Advanced Statistical Topics B

Network analysis

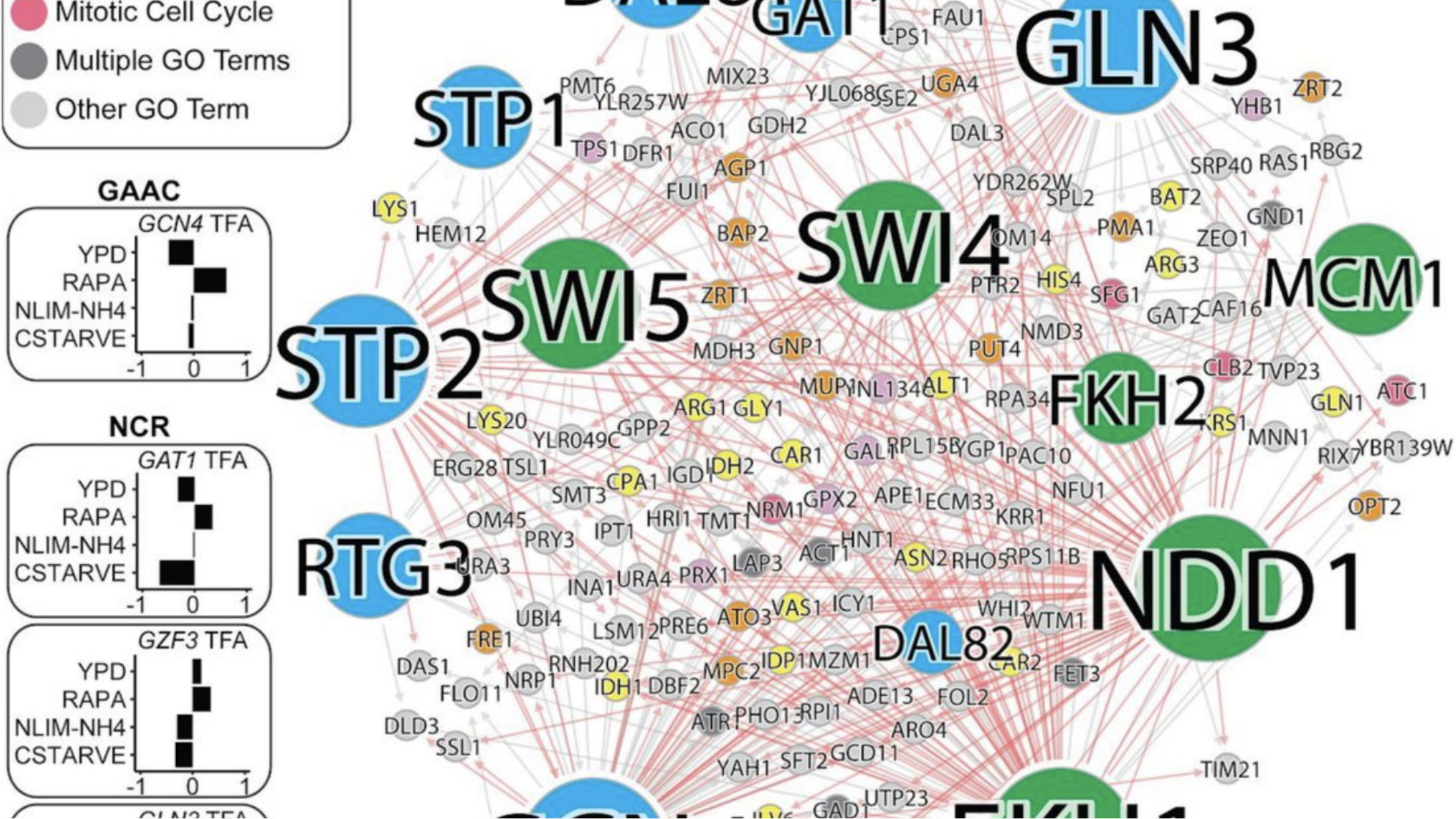
Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk

