Social Network Analysis for Computer Scientists

Frank Takes

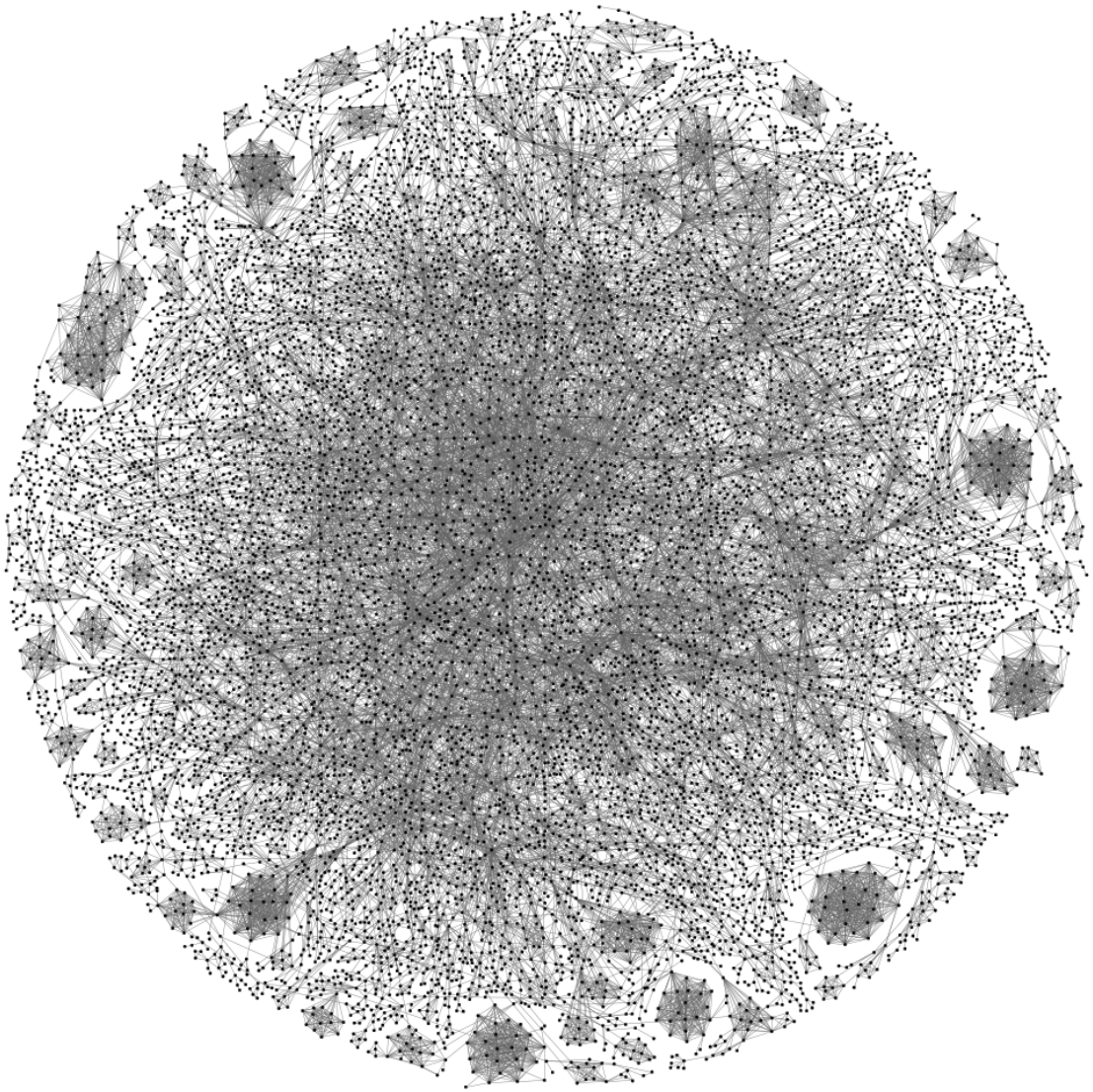
LIACS, Leiden University

https://liacs.leidenuniv.nl/~takesfw/SNACS

Lecture 2 — Advanced network concepts and centrality

Recap

Networks



Notation

Concept	**Symbol**
Network (graph)	$G = (V, E)$
Nodes (objects, vertices, ...)	V
Links (ties, relationships, ...)	E
Directed — $E \subseteq V \times V$ — "links"	
Undirected — "edges"	
Number of nodes — $\|V\|$	n
Number of edges — $\|E\|$	m
Degree of node u	$deg(u)$
Distance from node u to v	$d(u, v)$

Real-world networks

1. Sparse networks — density
2. Fat-tailed power-law degree distribution — degree
3. Giant component — components
4. Low pairwise node-to-node distances — distance
5. Many triangles — clustering coefficient

- Many examples: communication networks, citation networks, collaboration networks (Erdös, Kevin Bacon), protein interaction networks, information networks (Wikipedia), webgraphs, financial networks (Bitcoin) ...

Advanced concepts

Advanced concepts

- Assortativity
- Reciprocity
- Power law exponent
- Planar graphs
- Complete graphs
- Subgraphs
- Trees
- Spanning trees
- Diameter
- Bridges
- Graph traversal

Assortativity

- **Assortativity**: extent to which "similar" nodes attract each other
  Value close to -1 if dissimilar nodes more often attract each other
  Value close to 1 if similar nodes more often attract each other

Assortativity

- **Assortativity**: extent to which "similar" nodes attract each other
  Value close to -1 if dissimilar nodes more often attract each other
  Value close to 1 if similar nodes more often attract each other
- Degree assortativity: nodes with a similar degree connect more frequently
- Attribute assortativity: nodes with similar attributes attract each other
- Influence on connectivity of network, spreading of information, etc.
- Social networks: **homophily**
- Complex networks: **mixing patterns**

Degree assortativity

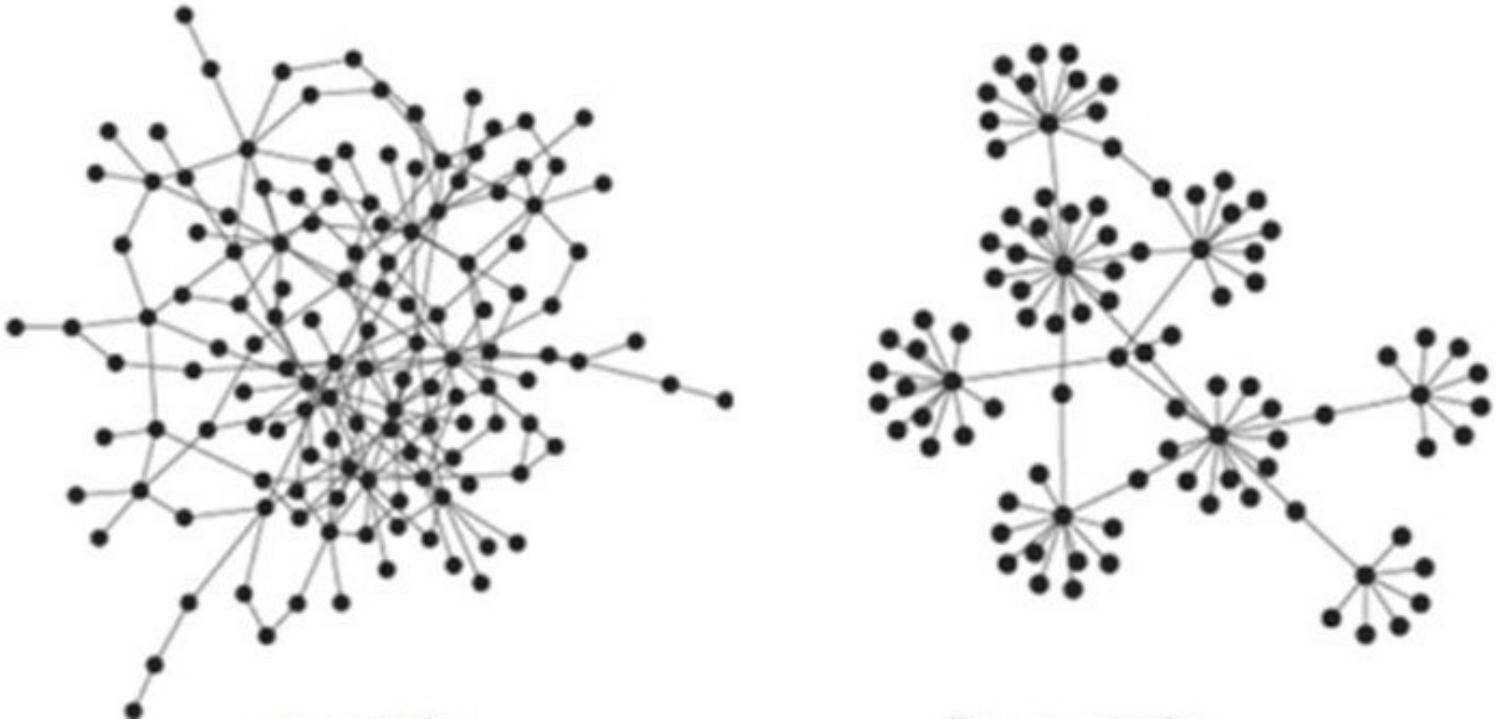


Figure: Degree **assortativity** (left) and degree **disassortativity** (right)

Image: Estrada et al., Clumpiness mixing in complex networks, J. Stat. Mech. Theor. Exp. P03008 (2008).

Attribute assortativity

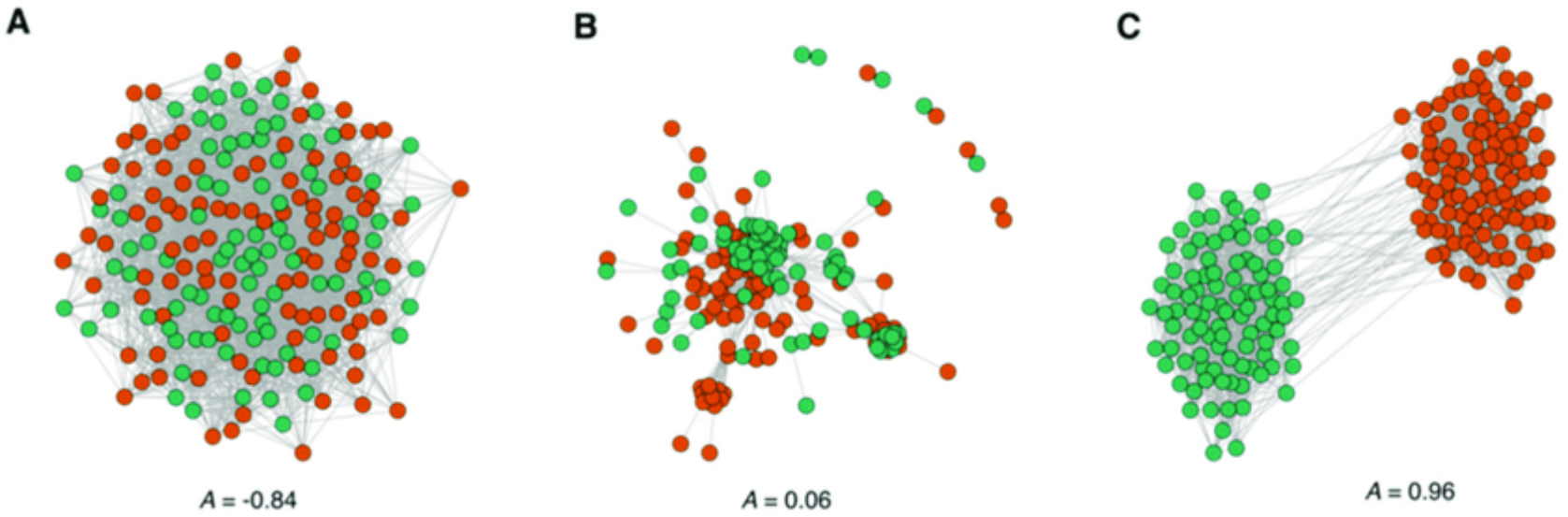


Figure: Attribute assortativity

Image: Moya-García, A. et al. Identification of New Toxicity Mechanisms ... *Genes*, 13(7), 1292, 2022.

Reciprocity

- **Reciprocity**: measure of the likelihood of nodes in a directed network to be mutually linked
- Let $m_{<->}$ be the number of links in the directed network for which there also exists a symmetric counterpart:

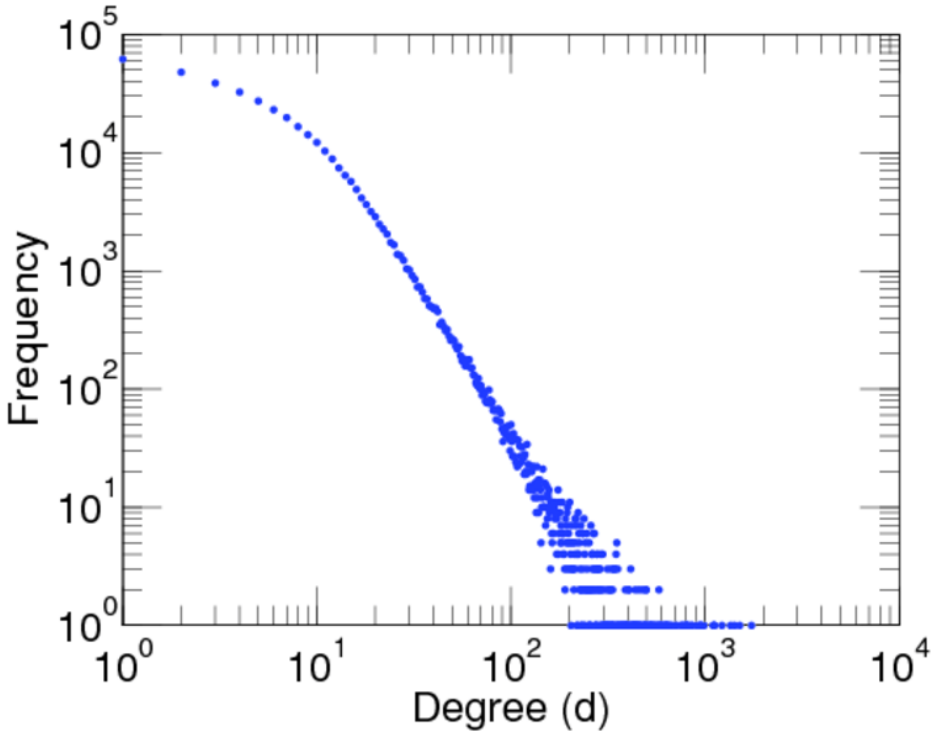
$$m_{<->} = |\{(u, v) \in E : (v, u) \in E\}|$$

- Reciprocity r is then the fraction of links that is symmetric:

$$r = \frac{m_{<->}}{m}$$

- Measures the extent to which relationships are mutual
- Useful to compare between networks

Power law degree distribution



Source: `http://konect.cc/networks/citeseer/`

Power law exponent in undirected networks

- The probability p_k of a node having degree k depends on the power law exponent γ:

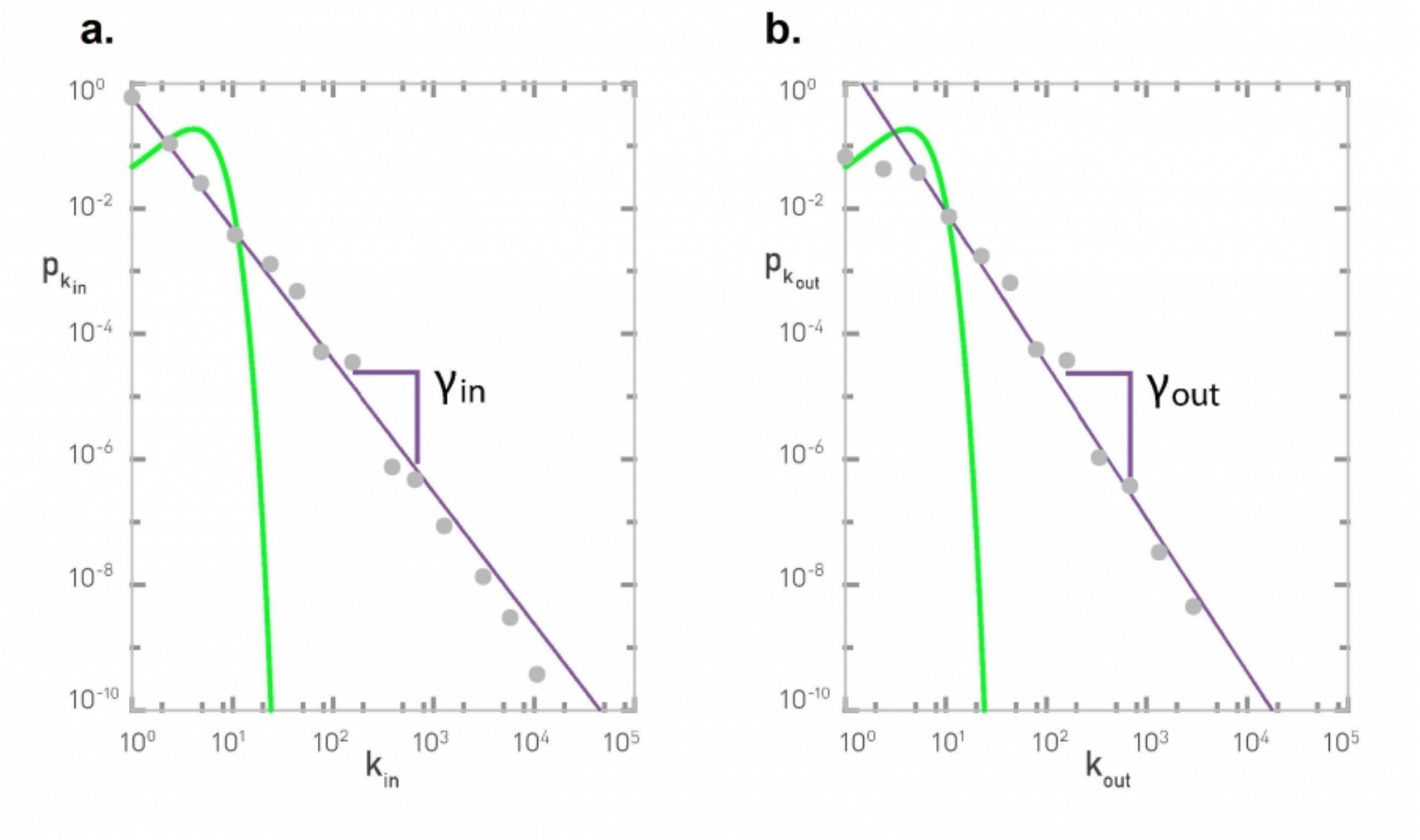
$$p_k \sim k^{-\gamma}$$

- This means that

$$\log p_k \sim -\gamma \log k$$

  And as such, the straight line in log-log scale plots is observed.
- In real-world networks, γ has a value of around 2 to 3
- Useful to compare between similar networks

