Binary Graph)



Edges are binary variables indicating presence/absence of directed relationship. (Ex: “is the son/daughter of…”)

Graph Diagram

(Weighted Graph)



Edges are ordinal or continuous variables indicating strength of relationship (Ex: "$ amount of trade between")

Examples of Networks

- Social
  - Personal affinity (friendship, respect, readership)
  - Interaction (phone records, email records, etc)
  - Organizational (managerial networks, organizational maps)
  - Affiliation (links to social events, clubs, or organizations)
  - Genealogy (family trees, kinship)
- Financial/Political
  - Trade networks
  - Business transactions, lending
- Transportation/Logistics
  - Manufacturing, warehousing, and retail networks
  - Highway and road systems

TopHat Quiz

Are the following networks weighted or unweighted?

A. A graph that links people together if they have the same birthday
B. A graph that links people together by the number of mutual friends they have

Is the following network directed or undirected?

A. A network that shows how universities transfer students to other universities

Statistical Considerations for Network data

Sampling, Independence

Sampling

- Suppose you're studying a massive network.
- How do you collect data to support your analysis?

Snowball sampling

- Data often collected via **snowball sampling**.
  - Select few individuals
  - Follow their ties/links to new individuals
  - Follow ties/links of new individuals
  - …Not so random…
- Advantages
  - Great for hidden populations
  - Or when trust needed to collect info
- Disadvantages
  - Bias! towards initial set of individuals.
    - ***Not necessarily representative of whole network distributions.***
  - Potential for inaccurate referrals

