Network Analysis

Advanced Statistical Topics B

Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk

Practicalities

- Every day from **8.15 to 15**
- R
- Additional software extensions
- Breaks/exercises as we go along
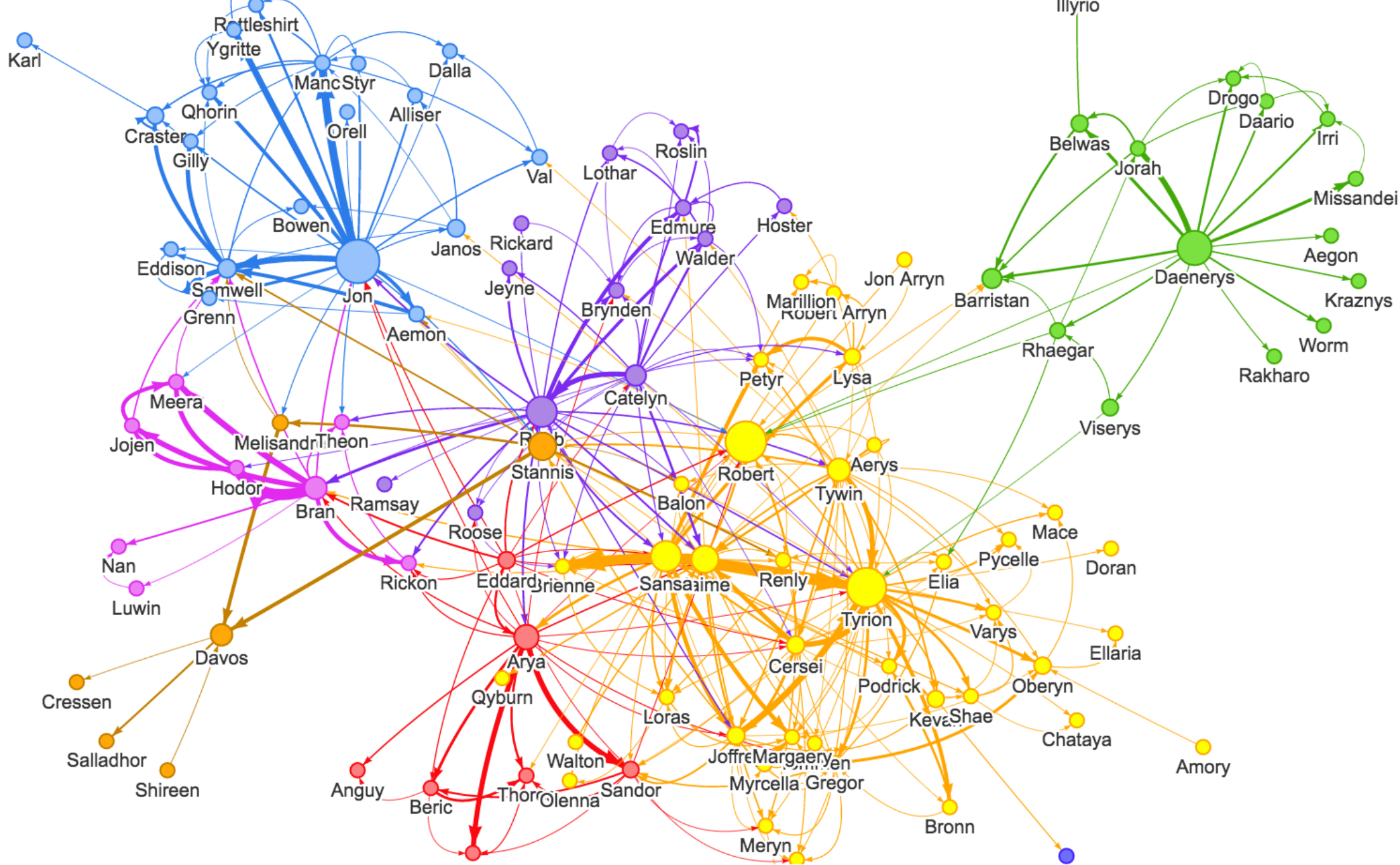- Lunch roughly at 11.
- Please "read" before classes