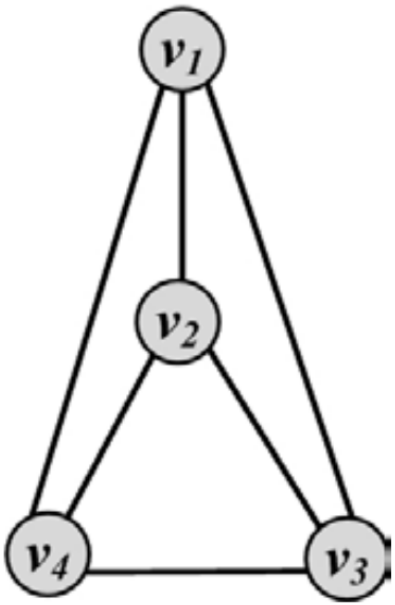
Power law exponent in directed networks



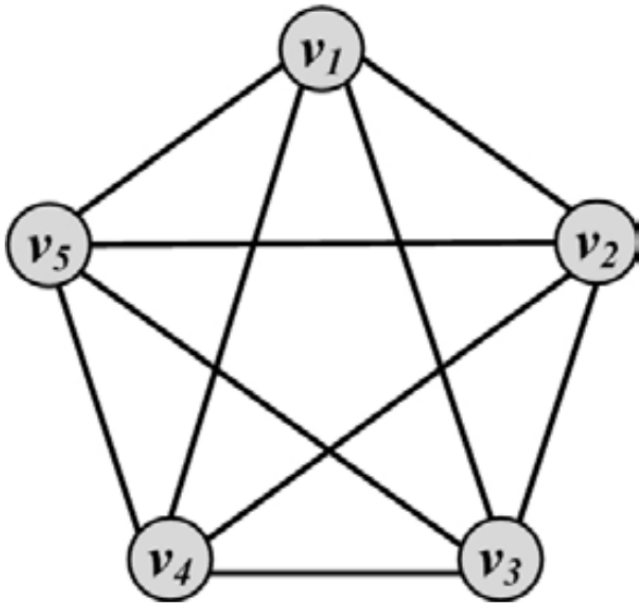
Source: A. Barabasi, Network Science, 2016.

Planar graphs

- **Planar graphs** can be visualized such that no two edges cross each other



(a) Planar Graph



(b) Non-planar Graph

Image: Zafarani et al., Social Media Mining, 2014.

Complete graphs

- In **complete graphs**, all pairs of nodes are connected
- The number of edges m is equal to $\frac{1}{2} \cdot n \cdot (n-1)$

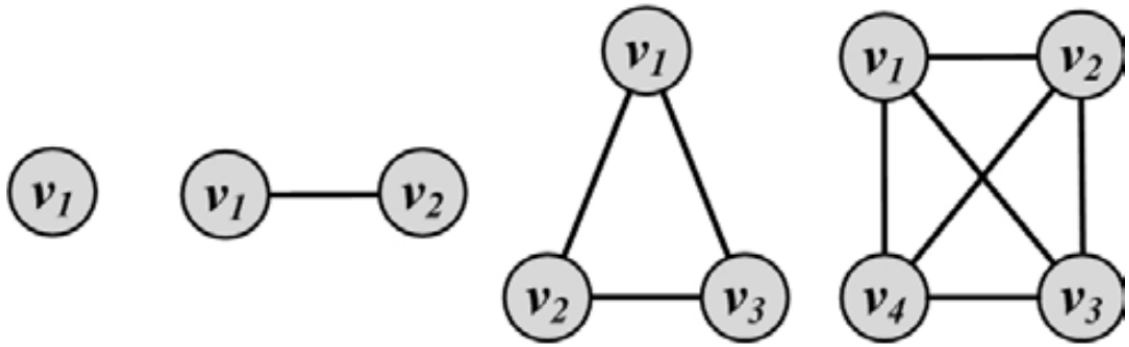


Figure: Complete graphs of size 1, 2, 3 and 4

Image: Zafarani et al., Social Media Mining, 2014.

Ego network

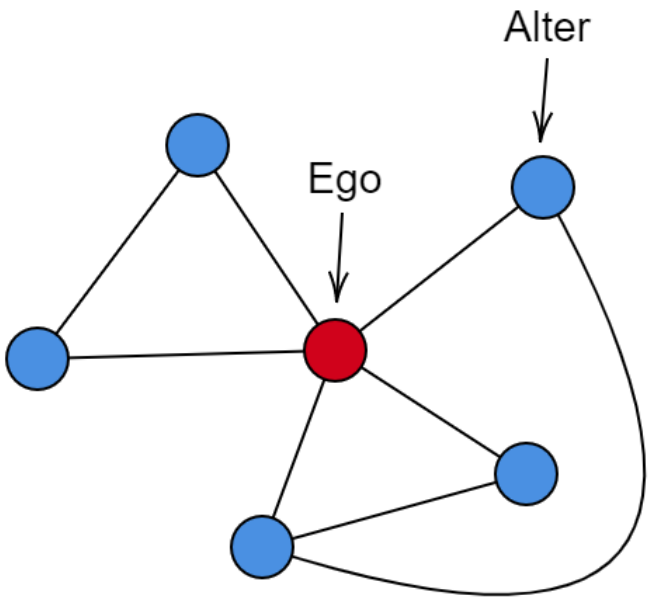


Figure: The **ego network** of a given node in a network consists of the set of nodes containing that node ("Ego") and its direct neighbors ("Alters"), and all edges present between the nodes in this set

Image: Wikipedia "Egocentric network.png", accessed 2022.

Trees

- A **tree** is a graph without cycles
- A set of disconnected trees is called a **forest**
- A tree with n nodes has $m = n - 1$ edges

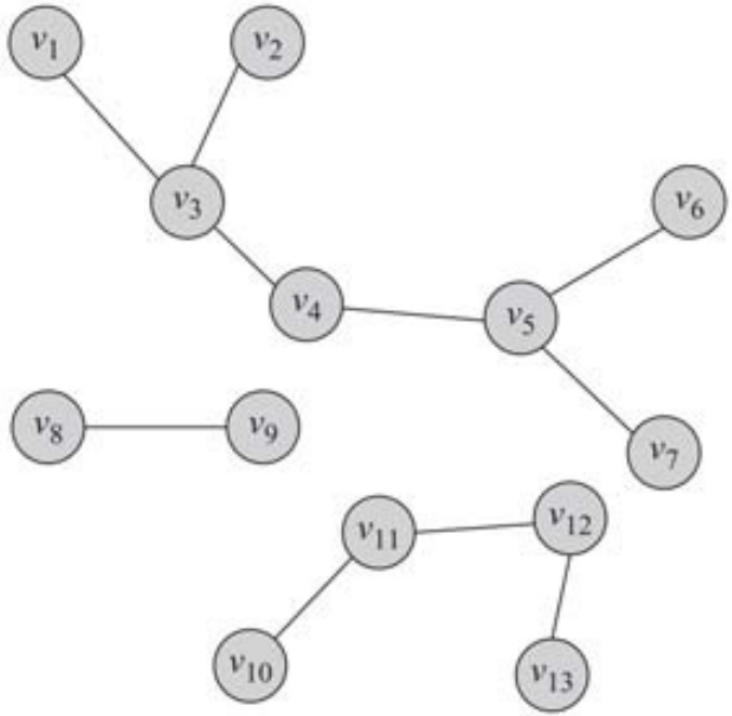


Image: Zafarani et al., Social Media Mining, 2014.

Trees

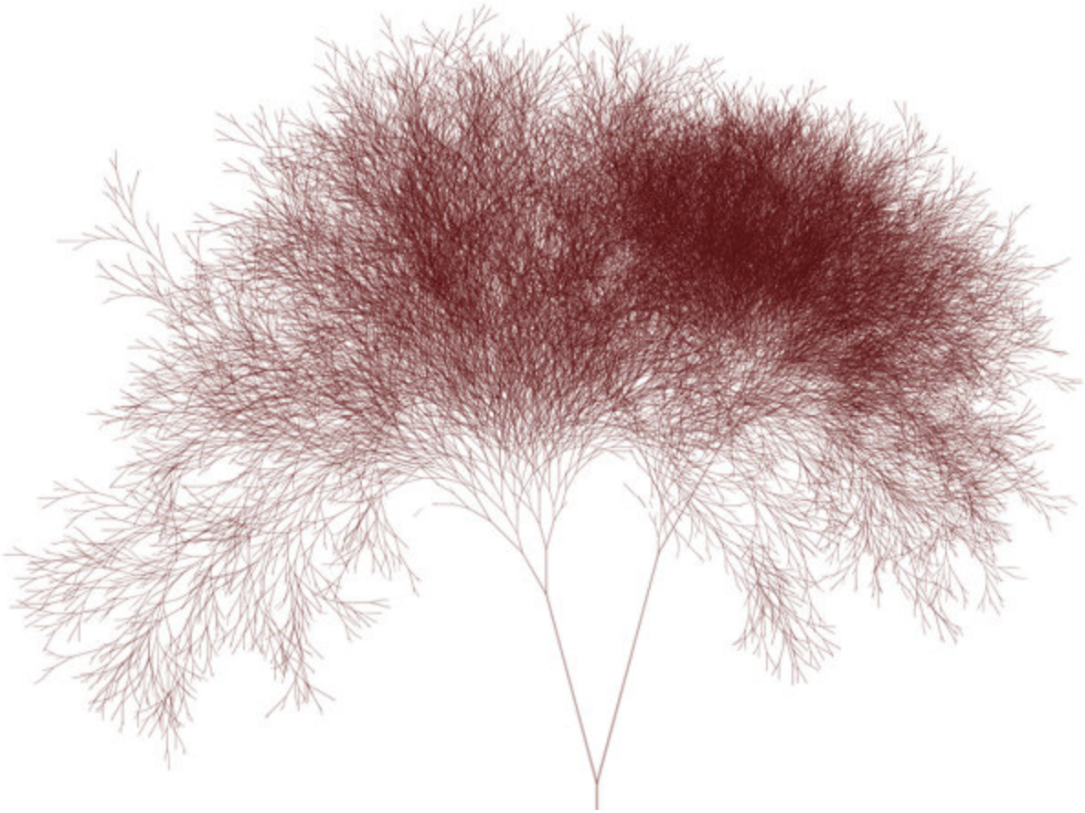


Image: M. Lima, Book of trees: Visualizing branches of knowledge, 2014.

Subgraphs

- Given a graph $G = (V, E)$
- **Subgraph** $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq (E \cap (V' \times V'))$
  (subset of the nodes and edges of the original network, commonly used when defining communities or clusters)
- **Subgraph** $G' = (V, E')$ with $E' \subseteq E$
  (only edges are left out, commonly used when modelling network evolution)
- Special subgraphs: spanning trees

Spanning trees

- A **spanning tree** is a tree and subgraph of a graph that covers all nodes of the graph
- In weighted graphs, a **minimal** spanning tree is one of minimal edge weight

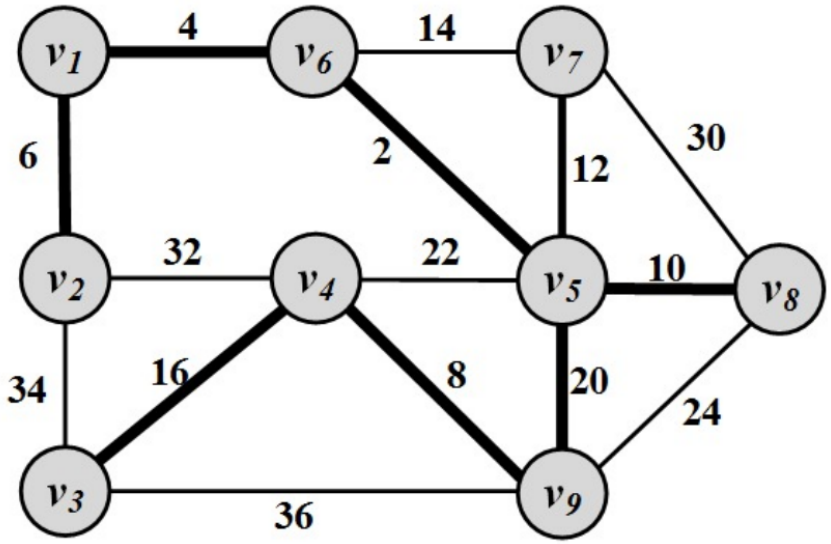


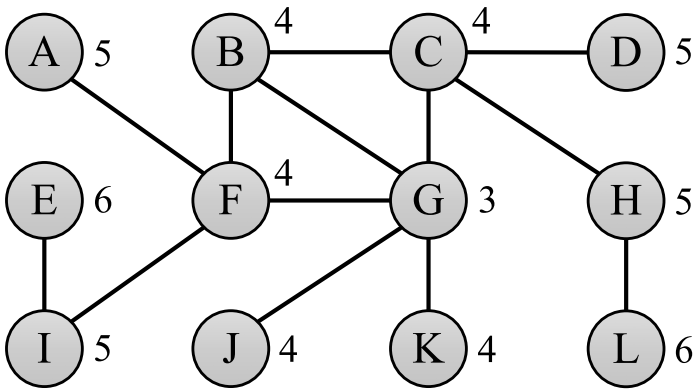
Image: Zafarani et al., Social Media Mining, 2014.

Diameter

- Distance $d(u, v) =$ length of shortest path from u to v
- Diameter $D(G) = \max_{u,v \in V} d(u, v) =$ maximal distance

Diameter

- Distance $d(u, v) =$ length of shortest path from u to v
- Diameter $D(G) = \max_{u,v \in V} d(u, v) =$ maximal distance
- Eccentricity $e(u) = \max_{v \in V} d(u, v) =$ length of longest shortest path from u
- Diameter $D(G) = \max_{u \in V} e(u) =$ maximal eccentricity
- Radius $R(G) = \min_{u \in V} e(u) =$ minimal eccentricity



Bridges

- **Bridge**: an edge whose removal will result in an increase in the number of connected components
- Also called **cut edges**, with applications in community detection

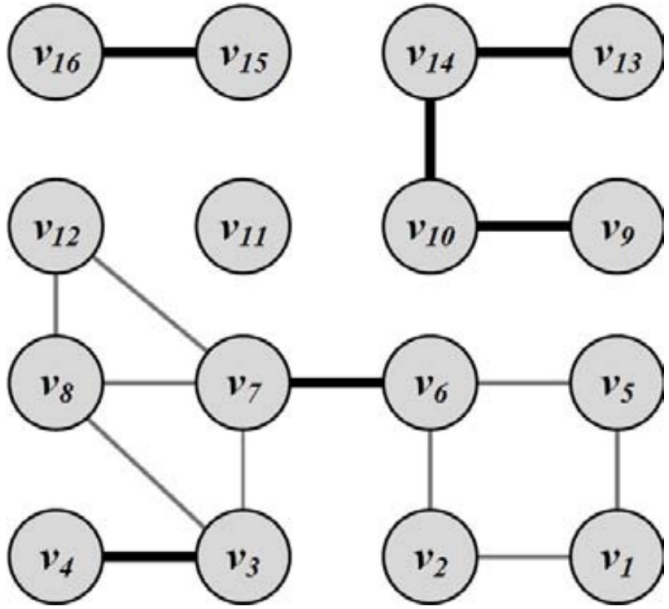


Image: Zafarani et al., Social Media Mining, 2014.