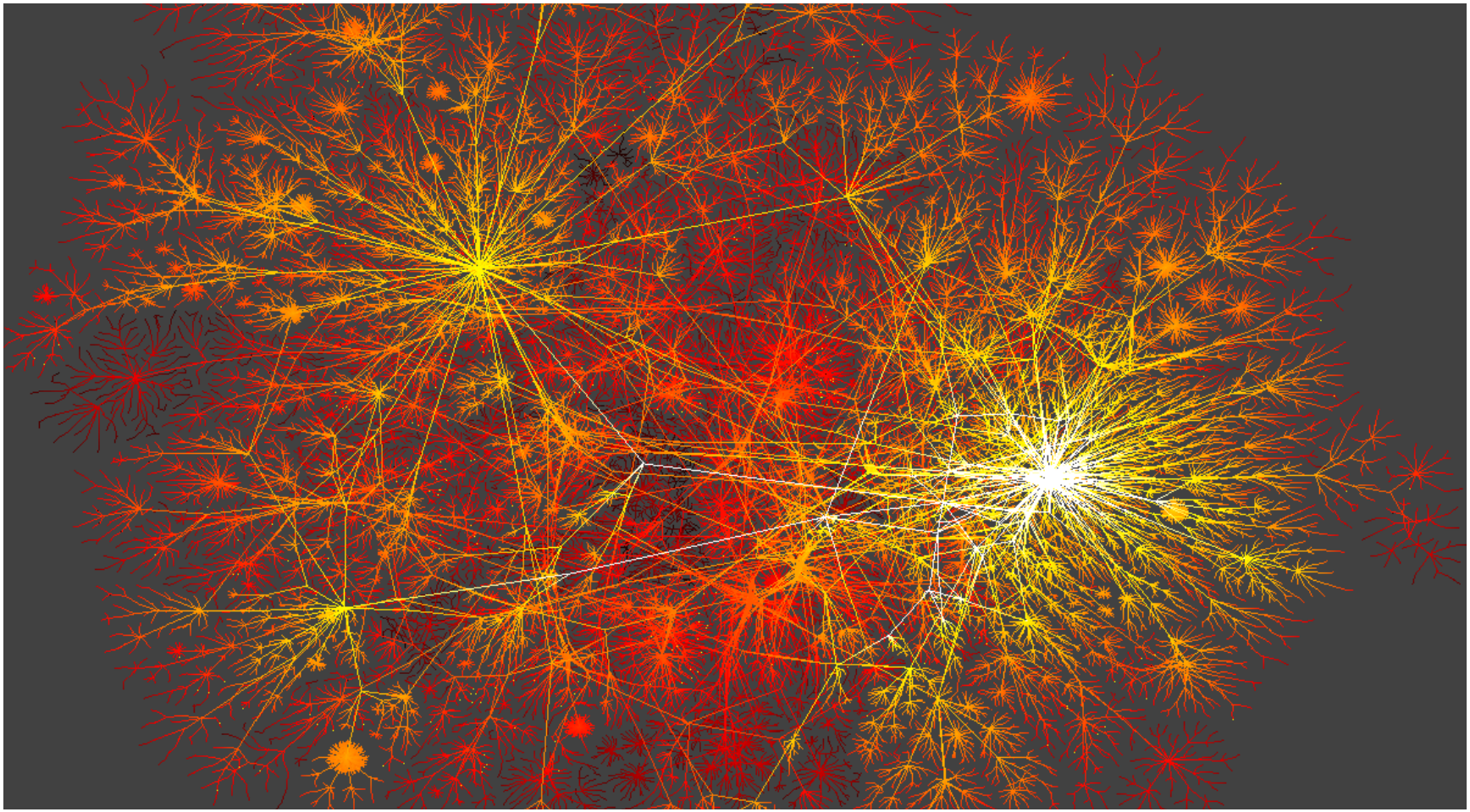
Independence in Networks

What independence??