Sister course: Advanced Statistical Topics A

Course overview

1. Network analysis (Monday)
2. Bayesian Statistics (Tuesday)
3. Principal component analysis (PCA) and partial least squares (PLS) (Wednesday)
4. Neural networks and deep learning (Thursday)

web.stanford.edu/~hastie/CASI/
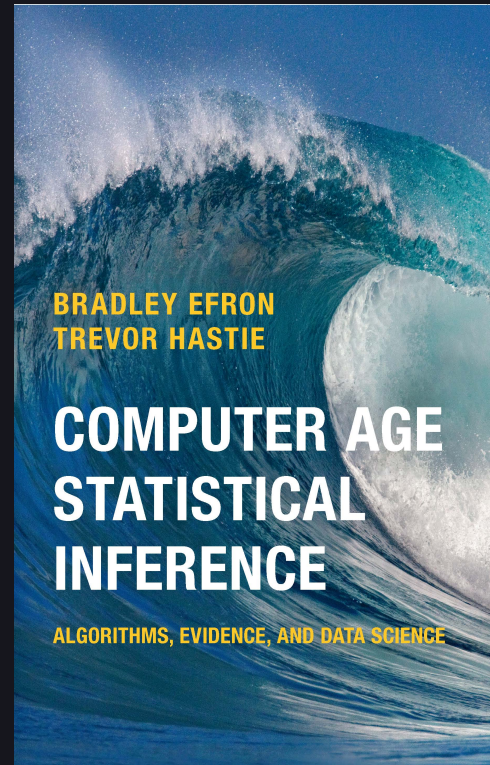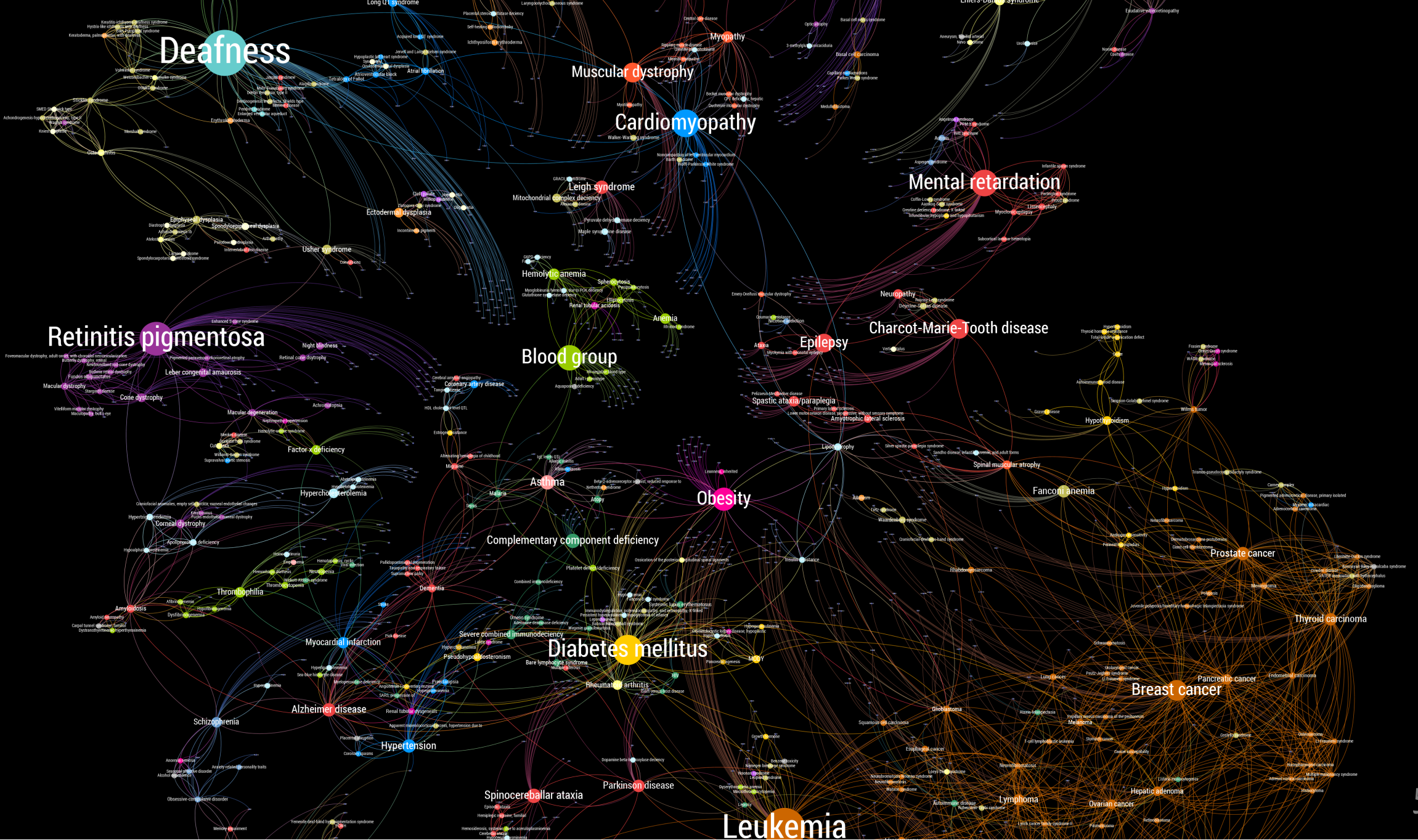
Statistical analysis of networks