Graph traversal

- Given a network, how can we explore it?
- Specifically: exploration starting from a particular source
- Node **adjacency**: two nodes are adjacent if there is an edge connecting them
- **Neighborhood**: set of nodes adjacent to a node $v \in V$:

$$N(v) = \{w \in V : (v, w) \in E\}$$

- Techniques to iteratively explore neighborhoods: DFS and BFS

Graph traversal: DFS

- Depth First Search (DFS)

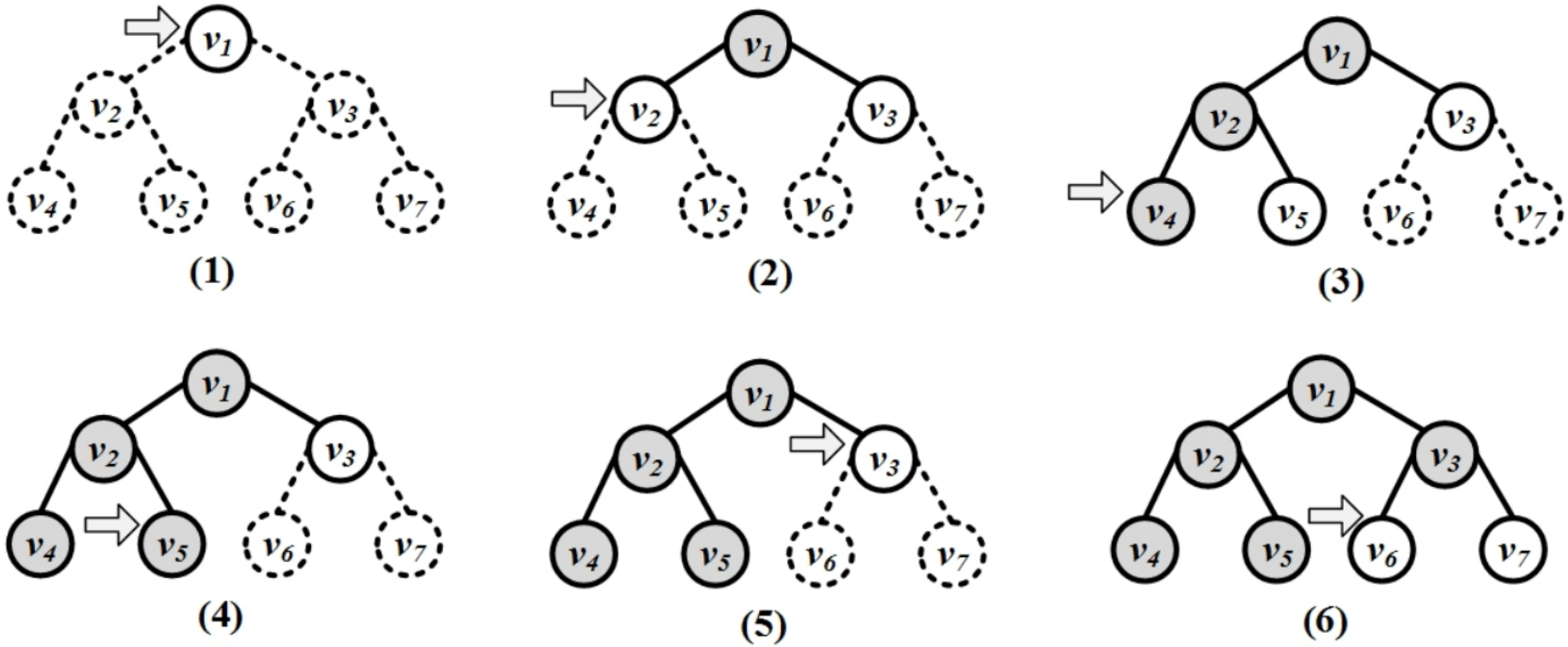
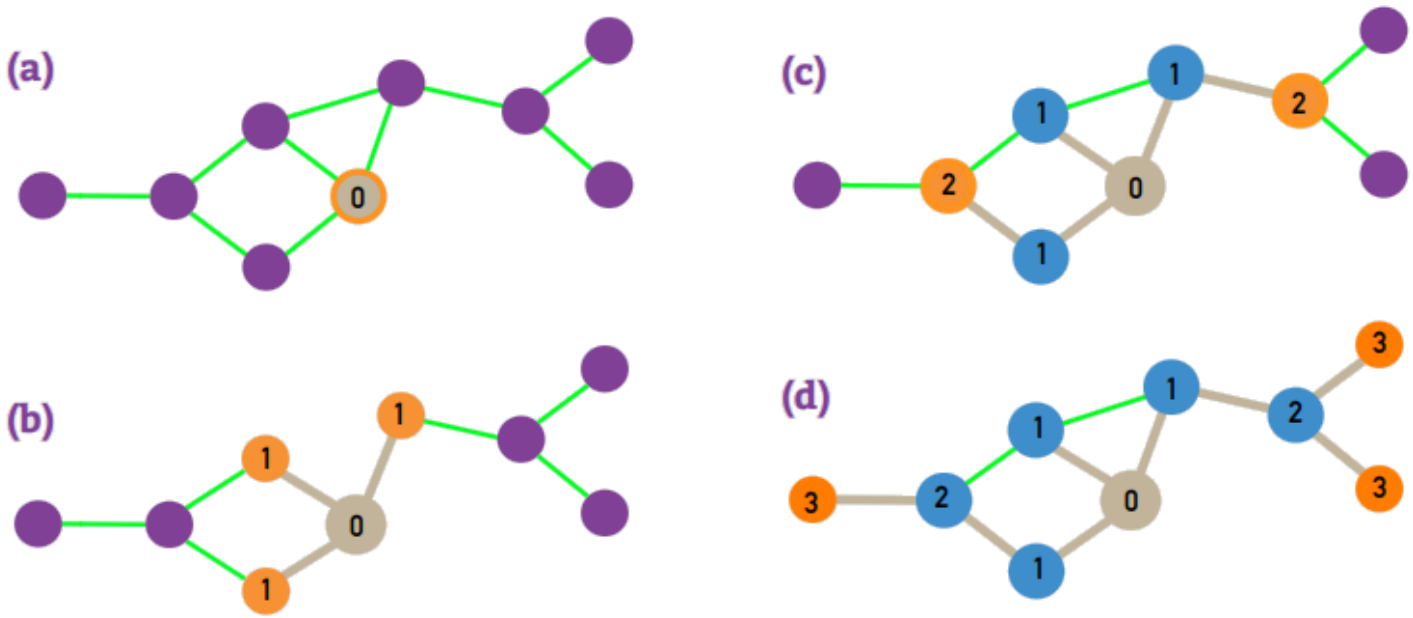


Image: Zafarani et al., Social Media Mining, 2014.

Graph traversal: BFS



Source: A. Barabasi, Network Science, 2016.

Graph traversal: BFS

- **Breadth First Search** (BFS)
- Graph traversal in level-order

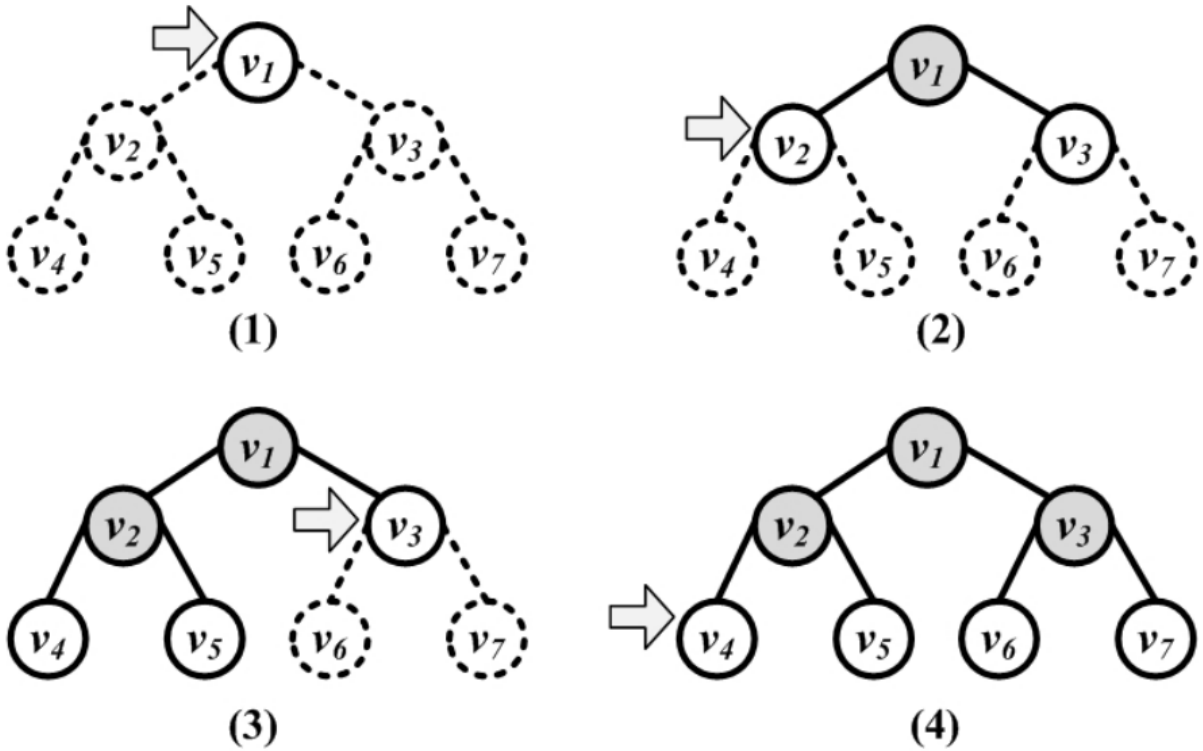


Image: Zafarani et al., Social Media Mining, 2014.

Graph traversal: BFS

- **Breadth First Search** (BFS)
- From source node, create a rooted spanning tree of the graph
- Graph traversal in level-order
- Often implemented using a queue
- BFS considers traversing each of the m edges once, so $O(m)$
- Important for computing various centrality measures

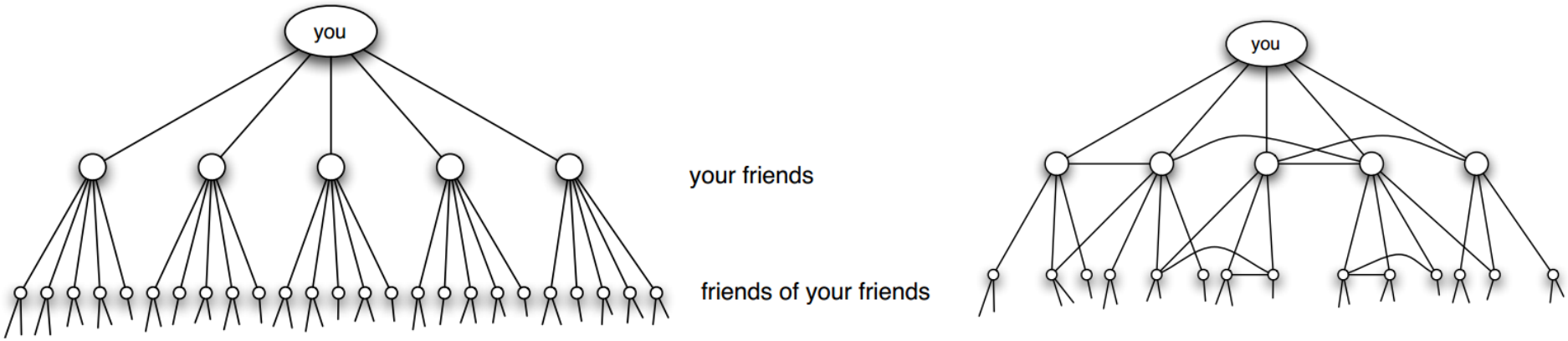


Image: Zafarani et al., Social Media Mining, 2014.

Centrality

Centrality

- Given a social network, which person is most important?
- What is the most important page on the web?
- Which protein is most vital in a biological network?
- Who is the most respected author in a scientific citation network?
- What is the most crucial router in an internet topology network?

Centrality

- **Node centrality**: the importance of a node with respect to the other nodes based on the structure of the network
- **Centrality measure**: computes the centrality value of all nodes in the graph
- For all $v \in V$ a measure M returns a value $C_M(v) \in [0; 1]$
- $C_M(v) > C_M(w)$ means that node v is more important than w


GAME OF THRONES
Watermark

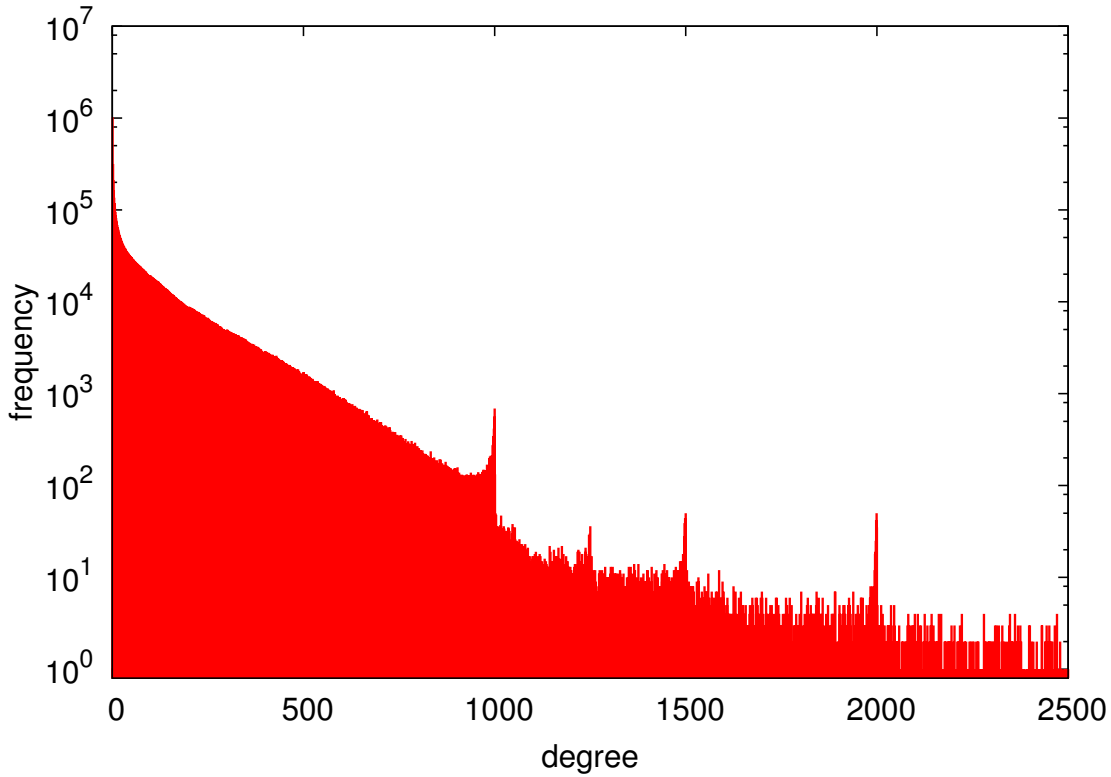
Degree centrality

- Undirected graphs – **degree centrality**: measure the number of adjacent nodes

$$C_d(v) = \frac{deg(v)}{n-1}$$

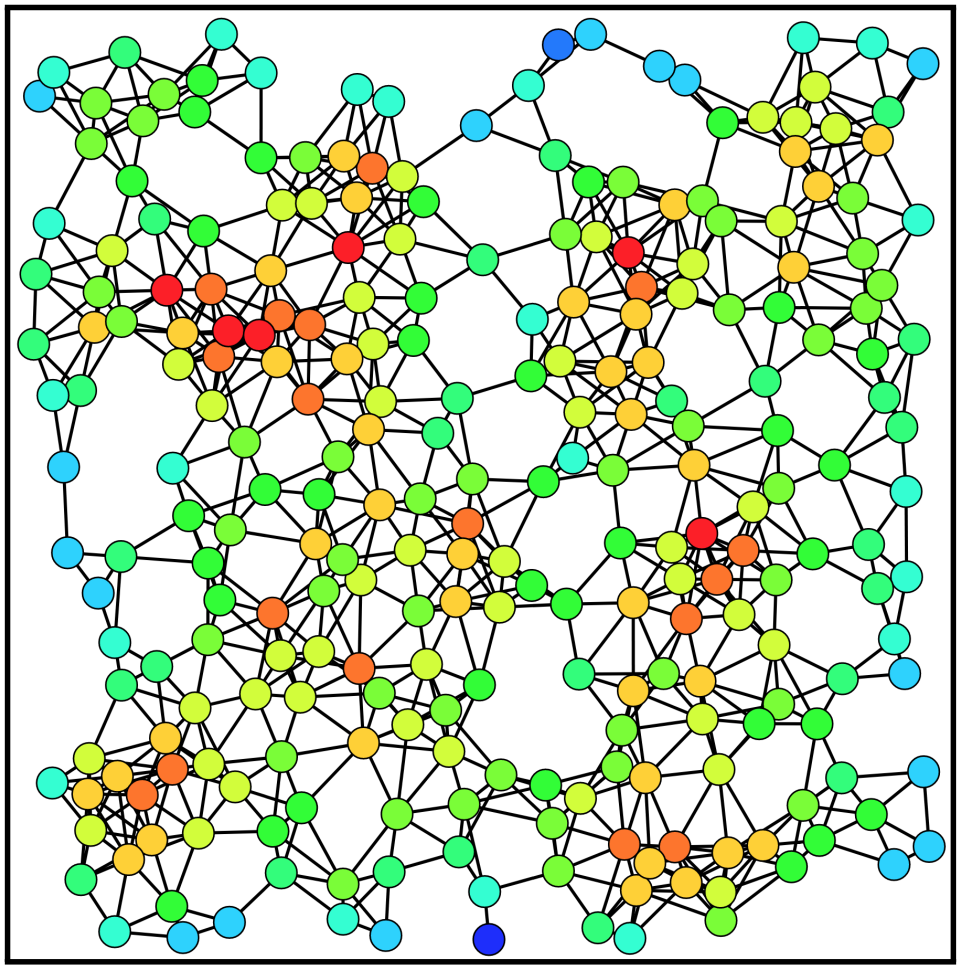
- Directed graphs — indegree centrality and outdegree centrality
- Local measure
- $O(1)$ time to compute