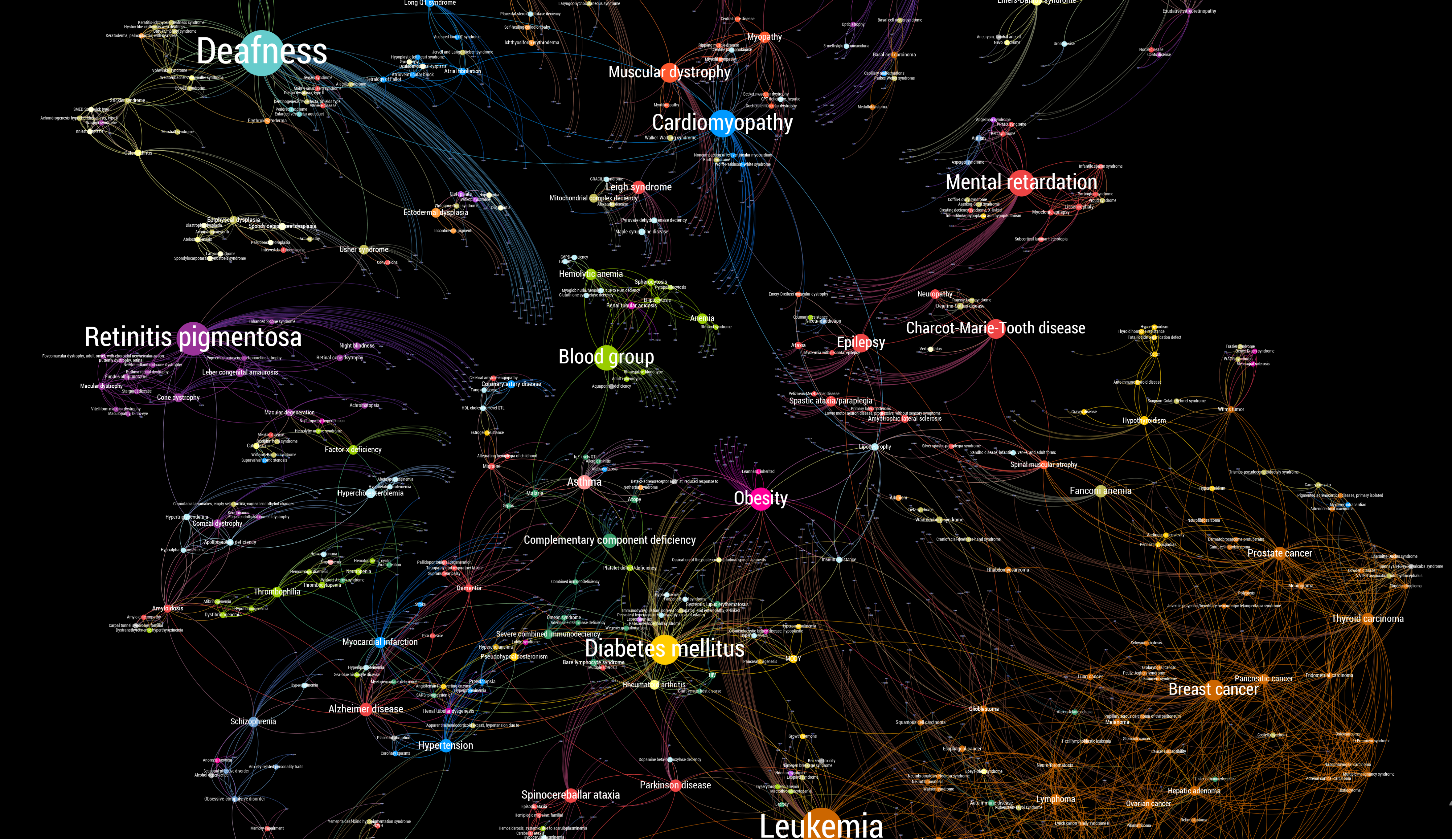
Slides @
biostatistics.dk/teaching/advtopicsB/