Things become less clear on networks:

What is the sample size?

How do we measure features? What is a feature?

How would we determine if a network was "random" or follows a particular model?
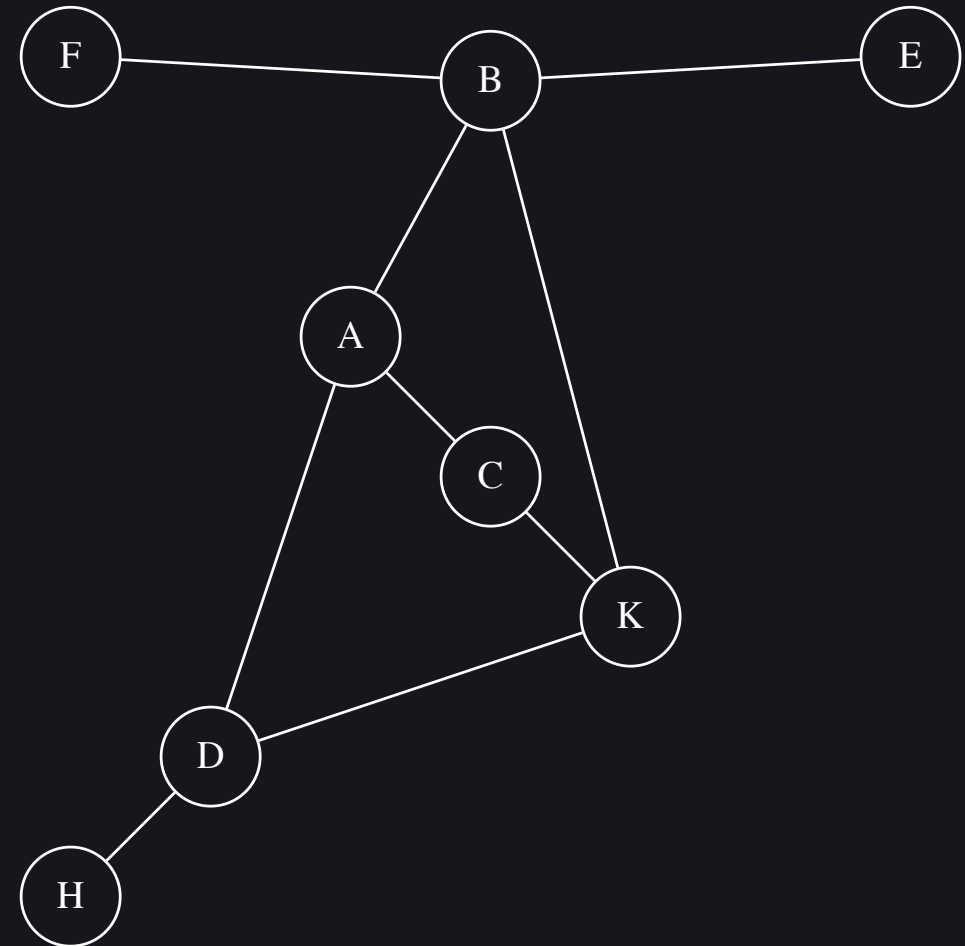
How do we define clusters / subgraphs / communities?

Graphs

Graphs consists of:

- **Vertices** or **nodes**, i.e., the numbers in the graph.
- **Edges** or **connections** i.e., the lines between the nodes.

Graphs can be *directed* or *undirected*, and can contain data on edge or node weights.

Overview

- Descriptive information about graphs

- Statistical models

- Fitting models on networks and statistical inference.

- Finding communities

Summarizing network graphs

Descriptive analysis; this is a standard first (and sometimes only!) step in characterising networks.