Degree distribution

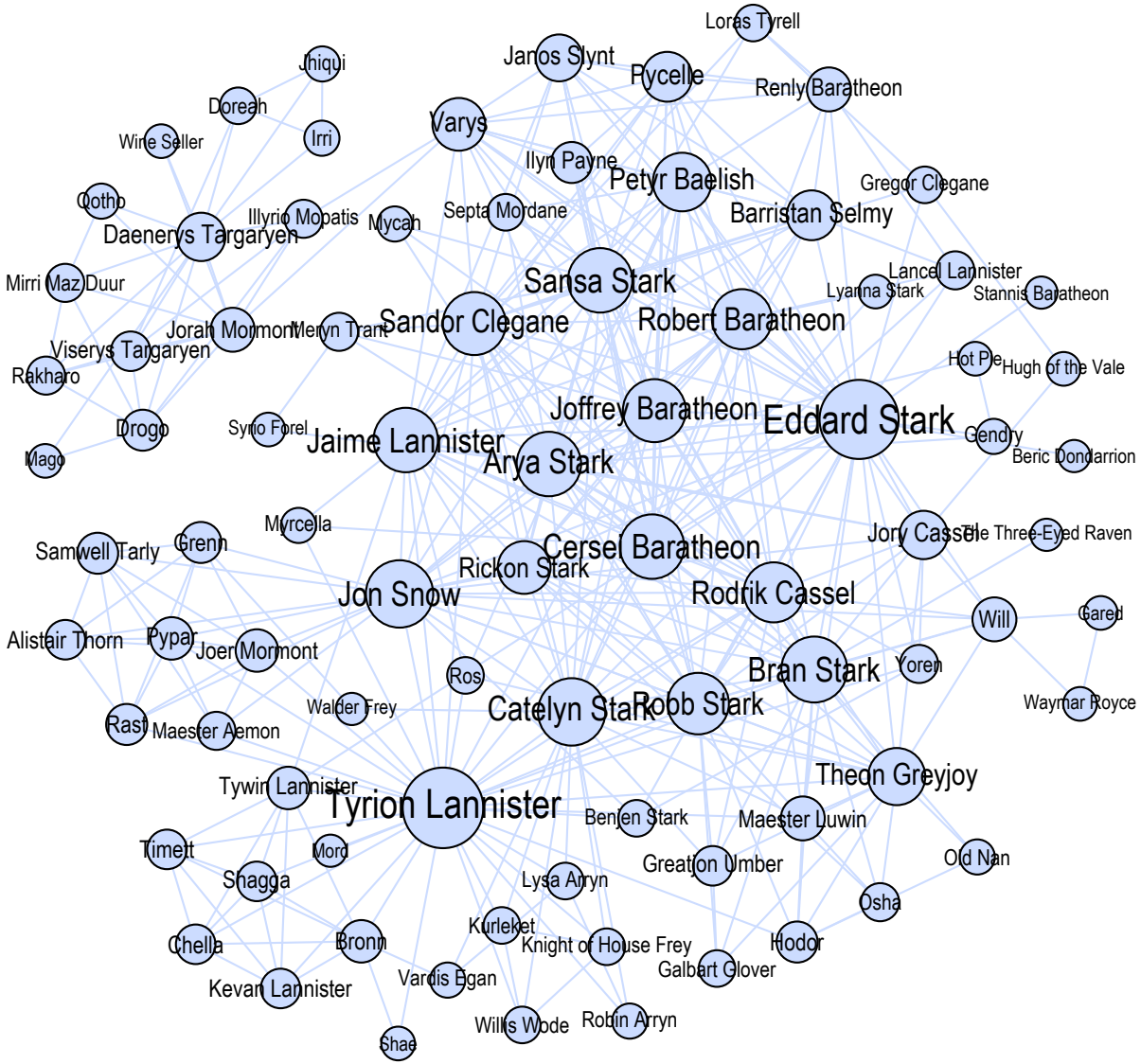


- Not so many distinct values in the lower ranges

Degree centrality



Degree centrality



Closeness centrality

- **Closeness centrality**: based on the average distance to all other nodes

$$C_c(v) = \left(\frac{1}{n-1} \sum_{w \in V} d(v, w) \right)^{-1}$$

- Global distance-based measure
- $O(mn)$ to compute: one BFS in $O(m)$ for each of the n nodes

Closeness centrality

- **Closeness centrality**: based on the average distance to all other nodes

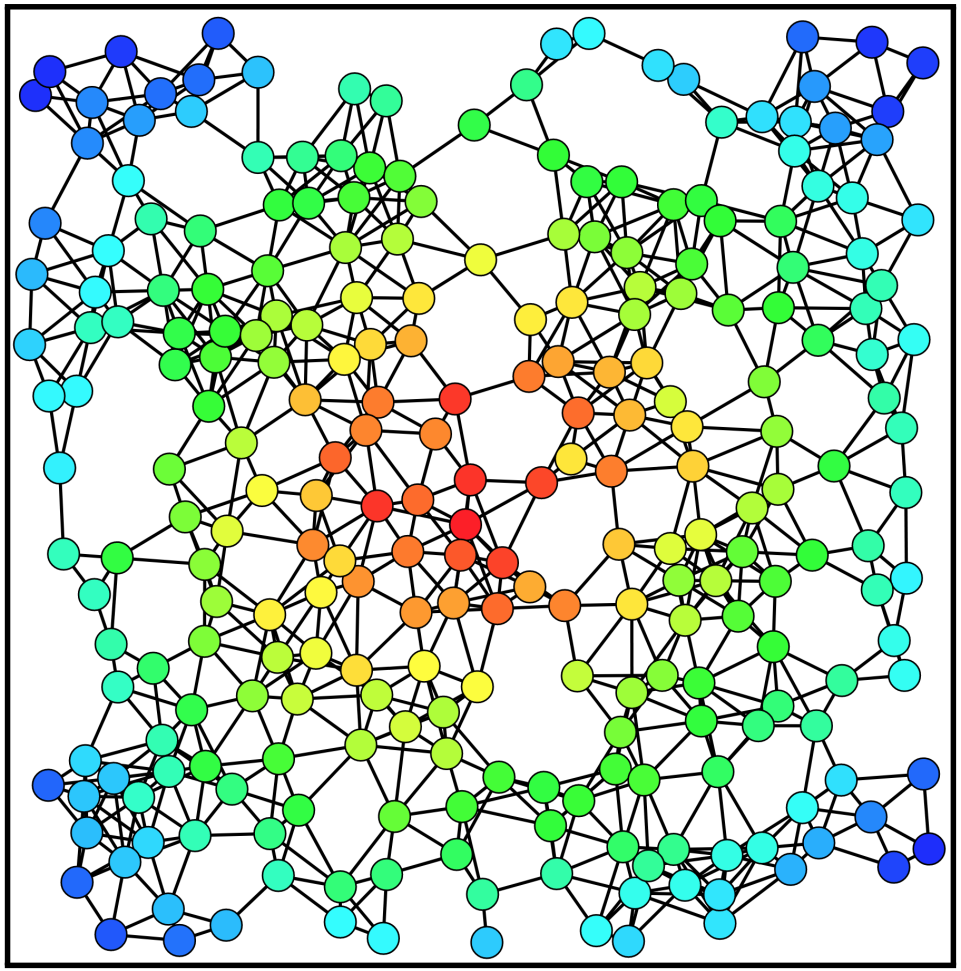
$$C_c(v) = \left(\frac{1}{n-1} \sum_{w \in V} d(v, w) \right)^{-1}$$

- Global distance-based measure
- $O(mn)$ to compute: one BFS in $O(m)$ for each of the n nodes
- Connected component(s)...
- **Harmonic centrality**: variant of closeness (not normalized)

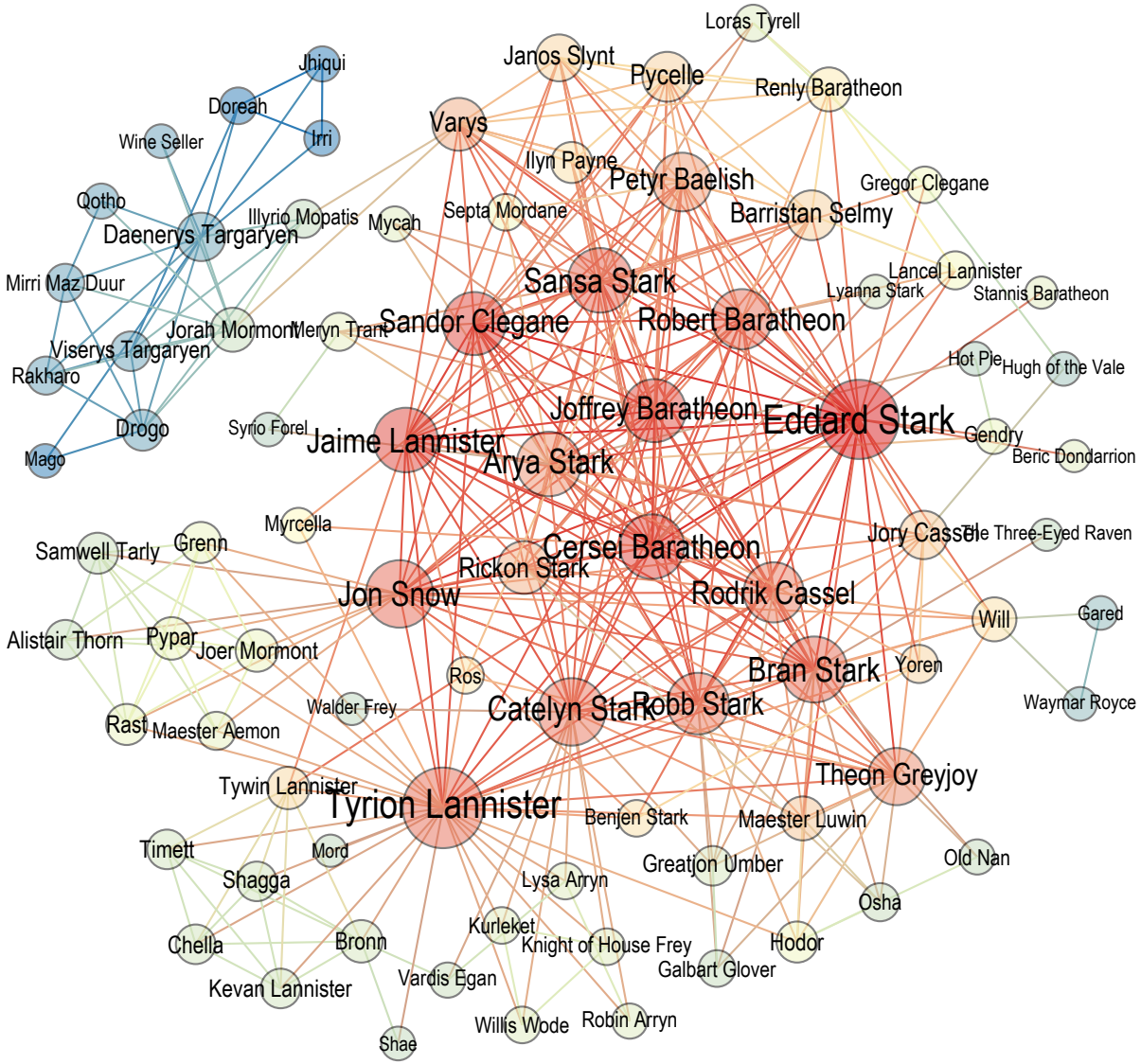
$$C_h(v) = \sum_{w \in V} \frac{1}{d(w, v)}$$

Sabidussi, G., The centrality index of a graph, *Psychometrika* 31(4): 581–603, 1966.

Closeness centrality



Degree vs. closeness centrality



Betweenness centrality

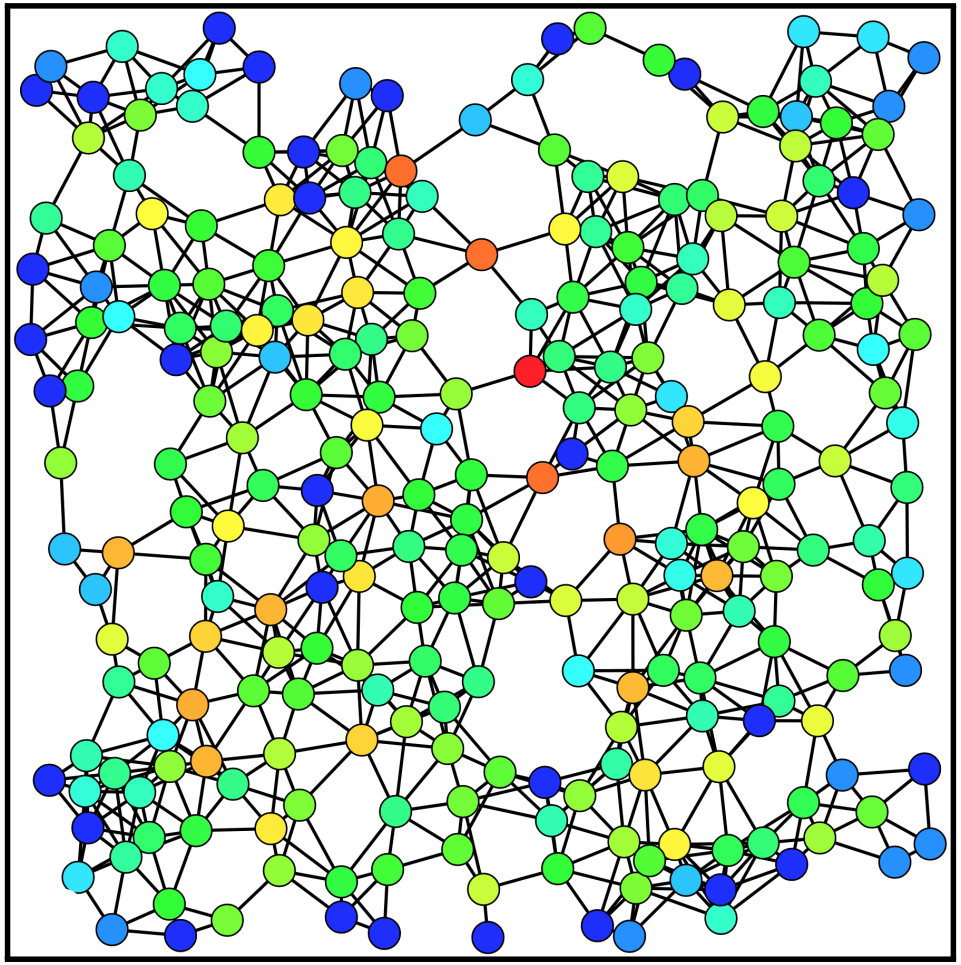
- **Betweenness centrality**: measure the number of shortest paths that run through a node

$$C_b(u) = \sum_{\substack{v,w \in V \\ v \neq w, u \neq v, u \neq w}} \frac{\sigma_u(v,w)}{\sigma(v,w)}$$

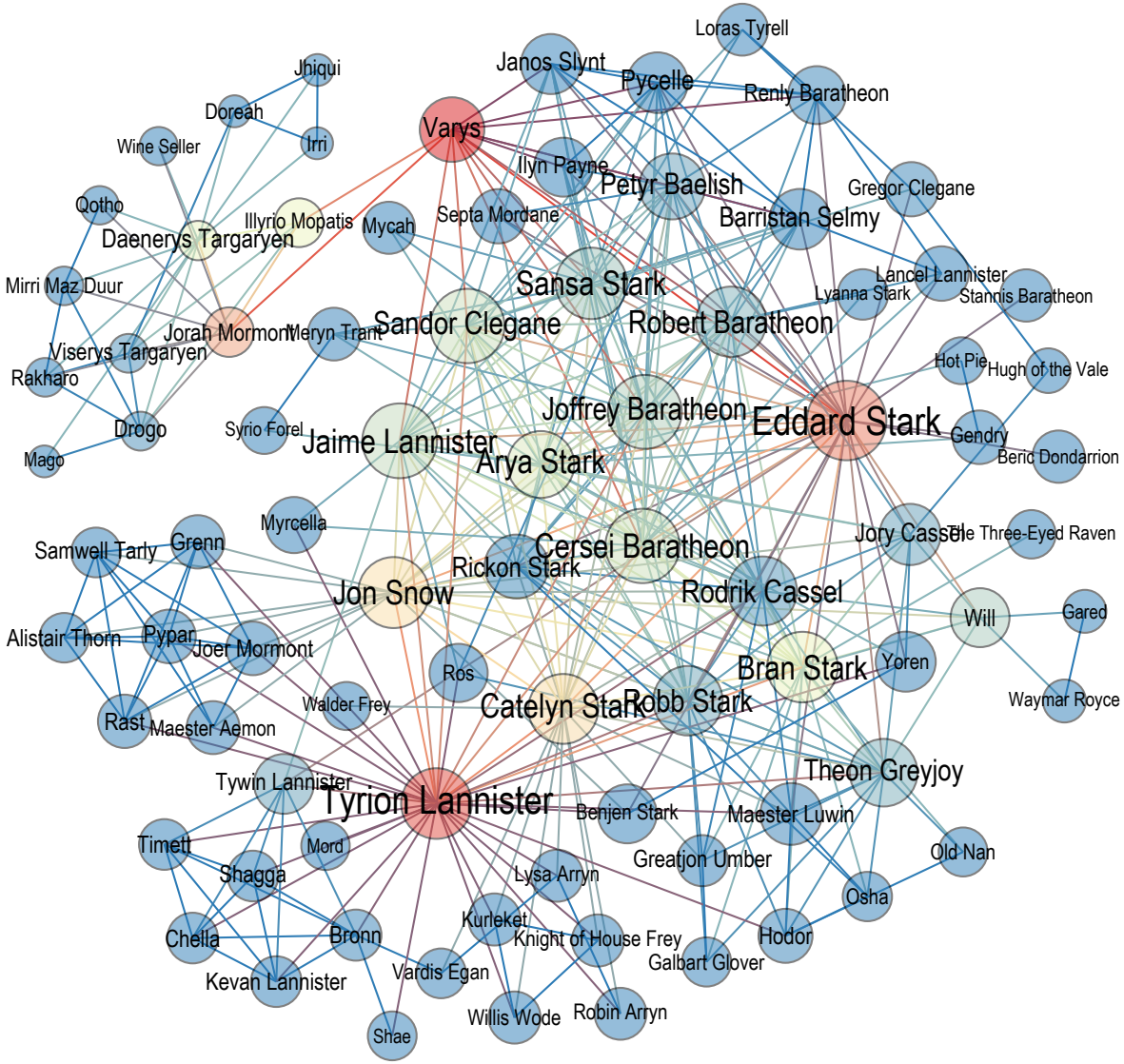
- $\sigma(v,w)$ is the number of shortest paths from v to w
- $\sigma_u(v,w)$ is the number of such shortest paths that run through u
- Divide by largest value to normalize to $[0;1]$
- Global path-based measure
- $O(2mn)$ time to compute (two "BFSes" for each node)