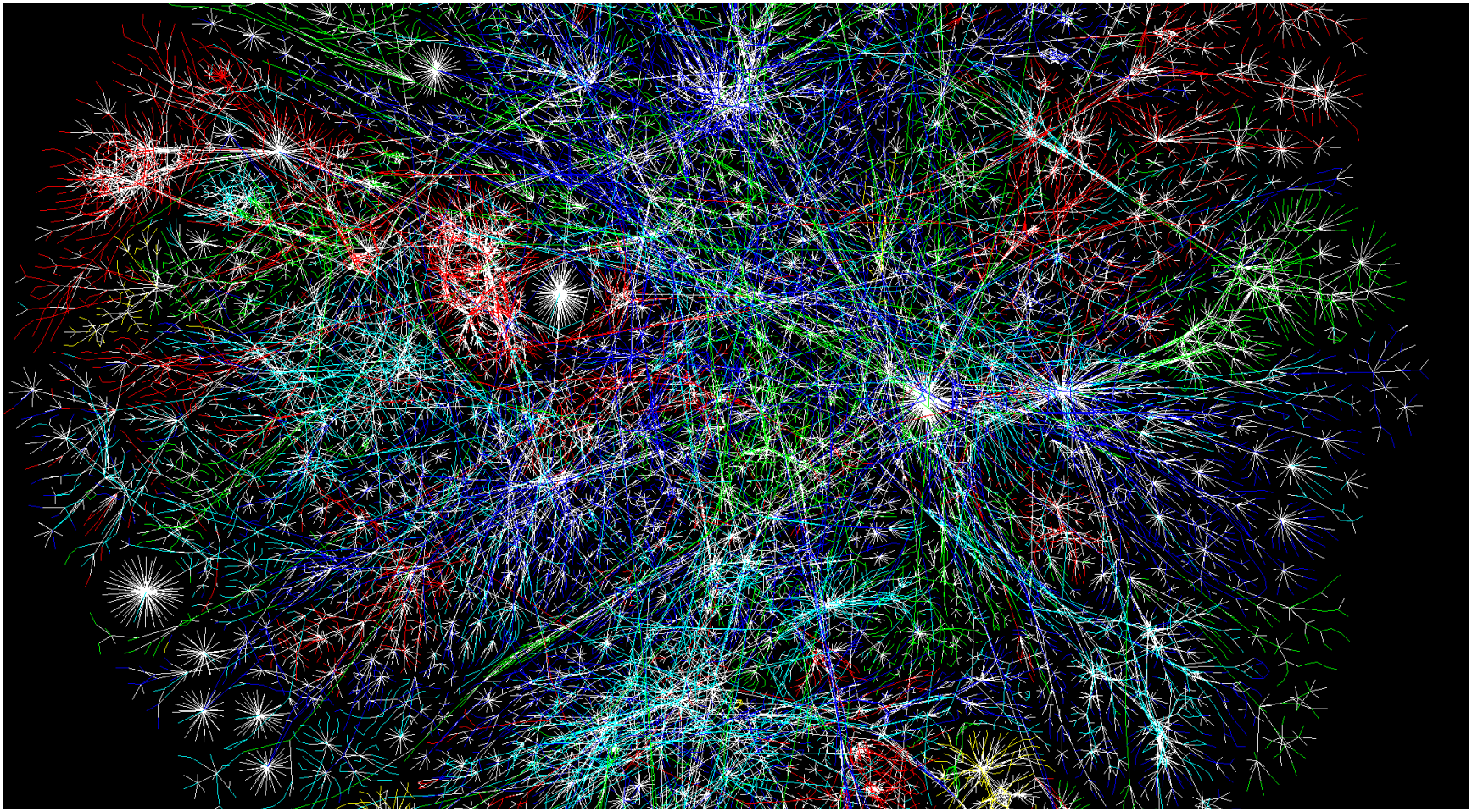
Network Visualization

One of the primary types of network analyses!