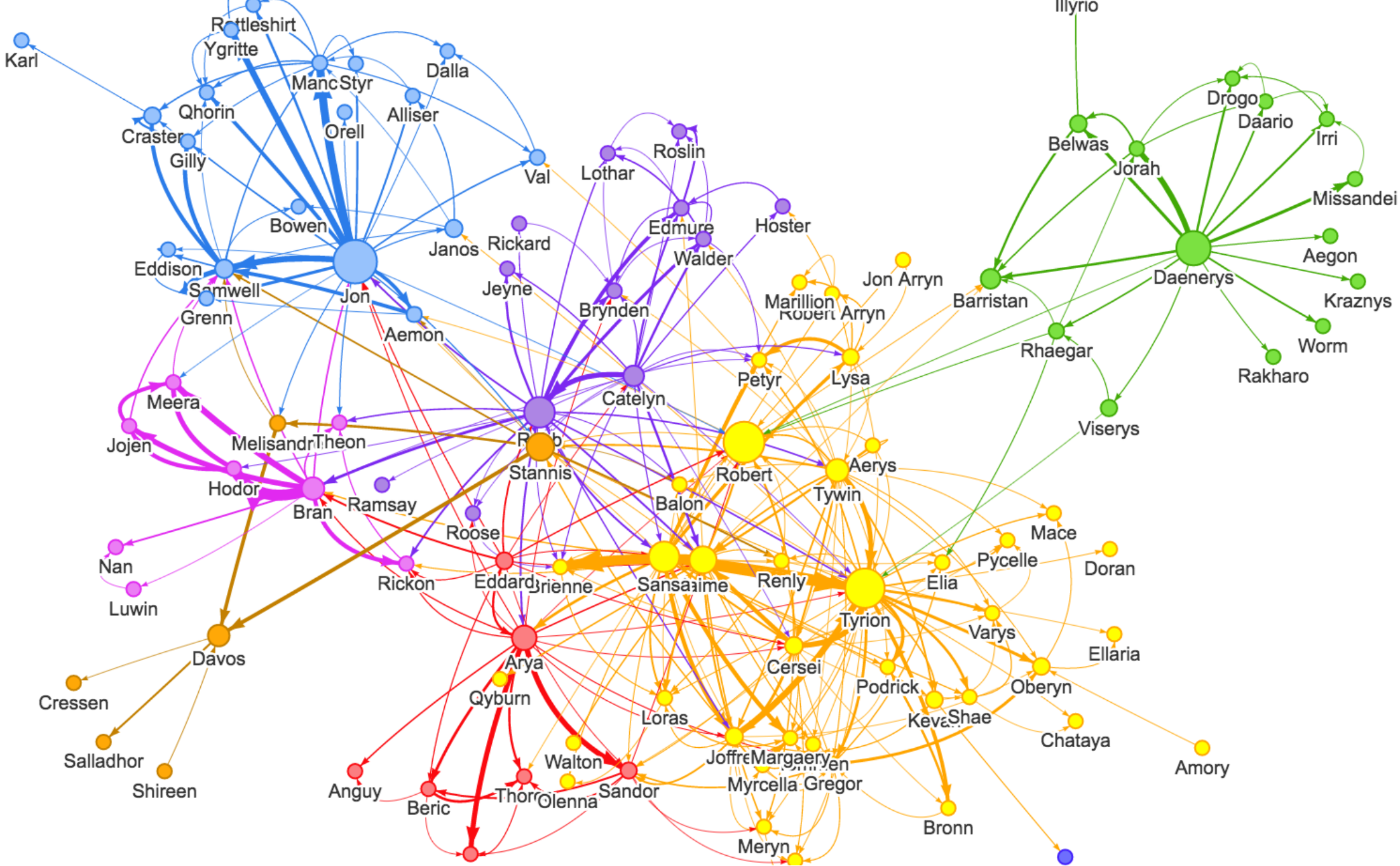
Network analysis

Large-scale connectivity data are available in many fields:

- Traffic and transportation
- Gene-gene-networks
- Social networks
- Websites
- Text mining in patient records
- Bibliometrics
- Spread of epidemic