U. Brandes, "A faster algorithm for betweenness centrality", Journal of Mathematical Sociology 25(2): 163–177, 2001

Betweenness centrality



Degree vs. betweeness centrality



Centrality measures compared

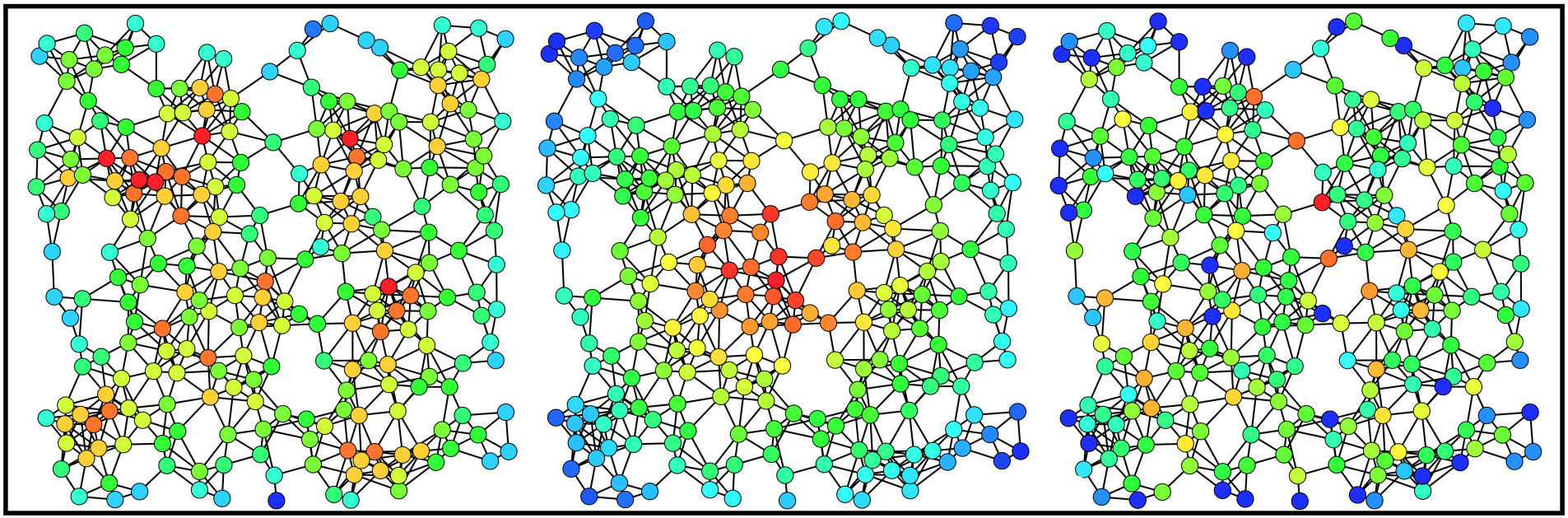


Figure: Degree, closeness and betweenness centrality

Source: "Centrality"' by Claudio Rocchini, Wikipedia `File:Centrality.svg`

Eccentricity centrality

- Node **eccentricity**: length of a longest shortest path (distance to a node furthest away)

$$e(v) = \max_{w \in V} d(v, w)$$

- **Eccentricity centrality**:

$$C_e(v) = \frac{1}{e(v)}$$

- Worst-case variant of closeness centrality
- $O(mn)$ to compute: one BFS in $O(m)$ for each of the n nodes

Eccentricity centrality

- Node **eccentricity**: length of a longest shortest path (distance to a node furthest away)

$$e(v) = \max_{w \in V} d(v, w)$$

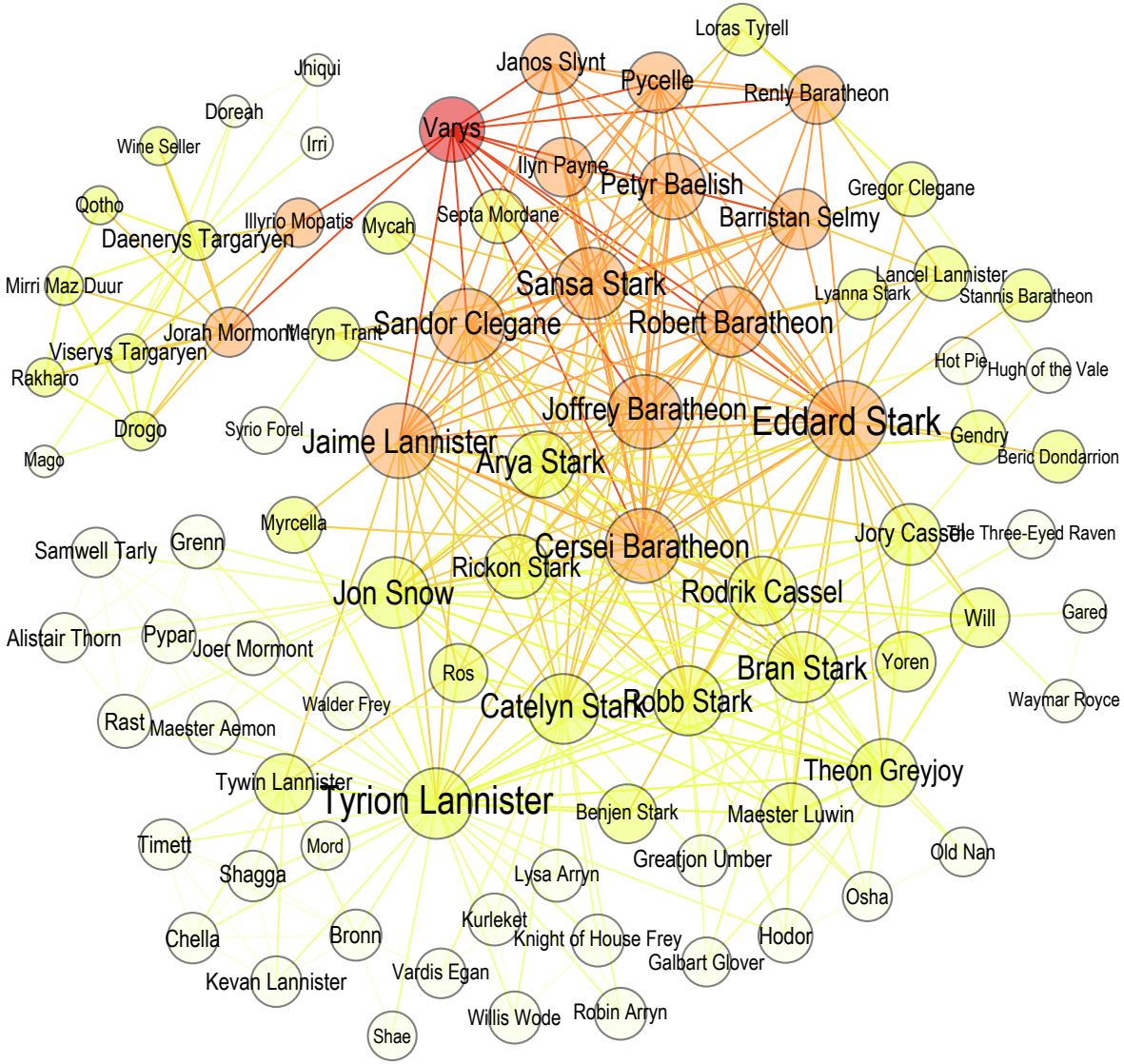
- **Eccentricity centrality**:

$$C_e(v) = \frac{1}{e(v)}$$

- Worst-case variant of closeness centrality
- $O(mn)$ to compute: one BFS in $O(m)$ for each of the n nodes
  - Large optimizations possible using lower and upper bounds, see F.W. Takes and W.A. Kosters, Computing the Eccentricity Distribution of Large Graphs, *Algorithms*, vol. 6, nr. 1, pp. 100-118, 2013.

Eccentricity centrality

Degree vs. eccentricity centrality

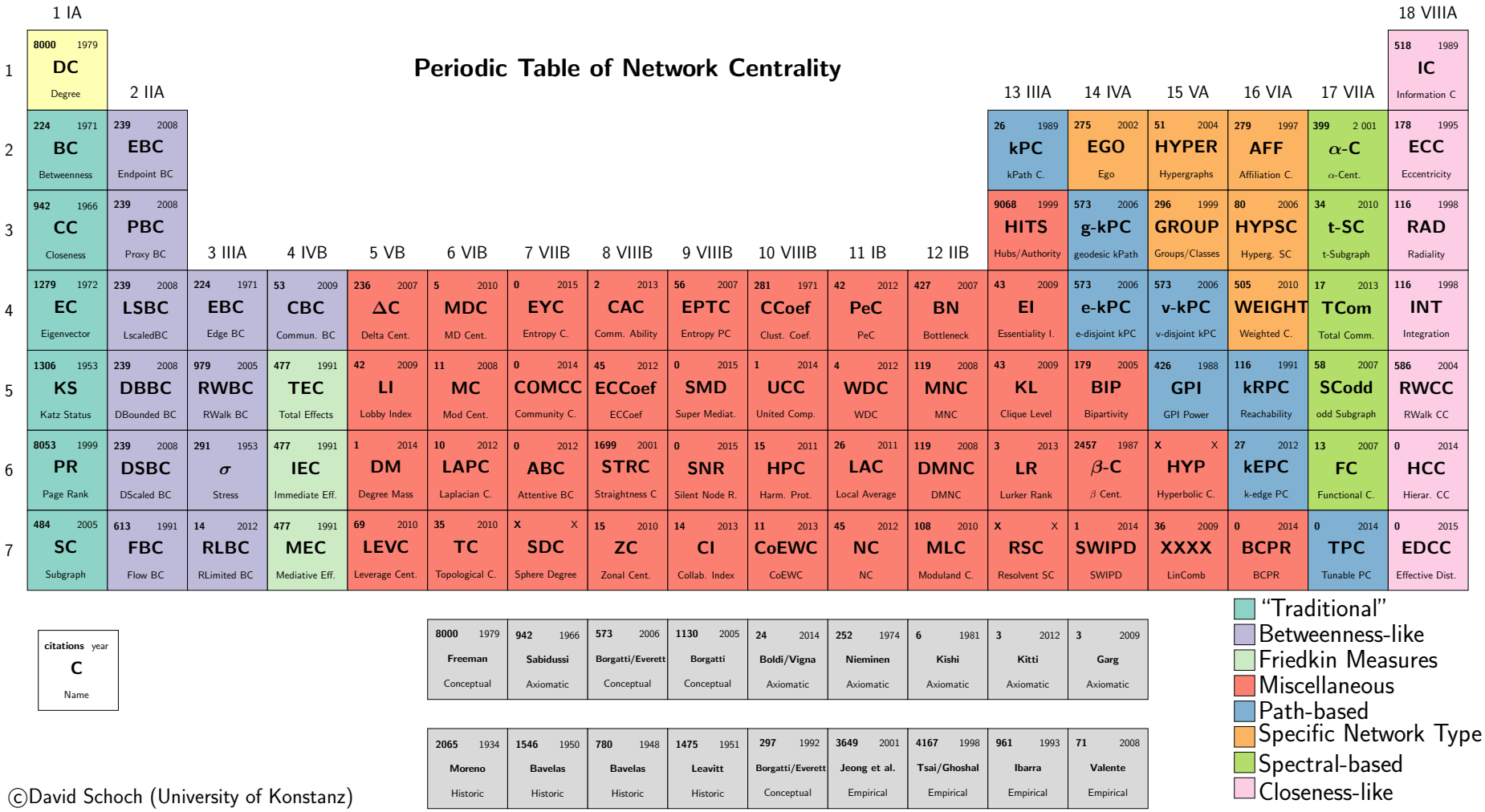


Centrality measures

- Distance/path-based measures:
  - Degree centrality $O(n)$
  - Closeness centrality $O(mn)$
  - Betweenness centrality $O(mn)$
  - Eccentricity centrality $O(mn)$

  (complexity is for computing centralities of all n nodes)
- Many more: Eigenvector centrality, Katz centrality, ...
- Approximating these measures is also possible
- Also: propagation-based centrality measures like PageRank

Periodic table of centrality

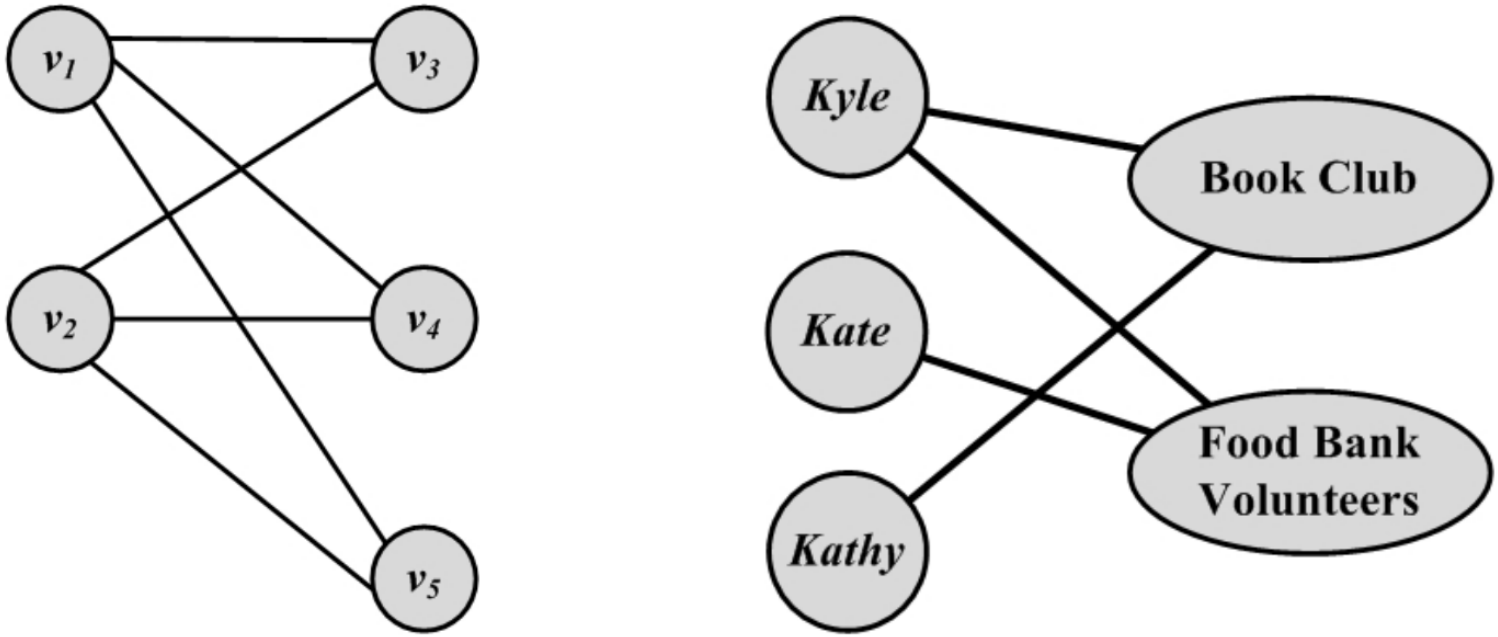


Network projection

Bipartite graphs

- In bipartite graphs the set of of nodes V can be split into two node sets V_L and V_R such that all edges E of the graph have their endpoints in different node sets. Specifically:
  - $V = V_L \cup V_R$
  - $V_L \cap V_R = \emptyset$
  - $E \subseteq V_L \times V_R$
- Also called **two-mode networks** or **heterogenic networks** (as opposed to respectively one-mode networks and homogenic networks)
- Called **affiliation networks** in a social network context
- So, two different types of nodes ...

Bipartite graphs



(a) Bipartite Graph (b) Affiliation Network

Image: Zafarani et al., Social Media Mining, 2014.