Typical measures include

- Density
- Centrality
- Closeness
- Betweenness

Graph density

The graph density of a graph $G = (V, E)$ with v nodes and e edges is the ratio of edges to the number of potential edges:

$$\text{density} = \frac{e}{v(v-1)/2}$$

A graph can be represented by its adjacency matrix

$$A_{s,t} = \begin{cases} 1 & \text{if } (s,t) \in E \\ 0 & \text{otherwise} \end{cases}$$

Centrality measures

Find important/central nodes in a graph. However, centrality is not uniquely defined.

The *degree* of a node v is its number of edges (arrows going in).

The *average degree of network* is the average of all node degrees.

The *degree distribution* is the relative frequency of all degrees in the network.

The degree summaries give simple characteristics of the network but tell very little about its specific structure.

Clustering coefficients

The *local clustering coefficient* for node v is the proportion of neighbours that are neighbours themselves. Describes the local density of a graph.

$$LCC(v) = \frac{\sum_{s,t} a_{v,s} a_{v,t} a_{s,t}}{\sum_{s,t} a_{v,s} a_{v,t}} = \frac{2 \text{ links between neighbours}}{\text{degree}_v (\text{degree}_v - 1)}$$

(with 0/0 = 0)

The *average clustering coefficient* is the average of $LCC(v)$ over all nodes V.

0 = star, 1 = clique

Global clustering

The global clustering coefficient is defined as

$$GCC = \frac{\text{number of closed triplet}}{\text{number of all triplets}}$$

The LCC measures how locally dense the network is.

The GCC measures globally denseness.

Distance

The *distance* between u, v is the shortest path between them if any such exists. Not necessarily unique.

The average shortest path is

$$\bar{\ell} = \frac{1}{|V||V-1|} \sum_{u,v \in V, u \neq v} \ell(u, v)$$

Describes how globally connected a graph is.

Unconnected groups make the average shortest path undefined as the distance is infinity.

Connectedness

The *betweenness of an edge e* is the proportion of shortest paths between any two nodes that pass through edge e.

The *betweenness of a node v* is the proportion of shortest paths between any two nodes that pass through node v.

The *connectivity* of a graph is the smallest number of edges to remove that results in a disconnected graph

Network models

Network generating models

If we want to judge if a network summary is "unusual" then we need to specify how a random network (from some model) would look.

Randomness in a network can be due to construction of the network (when is an edge an edge?), selection/sampling, errors in data, dynamic changes (if two people are related today they may not be related tomorrow), …

Bernoulli (Erdös-Renyi) random graphs

The nodeset V is given with $|V| = N$.

An edge between two nodes is present with probability p independently of all other edges. The expected number of edges is

$$\frac{N(N-1)}{2}p$$

and the expected clustering coefficient is

$$p$$

In practice

In real networks we often see results from *small world phenomenon*:

- A higher clustering than anticipated by the Bernoulli RG
- A shorter average shortest path than expected
- More nodes with higher degrees than expected (heavy tails)

The Watts-Strogatz

Arrange the $|V|$ nodes on a circle. Hard-wire each node to its k nearest neighbours on each side (k small).

Introduce random shortcuts between nodes which are not hard-wired. Chosen randomly all with same probability.

Average shortest distance is of order $|V|$. When shortcuts are introduced then the ASD is $\log(|V|)$.

Spread of epidemic.

Scale-free random graphs

Based on empirical observations by Barabasi and Albert. Resembles a power-law

$$P(\text{random node has degree } = k) \approx C \times k^{-\gamma}$$

"Rich gets richer"-model. Cannot (directly) produce any triangles.

The stochastic block model

Consider that nodes each belong to one of L classes. Edges are constructed independently, such that the probability for an edge depends only on the combined types of the two connecting nodes.

Quite flexible, but requires many assumptions.

Your network model should depend ...

... on the application.

Like any statistical model it should be tailored to the problem at hand.

Base the model on the context.

Fitting network models and testing hypotheses

Bernoulli random graphs

In Bernoulli random graph with $|V|$ nodes an edge is present with probability p independently of all other edges.

$$\hat{p} = \frac{|E|}{\binom{|V|}{2}}$$

This approach also works for stochastic block models (when the types are known) or for Watts-Strogatz (when the number of neighbours are known).