Mitotic Cell Cycle
Multiple GO Terms
Other GO Term
GAAC
GCN4 TFA
YPD
RAPA
NLIM-NH4
CSTARVE
-1
0
1
NCR
GAT1 TFA
GZF3 TFA
GAT1
GLN3
STP1
SWI4
SWI5
MCM1
STP2
FKH2
RTG3
NDD1
DAL82
FAU1
CPS1
MIX23
PMT6
YLR257W
YJL068C
SSE2
UGA4
ZRT2
YHB1
ACO1
GDH2
DAL3
TPS1
DFR1
AGP1
FUI1
SRP40
RAS1
RBG2
YDR262W
SPL2
BAT2
GND1
LYS1
HEM12
BAP2
OM14
PMA1
ZEO1
ARG3
PTR2
HIS4
SFG1
GAT2
CAF16
ZRT1
NMD3
MDH3
GNP1
PUT4
CLB2
TVP23
MUP1
INL134C
ALT1
RPA34
ATC1
GLN1
ARG1
GLY1
LYS20
YLR049C
GPP2
KRS1
MNN1
RIX7
YBR139W
ERG28
TSL1
CAR1
GAL1
RPL15B
YGP1
PAC10
IDH2
IGD1
CPA1
SMT3
NFU1
OPT2
OM45
PRY3
IPT1
HRI1
TMT1
NRM1
GPX2
APE1
ECM33
KRR1
HNT1
ACT1
ASN2
RHO5
RPS11B
URA3
INA1
URA4
PRX1
LAP3
UBI4
VAS1
ICY1
WHI2
WTM1
FRE1
LSM12
PRE6
ATO3
DAS1
RNH202
IDP1
MZM1
CAR2
FET3
NRP1
IDH1
DBF2
MPC2
FLO11
ADE13
FOL2
DLD3
ATR1
PHO13
RPI1
ARO4
SSL1
YAH1
SFT2
GCD11
TIM21
UTP23
GAD1


Deafness
Muscular dystrophy
Myopathy
Cardiomyopathy
Mental retardation
Leigh syndrome
Mitochondrial complex deficiency
Ectodermal dysplasia
Usher syndrome
Epiphyseal dysplasia
Spondyloepiphyseal dysplasia
Hemolytic anemia
Spherocytosis
Renal tubular acidosis
Anemia
Neuropathy
Charcot-Marie-Tooth disease
Epilepsy
Ataxia
Retinitis pigmentosa
Night blindness
Leber congenital amaurosis
Macular dystrophy
Cone dystrophy
Blood group
Coronary artery disease
Spastic ataxia/paraplegia
Amyotrophic lateral sclerosis
Macular degeneration
Factor x deficiency
Hypothyroidism
Wilms tumor
Spinal muscular atrophy
Lipodystrophy
Asthma
Obesity
Fanconi anemia
Hypercholesterolemia
Corneal dystrophy
Malaria
Complementary component deficiency
Prostate cancer
Thrombophilia
Dementia
Amyloidosis
Myocardial infarction
Severe combined immunodeficiency
Diabetes mellitus
Thyroid carcinoma
Pseudohypoaldosteronism
Rheumatoid arthritis
Breast cancer
Pancreatic cancer
Alzheimer disease
Schizophrenia
Hypertension
Spinocereballar ataxia
Parkinson disease
Lymphoma
Hepatic adenoma
Ovarian cancer
Leukemia


Illyrio
Karl
Rattleshirt
Ygritte
Manc
Styr
Dalla
Qhorin
Alliser
Craster
Orell
Gilly
Val
Lothar
Roslin
Belwas
Jorah
Drogo
Daario
Irri
Missandei
Bowen
Edmure
Hoster
Rickard
Janos
Walder
Eddison
Samwell
Jon
Jeyne
Brynden
Jon Arryn
Marillion
Robert Arryn
Barristan
Daenerys
Aegon
Kraznys
Grenn
Aemon
Rhaegar
Worm
Rakharo
Petyr
Lysa
Catelyn
Meera
Viserys
Jojen
Melisandr
Theon
Stannis
Robert
Aerys
Tywin
Hodor
Bran
Ramsay
Balon
Roose
Mace
Pycelle
Doran
Nan
Rickon
Eddard
Brienne
Sansa
Jaime
Renly
Elia
Luwin
Tyrion
Varys
Davos
Arya
Cersei
Ellaria
Qyburn
Podrick
Oberyn
Loras
Chataya
Cressen
Walton
Joffre
Margaery
Amory
Salladhor
Anguy
Sandor
Myrcella
Gregor
Shireen
Beric
Thor
Olenna
Bronn
Meryn

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?