Projecting networks

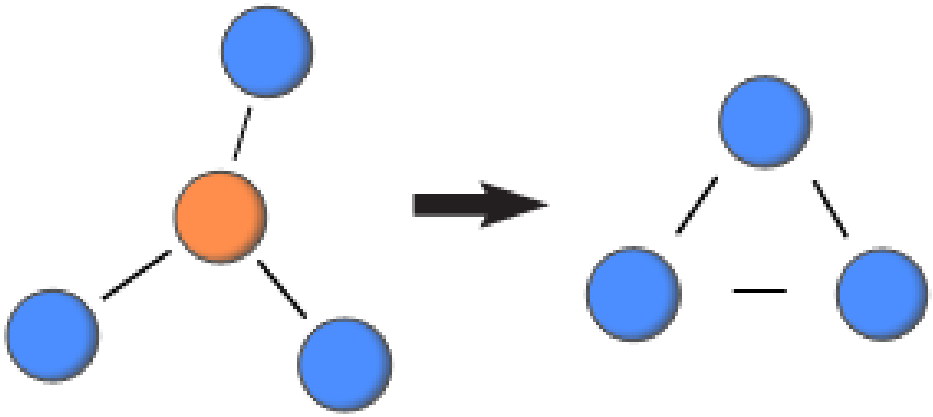
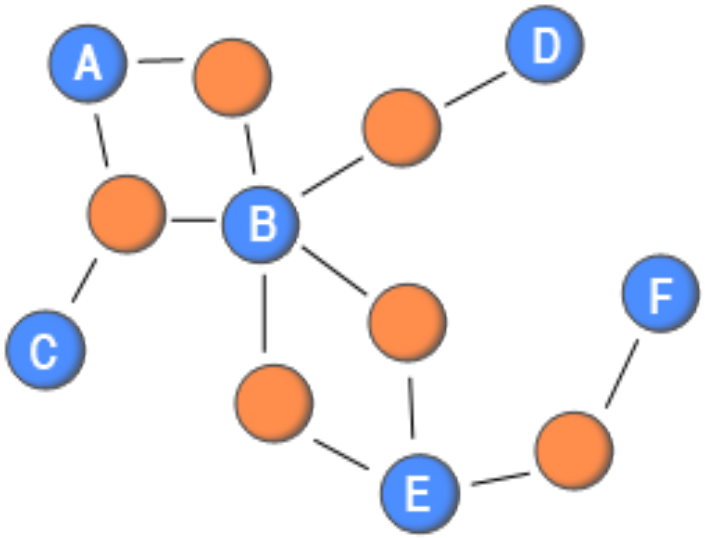


Image: `http://toreopsahl.com/`

Weighted projection



$\rightarrow$

Weighted projection

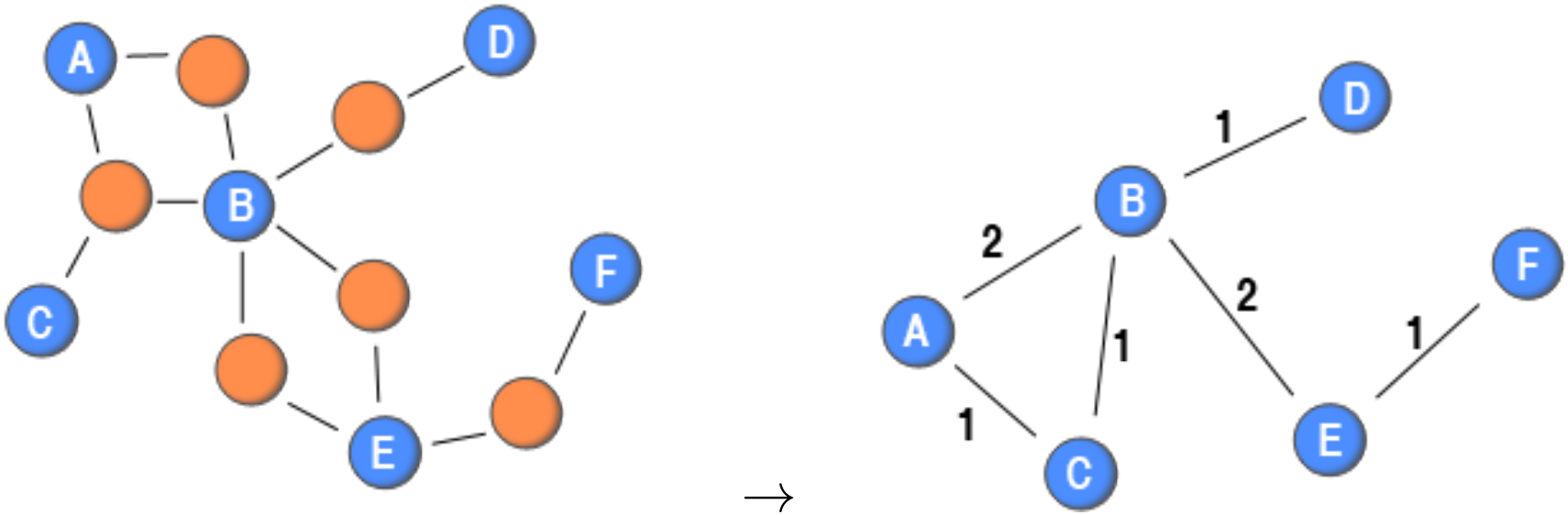
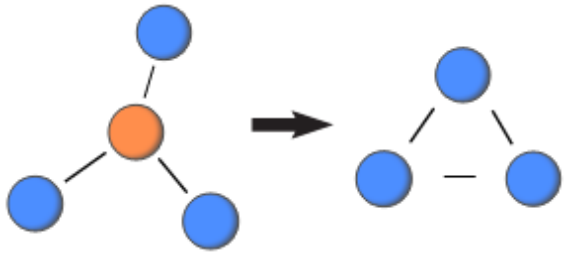


Image: `http://toreopsahl.com`

Projection algorithm



- Given a bipartite graph $G = (V_L \cup V_R, E)$ with $E \subseteq V_L \times V_R$, generate the projected graph $G' = (V_L, E')$
  - Initialize $G' = (V_L, E')$ with $E' = \emptyset$
  - For each node $v \in V_R$, determine its neighborhood $N(v) \subseteq V_L$
    - For each distinct node pair $v_i, v_j \in N(v)$, add the edge (v_i, v_j) to E'
    - Optionally, assign a weight to edge (v_i, v_j) based on how often it occurs
- Analogously, the projection from $G = (V_L \cup V_R, E)$ to $G'' = (V_R, E'')$ can be made

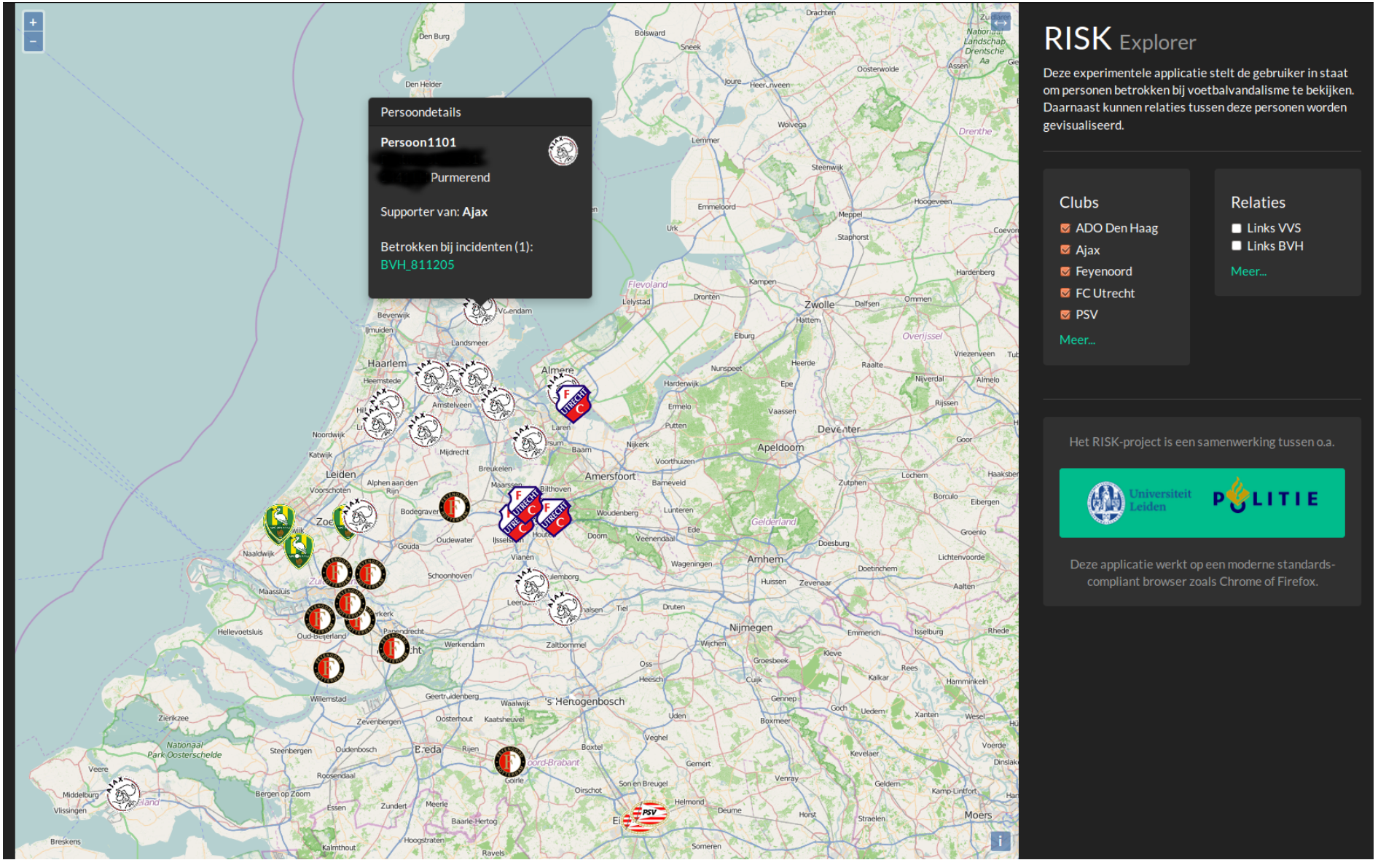
Network analysis on (almost) any dataset

- Different data objects typically have **attributes with identical values**
- The unique object identifier and the common attribute value are the two node types in a two-mode network representing the data
- The two-mode network can be converted into a one-mode network based on the common attribute
- Many projections of a dataset to a network are possible

Criminal networks

Example: Criminal networks

- Data science project with Dutch National Police
- Gain insight in social networks of soccer fans, group formation and organization
- Dataset: all entries in police systems of law violations of a particular group of people involved in soccer violence



RISK Explorer

Deze experimentele applicatie stelt de gebruiker in staat om personen betrokken bij voetbalvandalisme te bekijken. Daarnaast kunnen relaties tussen deze personen worden gevisualiseerd.

Clubs

- [x] ADO Den Haag
- [x] Ajax
- [x] Feyenoord
- [x] FC Utrecht
- [x] PSV

Meer...

Relaties

- [ ] Links VVS
- [ ] Links BVH

Meer...

Het RISK-project is een samenwerking tussen o.a.

Universiteit Leiden — POLITIE

Deze applicatie werkt op een moderne standards-compliant browser zoals Chrome of Firefox.

Criminal networks

Person ID	Incident ID	Incident Type
P000001	X00011	Straatroof/diefstal
P000001	X00014	Eenv. Mishandeling
P000002	X00011	Straatroof/diefstal
P000002	X00012	Eenv. Mishandeling
P000003	X00012	Eenv. Mishandeling
P000003	X00016	Bedreiging
P000004	X00012	Eenv. Mishandeling
P000004	X00017	Eenv. Mishandeling
P000005	X00013	Bedreiging
P000005	X00014	Eenv. Mishandeling
P000005	X00015	Straatroof/diefstal
P000006	X00013	Bedreiging
P000007	X00013	Bedreiging
P000008	X00013	Bedreiging
P000009	X00015	Straatroof/diefstal
P000010	X00016	Bedreiging
P000010	X00017	Eenv. Mishandeling
P000011	X00016	Bedreiging

Table: Data on suspects involved in incidents

Network formation

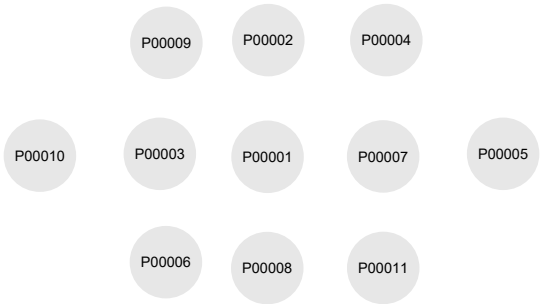
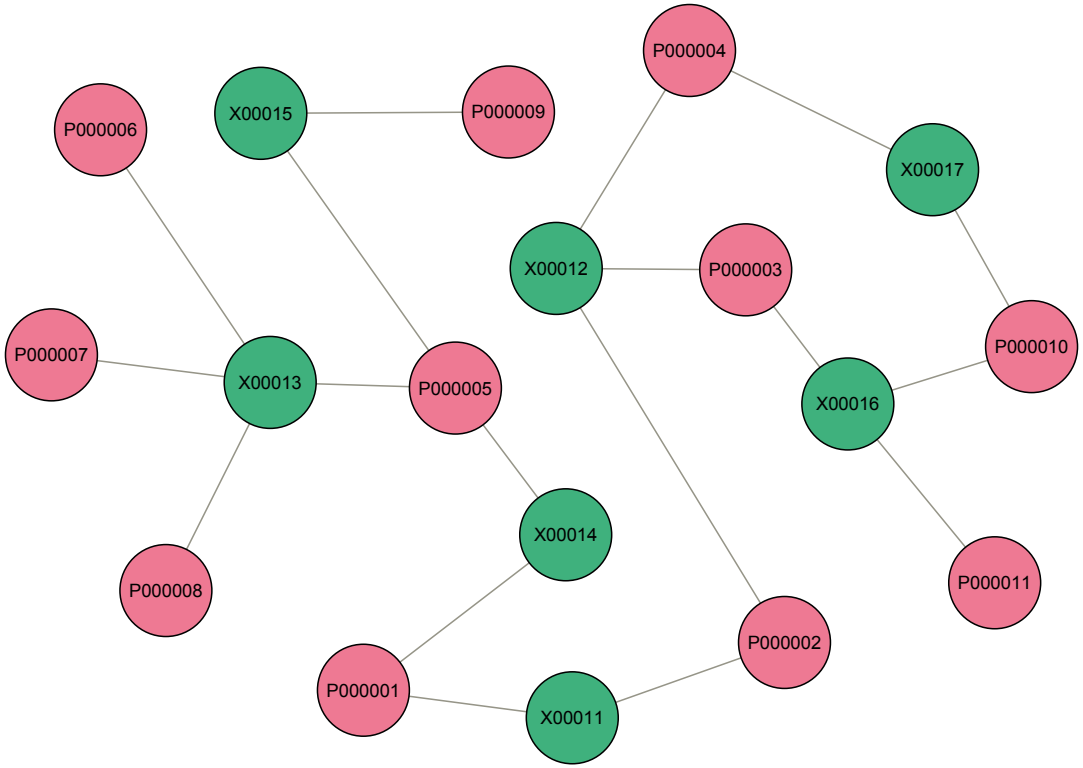