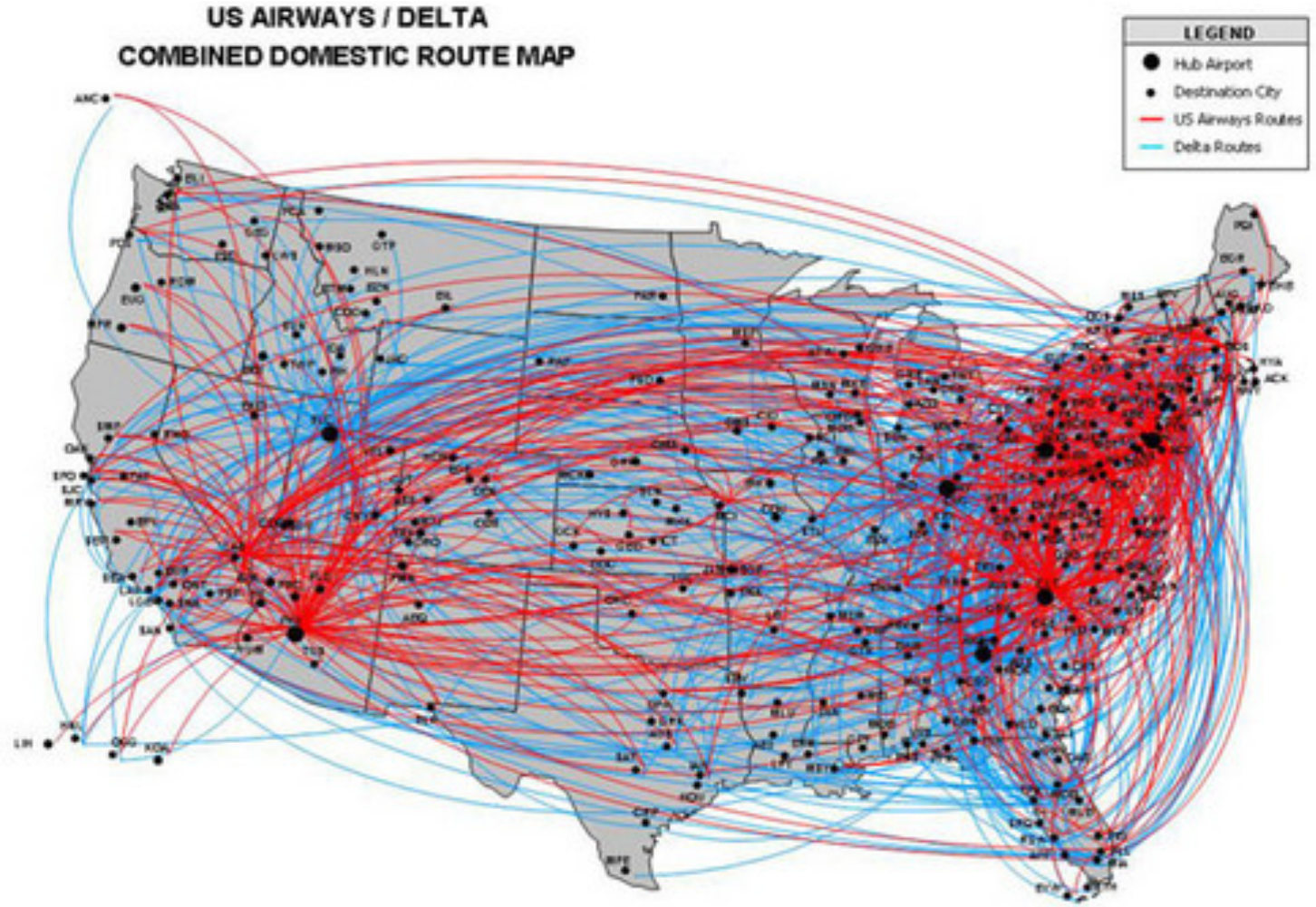
Sample of The Internet



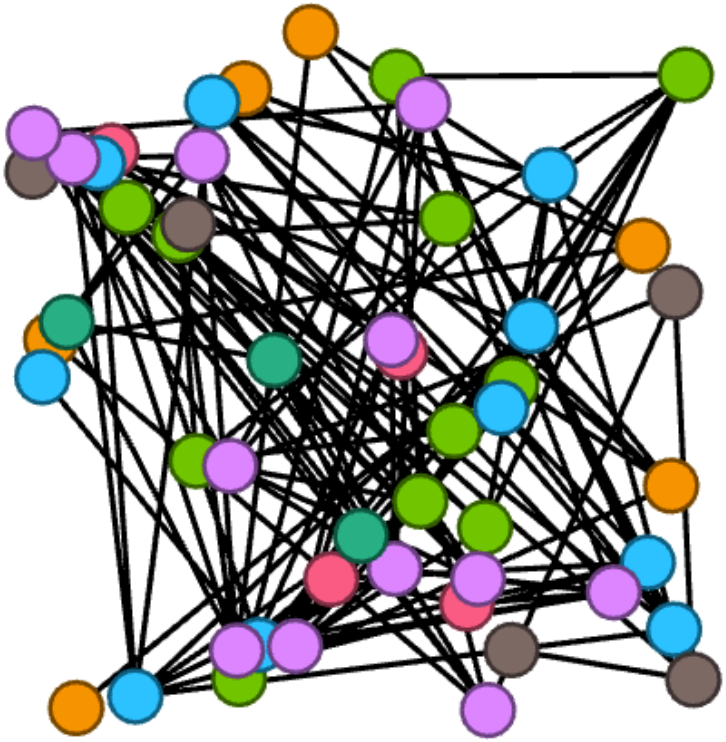
Sample of The Internet



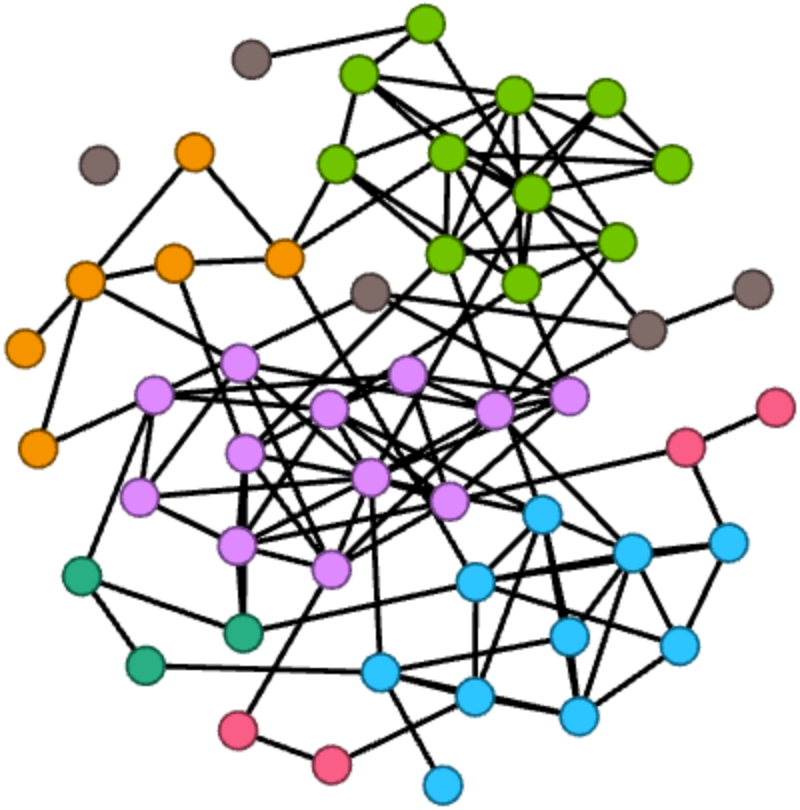
Air Transportation



Layout is important!

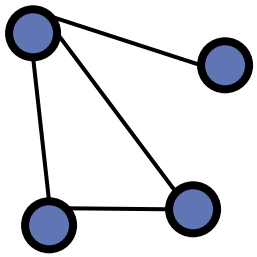


vs.