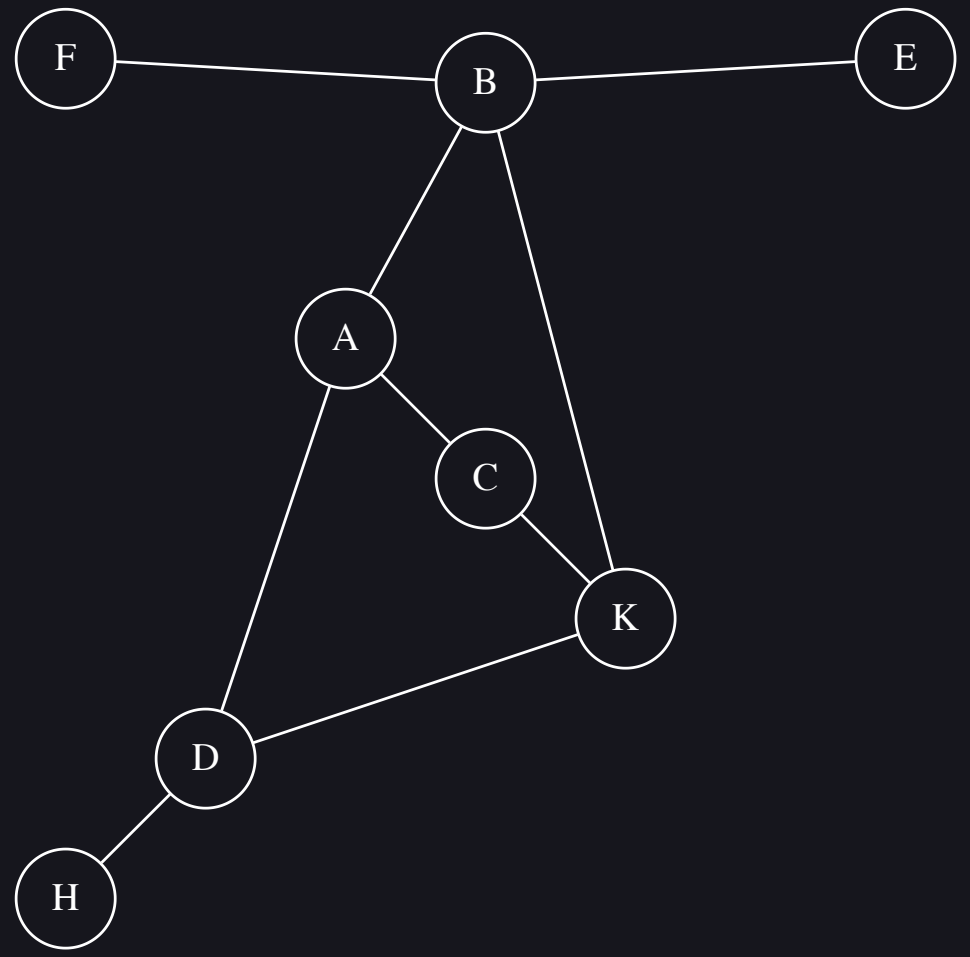
How do we define clusters / subgraphs / communities?

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 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 and takes the "zeros" in the adjacency matrix into account.

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.

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 fully 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):