Figure: Suspects are nodes

Two-mode criminal network



Network formation

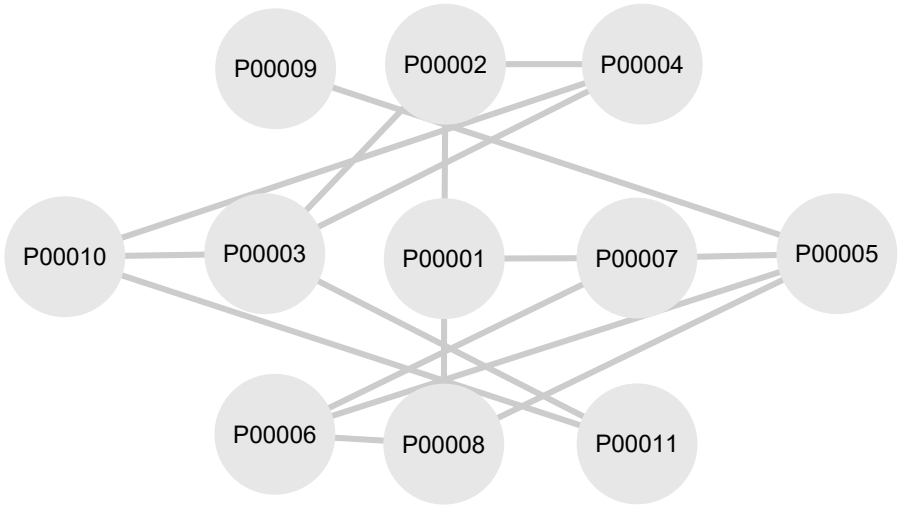


Figure: Edges are based on common involvement as a suspect

Network visualization

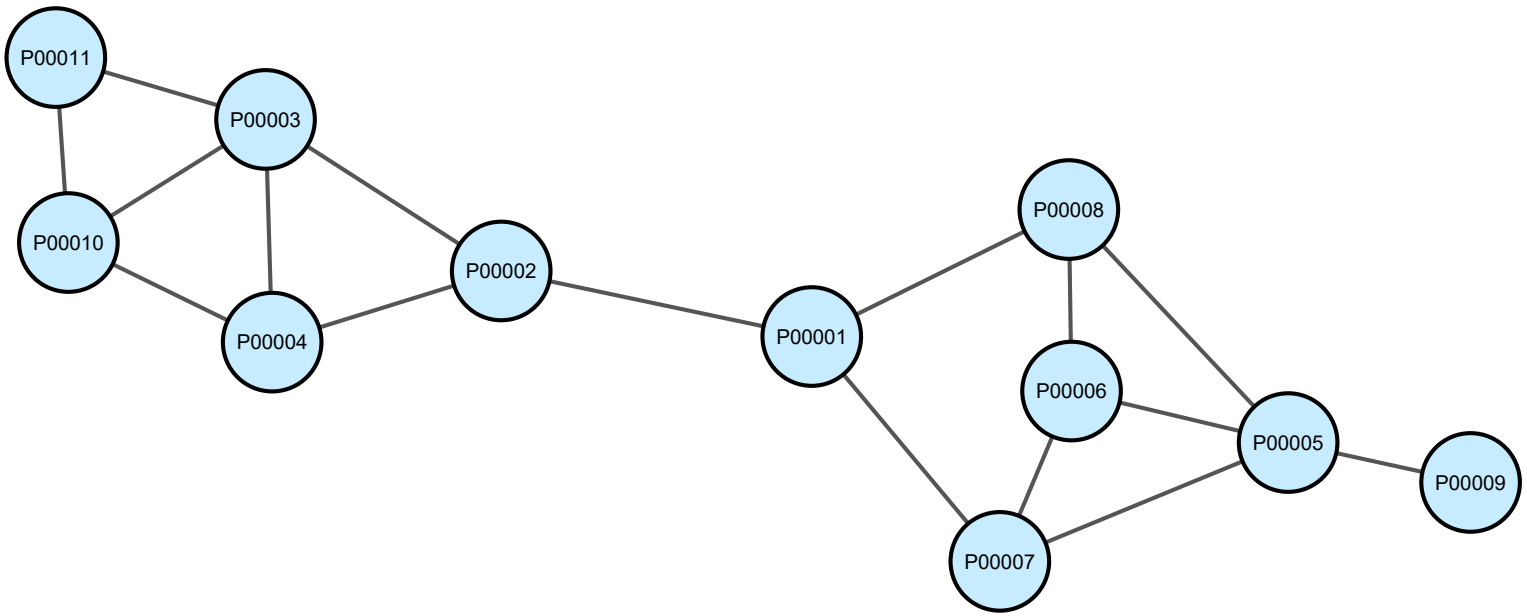


Figure: Force-directed visualization algorithm reveals structure

Network analysis: Centrality

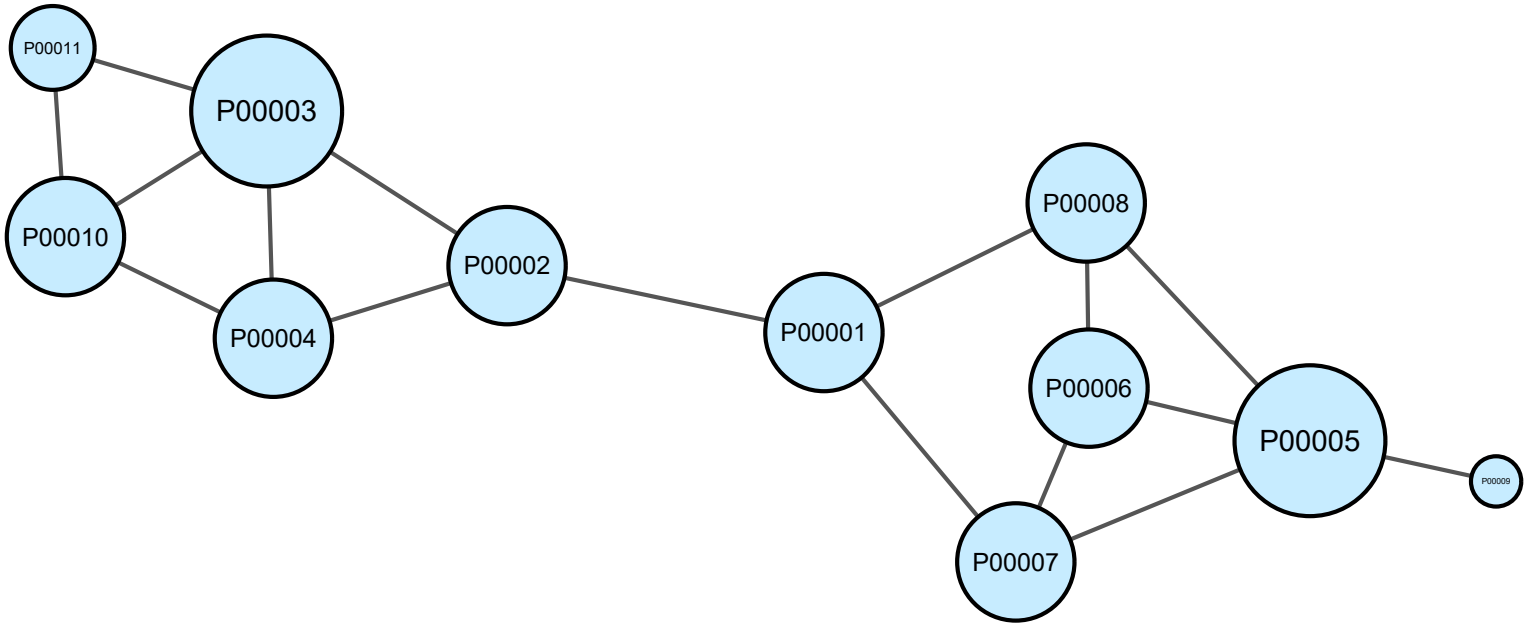


Figure: Degree centrality finds locally important nodes

Network analysis: Centrality

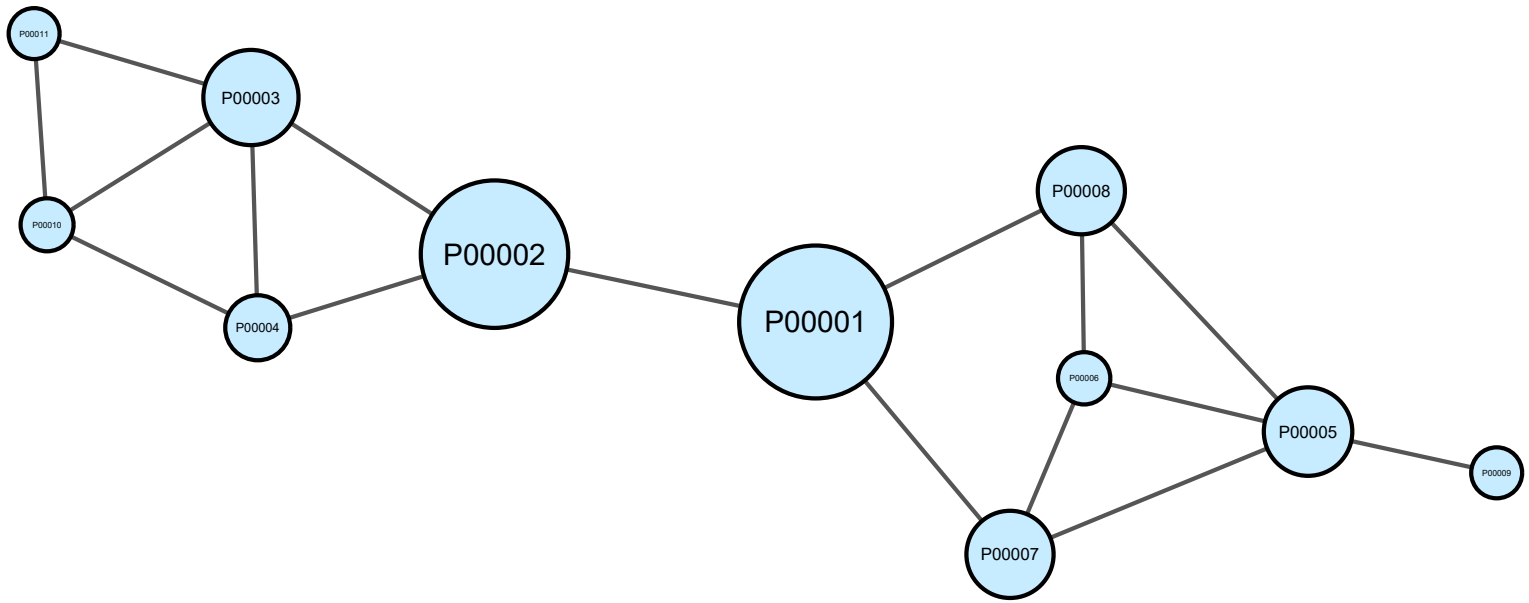


Figure: Betweenness centrality reveals globally important nodes

Network analysis: Community detection

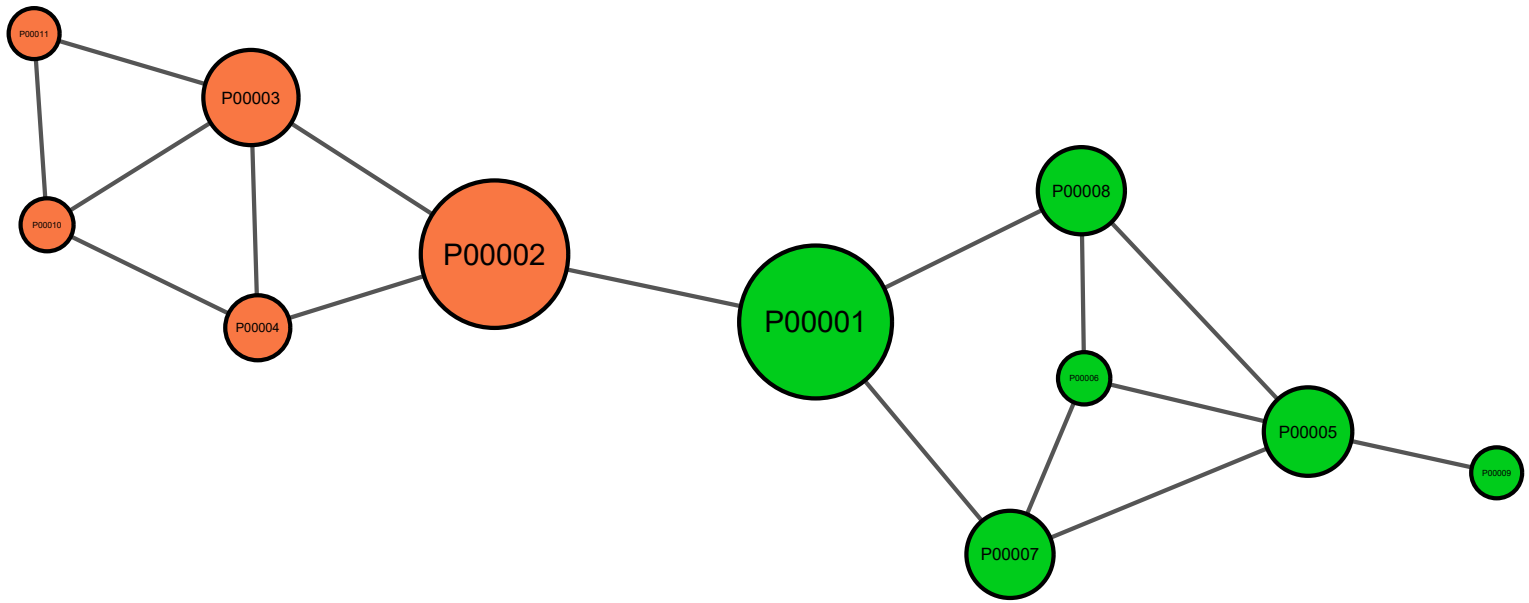
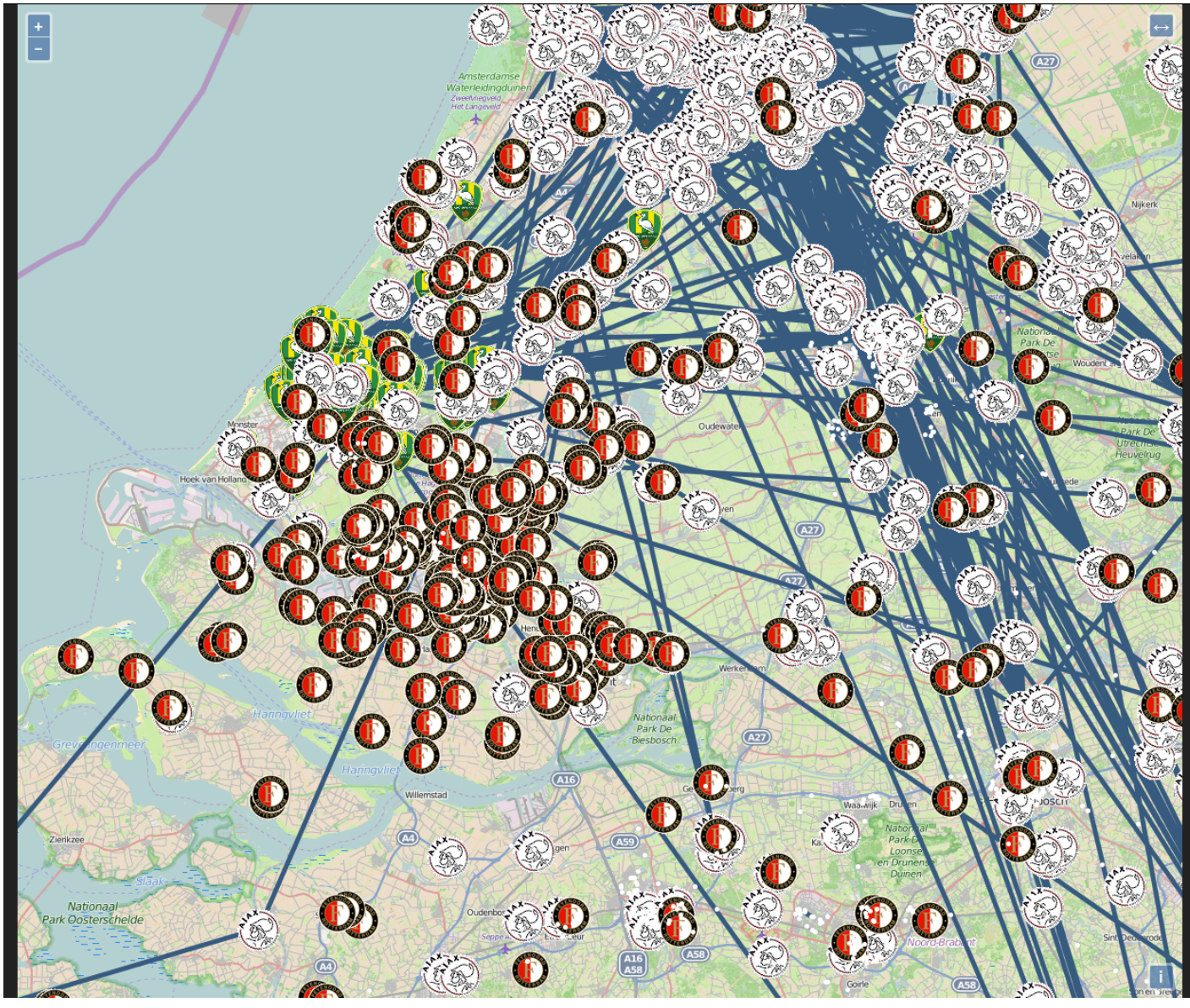


Figure: Community detection finds groups of tightly connected nodes



RISK Explorer

Deze experimentele applicatie stelt de gebruiker in staat om personen betrokken bij voetbalvandalisme te bekijken. Daarnaast kunnen relaties tussen deze personen worden gevisualiseerd.

Clubs

- [x] ADO Den Haag
- [x] Ajax
- [x] Feyenoord
- [ ] FC Utrecht
- [ ] PSV

Meer...