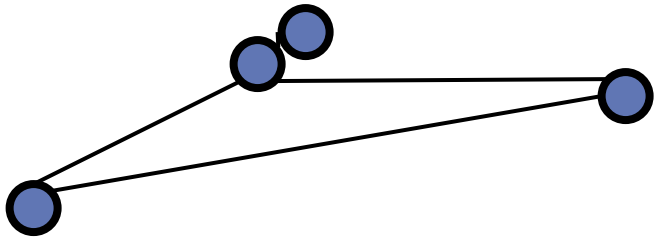


Considerations for Network Layout

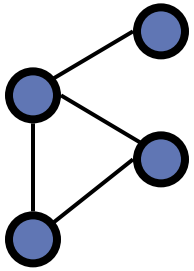
- Edge lengths



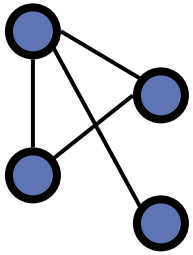
vs.



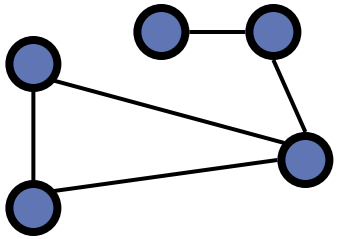
- Edge crossings



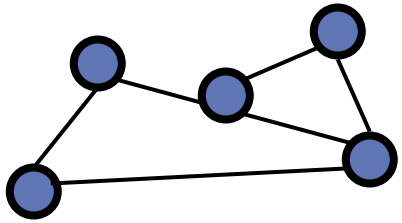
vs.



- Node-edge overlap



vs.



Force-Directed Graph Drawing

- Family of algorithms designed to layout graphs
- Some attractive force is placed on adjacent vertices while a repulsive force placed on all vertices. Then the *energy* of the system is minimized.
- Force Directed Slack Network

Force-Directed Graph Drawing

- Advantages
  - Quality
  - Versatility
  - Simplicity
- Disadvantages
  - ***Long*** Run-time ~ $O(n^3)$ (Problem for large graphs)
    - Standard algorithm: Hundreds-Thousand nodes
    - Barnes-Hut Simulation O(n log(n)): ~100K nodes
    - Multi-level approaches: ~ million of nodes
  - Some algorithms find poor local minima
  - Rarely do algorithms factor in the weights of edges

Other Graph Drawings

- **Multidimensional Scaling**
  - Class of methods to best represent pairwise distances in 2-space.
  - The distances between points in the resulting 2-dimensional approximation are as close as possible to the distances between points in the original space.
  - Gives (x,y) coordinates based on the SVD/PCA of a similarity matrix
  - For a network, the distance between 2 nodes is the length of shortest path between them.

Visualization Software

- Gephi
- R
  - igraph
  - network
  - networkD3
- Python
  - networkX

Data Formats

• • •

Edgelists, Adjacency Matrices, Lists of Neighbors

Graph Data Formats

- Many possible formats (.gml, .csv, .gephi, etc)
- Many include built-in ways to include individual variables (e.g. gender, age)
  - Node table
  - Node attributes
- All have some way to list edges (structural variables)
  - Edge list
  - List of neighbors
  - Adjacency Matrix

Edge List Format

- Lists the source and target of each edge. Numbering of nodes dependent on software. (python: 0…n) (R: 1…n package dependent)
- Some formats allow edge attributes (type/weight)

Source	Target	Weight
1	3	2
1	4	1
1	5	2
2	1	1
2	3	1
3	4	1
5	4	2

Neighbors List Format

- Best choice for large graphs
- 1 data line for each node, listing all of its neighbors:

```
1:  3,4,5
2:  1,3
3:  4
4:
5:  4
```

- Cannot easily incorporate edge attributes to this data structure.

Adjacency Matrix

Generally, all of the mathematical analysis is going to involve the **adjacency matrix**, **A**.

$\mathbf{A}_{ij}$ = weight of edge between vertex *i* and vertex *j*

TopHat Quiz

Which of the following words accurately describes the adjacency matrix for an undirected graph?

A. Diagonal
B. Symmetric
C. Identity
D. Negative
E. Rectangular
F. Sassy