Scale-free models

If $P(\text{random node has degree } = k) \approx C \times k^{-\gamma}$ then

$$\log(P(\text{random node has degree } = k)) \approx \log(C) - \gamma \times \log(k)$$

Plot log relative frequency of degree k against $\log(k)$

Alternatively (more stable),

$$\log(P(\text{random node has degree } \geq k)) \approx C' - (\gamma - 1) \times \log(k)$$

Simulation using MCMC approaches (e.g., Stoc. block)

Lives on graphs and "moves" consist of

- adding or deleting edges, or
- adding or reducing types.

Not clear when the Markov chain has reached its stationary distribution.

Providing rules to modify edges or change node types requires (prior) assumptions about the shape of model to use for this!

Statistical tests

Study the asymptotic distribution of summary statistics by comparing the observed summary to how it would look under an assumed model.

E.g., degree, clustering coefficient, average shortest distance, …

Except in Bernoulli random graphs, the theoretical distribution of summary statistics is usually not easy to derive.

Parametric bootstrap of model

Simulate samples under the null hypothesis (model) from a parametric model with estimated parameters.

If we have a complex model then we might condition on a summary and focus on another summary:

For example, draw many random networks with same degree sequence. Count the simulated datasets where the shortest average path is as extreme as our observed shortest average path.

Note: Dependence between summaries might prove a problem.

Parametric bootstrapping algorithm

1. Compute the measure of interest in actual network, T^*
2. Fit the parameters of the network (if possible)
3. Simulate from network model *or* simulate from conditional random graph given a summary and compute the measure of interest for the sampled graph, T^b
4. Do 3 B times
5. Count the simulated measures larger (or smaller) than T^*, m
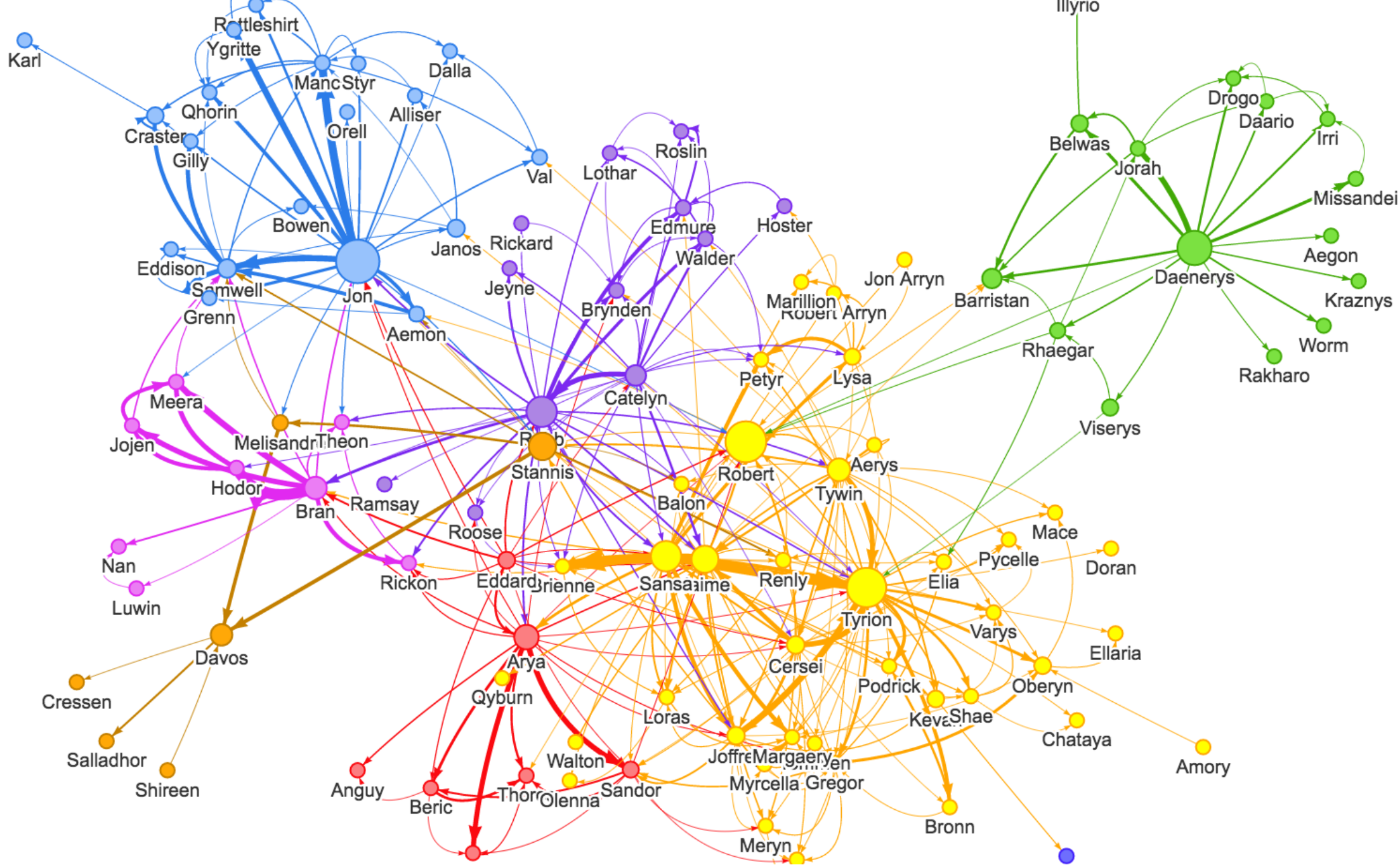6. Compute a p-value (check direction of statistic):