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

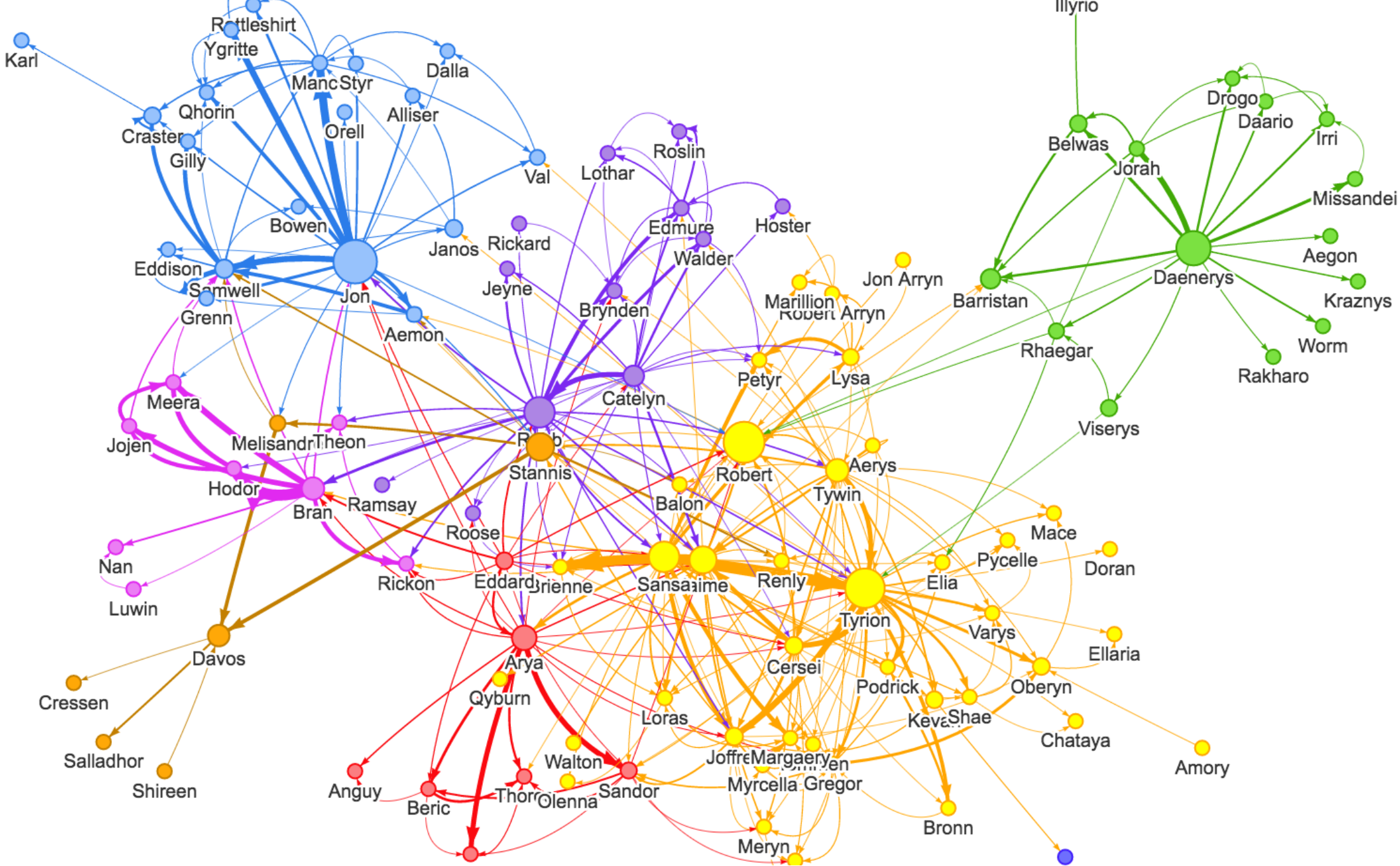
Communities

Community detection

Essentially the same as standard clustering.

Abundance of community detection methods

BEWARE of clustering


Illyrio
Karl
Rattleshirt
Ygritte
Manc
Styr
Dalla
Qhorin
Alliser
Craster
Orell
Gilly
Val
Lothar
Roslin
Belwas
Jorah
Drogo
Daario
Irri
Missandei
Bowen
Edmure
Hoster
Rickard
Janos
Walder
Eddison
Samwell
Jon
Jeyne
Brynden
Jon Arryn
Marillion
Robert Arryn
Barristan
Daenerys
Aegon
Kraznys
Grenn
Aemon
Rhaegar
Worm
Rakharo
Petyr
Lysa
Catelyn
Meera
Viserys
Jojen
Melisandr
Theon
Stannis
Robert
Aerys
Tywin
Hodor
Bran
Ramsay
Balon
Roose
Mace
Pycelle
Doran
Nan
Rickon
Eddard
Brienne
Sansa
Jaime
Renly
Elia
Luwin
Tyrion
Varys
Davos
Arya
Cersei
Ellaria
Qyburn
Podrick
Oberyn
Loras
Chataya
Cressen
Walton
Joffre
Margaery
Amory
Salladhor
Anguy
Sandor
Myrcella
Gregor
Shireen
Beric
Thor
Olenna
Bronn
Meryn

The Newman-Girman 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, ...