$$\frac{m}{B+1}$$

Communities

Community detection
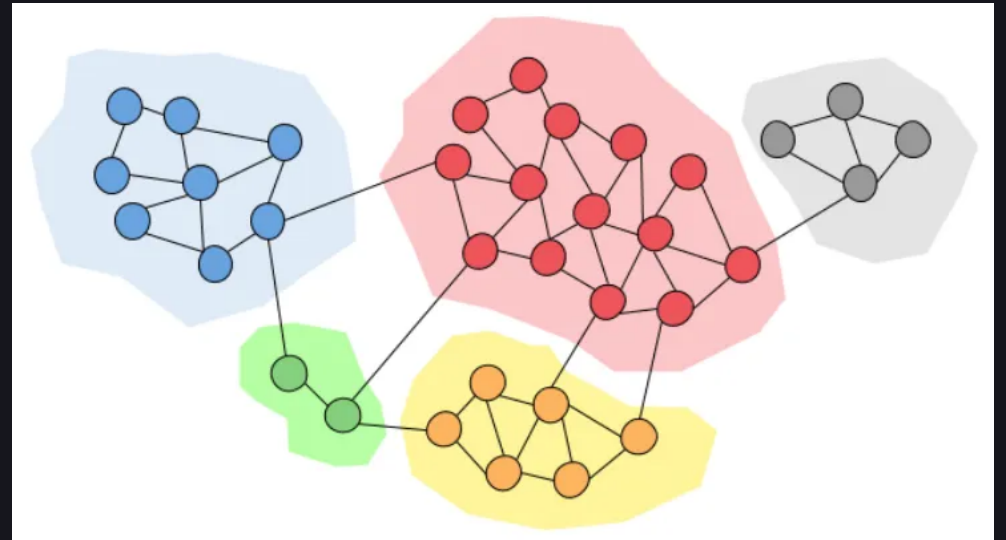
Essentially the same as standard clustering.

Abundance of community detection methods

BEWARE of clustering

The Girvan-Newman approach

1. Calculate betweenness of all edges in the network.
2. Remove the edge with highest betweenness. In case of ties either choose one at random or remove all
3. Repeat until no edges remain

How many communities?

The modularity is defined as

$$Q = \frac{1}{2|E|} \sum_{i,j} (A_{ij} - \frac{\deg(i)\deg(j)}{2|E|})I(c_i = c_j)$$

where A is the adjacency matrix, c_i is the community of node i.

$Q = 0$ indicates the community is no stronger than expected by random shuffling since $\frac{\deg(i)\deg(j)}{2|E|}$ is roughly the probability that there is an edge between i and j.

Communities for the stochastic block model

Nodes with same type can be considered to be of the same community.

Different types can be aggregated into larger communities.

If the *number* of types is known but not the exact class for each node then partial or exact recovery might be possible.

(Classification problem - requires that there is detectable differences in the proportion of groups *and* the edge probabilities)

Potential problems (with no easy solution)

If there is uncertainty in the determination of edge status (i.e., presence/absence) then that uncertainty propagates through the any calculations on the network.

Very little work addressing this

- Characterization of propagation of errors from networks to summaries
- Adjusting for errors (improved estimators)

Sampling from (very large) networks

What is the sample size?

When we sample from a network the sampled network might not be of the same type as the original network.

How to sample:

- Random sample of nodes and their corresponding edges.
- Snowball sampling. Sample 1 node (with edges), follow edges to neighbours, …