Relaties

- [ ] Links VVS
- [x] Links BVH

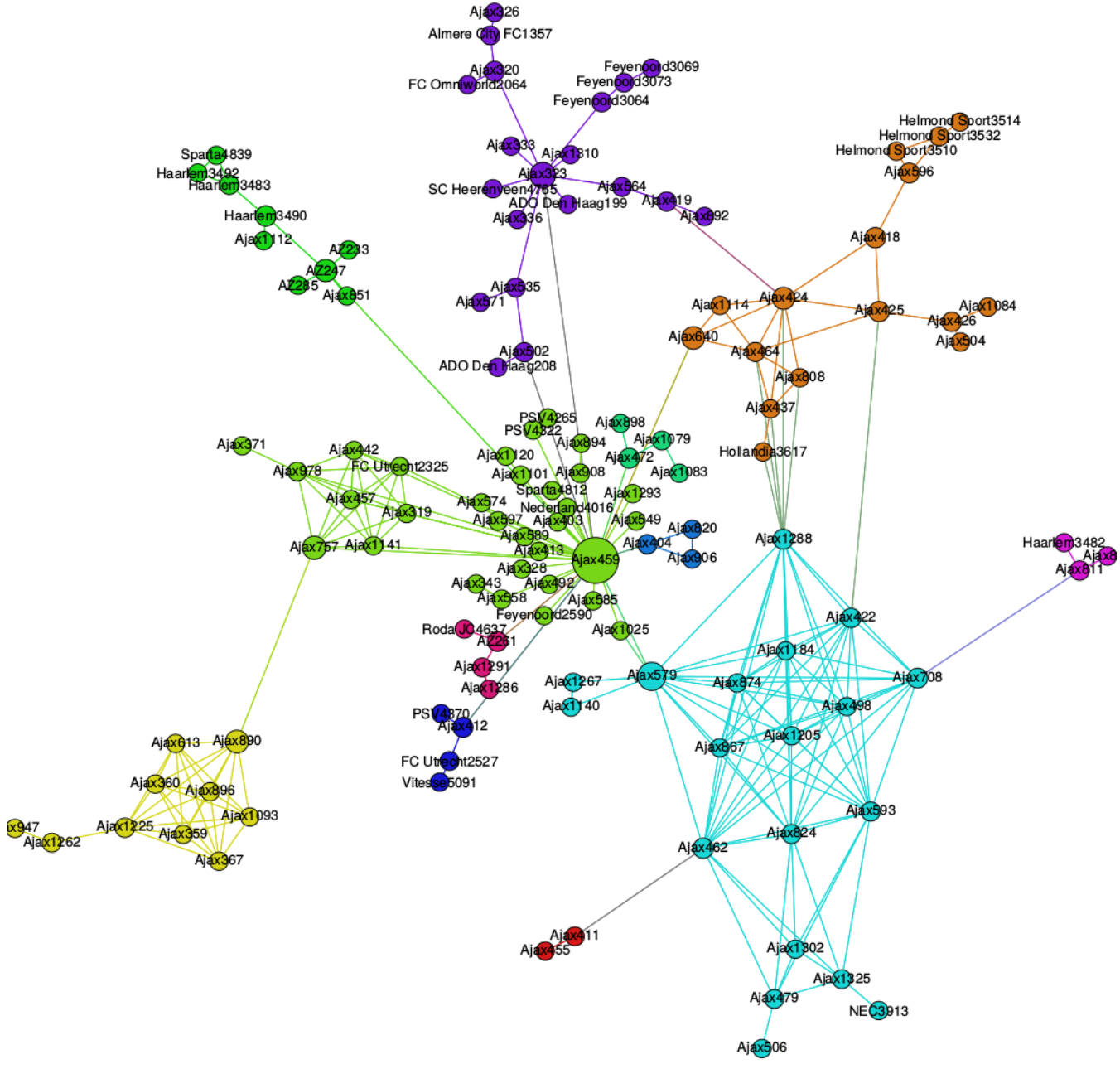
Meer...

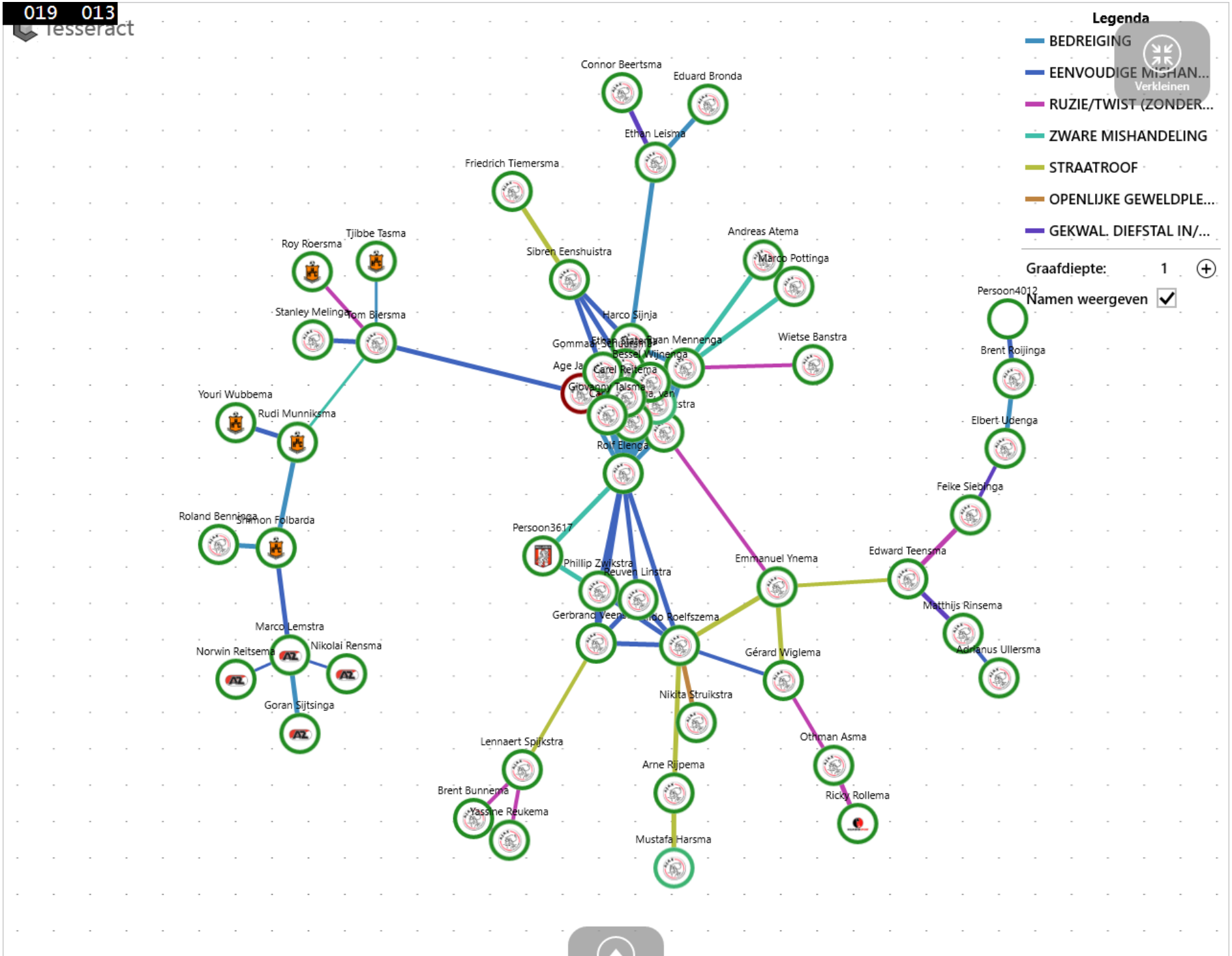
Het RISK-project is een samenwerking tussen o.a.

Universiteit Leiden

POLITIE

Deze applicatie werkt op een moderne standards-compliant browser zoals Chrome of Firefox.


Ajax326
Almere City FC1357
Ajax320
FC Omniworld2064
Feyenoord3069
Feyenoord3073
Feyenoord3064
Ajax333
Ajax1310
Ajax323
SC Heerenveen4765
Ajax964
ADO Den Haag199
Ajax419
Ajax892
Ajax336
Helmond Sport3514
Helmond Sport3532
Helmond Sport3510
Ajax596
Ajax418
Sparta4839
Haarlem3492
Haarlem3483
Haarlem3490
Ajax1112
AZ233
AZ247
AZ235
Ajax951
Ajax535
Ajax571
Ajax1114
Ajax424
Ajax425
Ajax1084
Ajax426
Ajax504
Ajax640
Ajax464
Ajax808
Ajax437
Ajax502
ADO Den Haag208
PSV4265
PSV4322
Ajax898
Ajax894
Ajax1079
Ajax472
Ajax1083
Ajax1120
Ajax1101
Ajax908
Sparta4812
Ajax1293
Nederland4016
Ajax549
Ajax403
Ajax597
Ajax574
Ajax589
Ajax413
Ajax459
Ajax404
Ajax820
Ajax906
Ajax328
Ajax343
Ajax492
Ajax558
Ajax585
Feyenoord2590
Ajax1025
Roda JC4637
AZ261
Ajax1291
Ajax1286
PSV4870
Ajax412
FC Utrecht2527
Vitesse5091
Ajax371
Ajax442
Ajax978
FC Utrecht2325
Ajax457
Ajax319
Ajax757
Ajax1141
Hollandia3617
Ajax1288
Haarlem3482
Ajax811
Ajax422
Ajax1184
Ajax579
Ajax874
Ajax708
Ajax1267
Ajax1140
Ajax498
Ajax867
Ajax1205
Ajax613
Ajax890
Ajax360
Ajax896
Ajax1093
Ajax1225
Ajax359
Ajax1262
Ajax367
Ajax593
Ajax824
Ajax462
Ajax411
Ajax455
Ajax1302
Ajax1325
Ajax479
NEC3913
Ajax506


Legenda
BEDREIGING
EENVOUDIGE MISHAN...
RUZIE/TWIST (ZONDER...
ZWARE MISHANDELING
STRAATROOF
OPENLIJKE GEWELDPLE...
GEKWAL. DIEFSTAL IN/...
Verkleinen
Graafdiepte: 1
Namen weergeven
Connor Beertsma
Eduard Bronda
Ethan Leisma
Friedrich Tiemersma
Sibren Eenshuistra
Andreas Atema
Marco Pottinga
Roy Roersma
Tjibbe Tasma
Stanley Melinga
Tom Biersma
Harco Sijnja
Wietse Banstra
Youri Wubbema
Rudi Munniksma
Rolf Elenga
Persoon4012
Brent Roijinga
Elbert Udenga
Feike Siebinga
Roland Benninga
Shimon Folbarda
Persoon3617
Phillip Zwikstra
Reuven Linstra
Emmanuel Ynema
Edward Teensma
Matthijs Rinsema
Adrianus Ullersma
Marco Lemstra
Norwin Reitsema
Nikolai Rensma
Goran Sijtsinga
Gerbrand Veen
Gérard Wiglema
Nikita Struikstra
Lennaert Spijkstra
Arne Rijpema
Otsman Asma
Ricky Rollema
Brent Bunnema
Yassine Reukema
Mustafa Harsma

Community detection — very brief introduction

Related: clustering

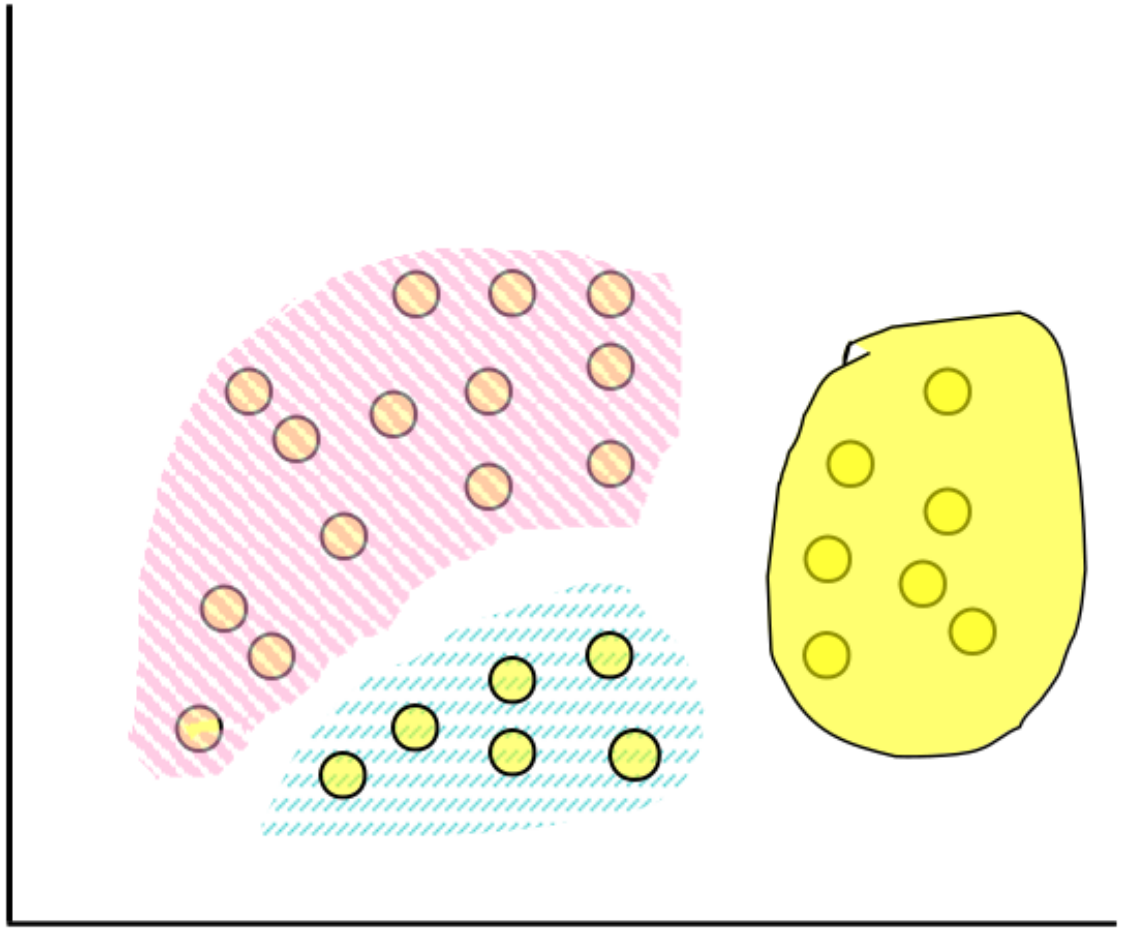


Image: KDnuggets - Clustering, 2009.

Community detection

- **Community**: subset of nodes connected more strongly with eachother than with the rest of the network
- Community detection algorithms:
  - Clique-based methods
  - Divisive algorithms (centrality-based)
  - Label propagation algorithms
  - Random walk algorithms
  - **Modularity maximization** algorithms

Community detection

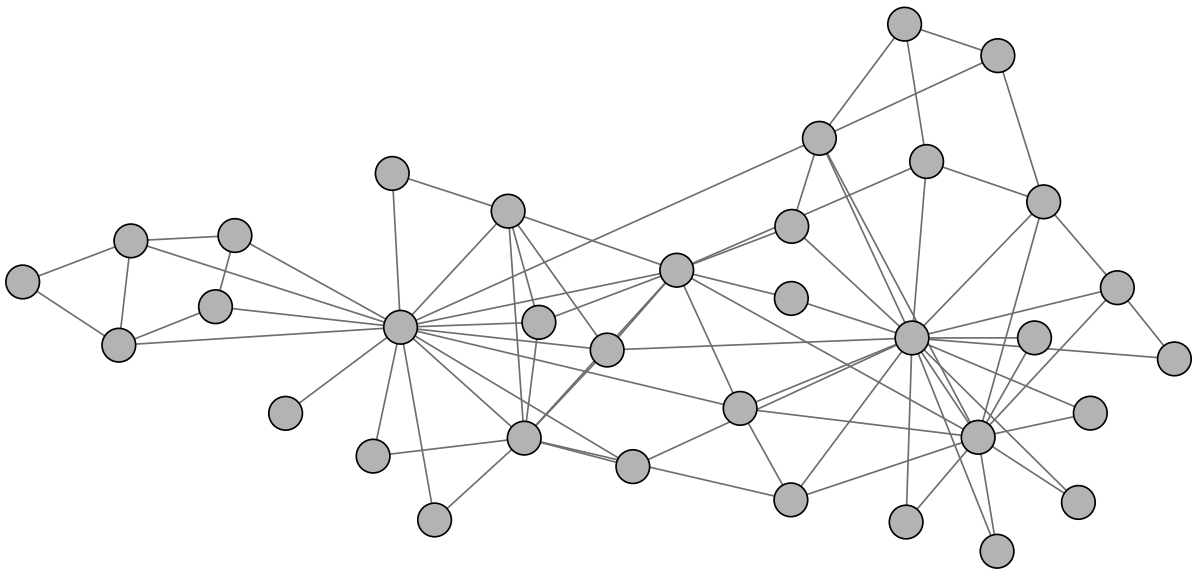


Figure: Communities: node subsets connected more strongly with each other

Community detection

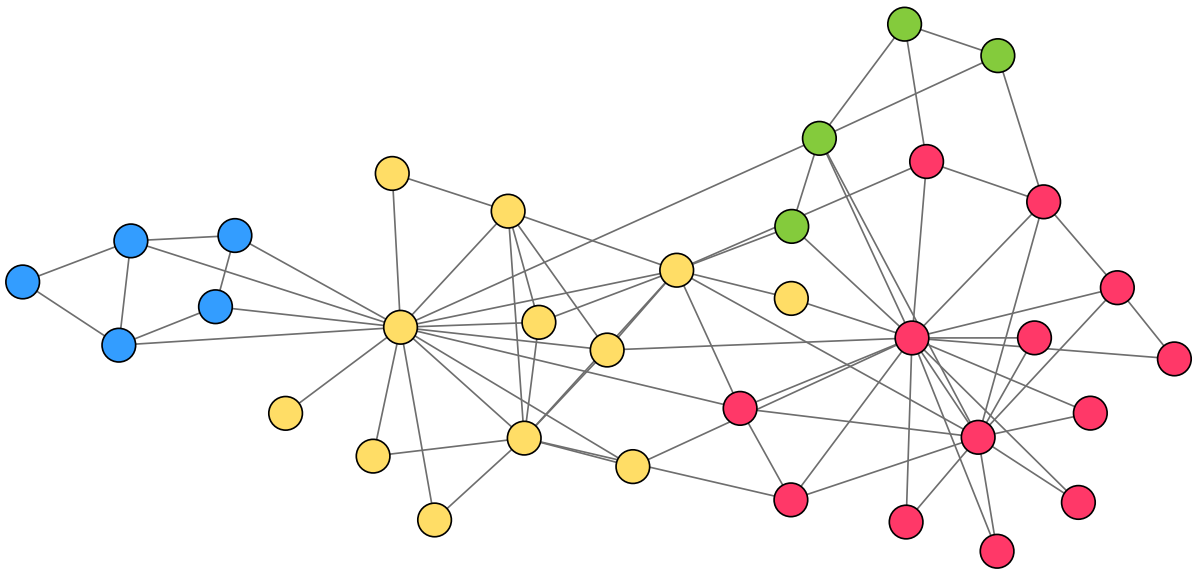


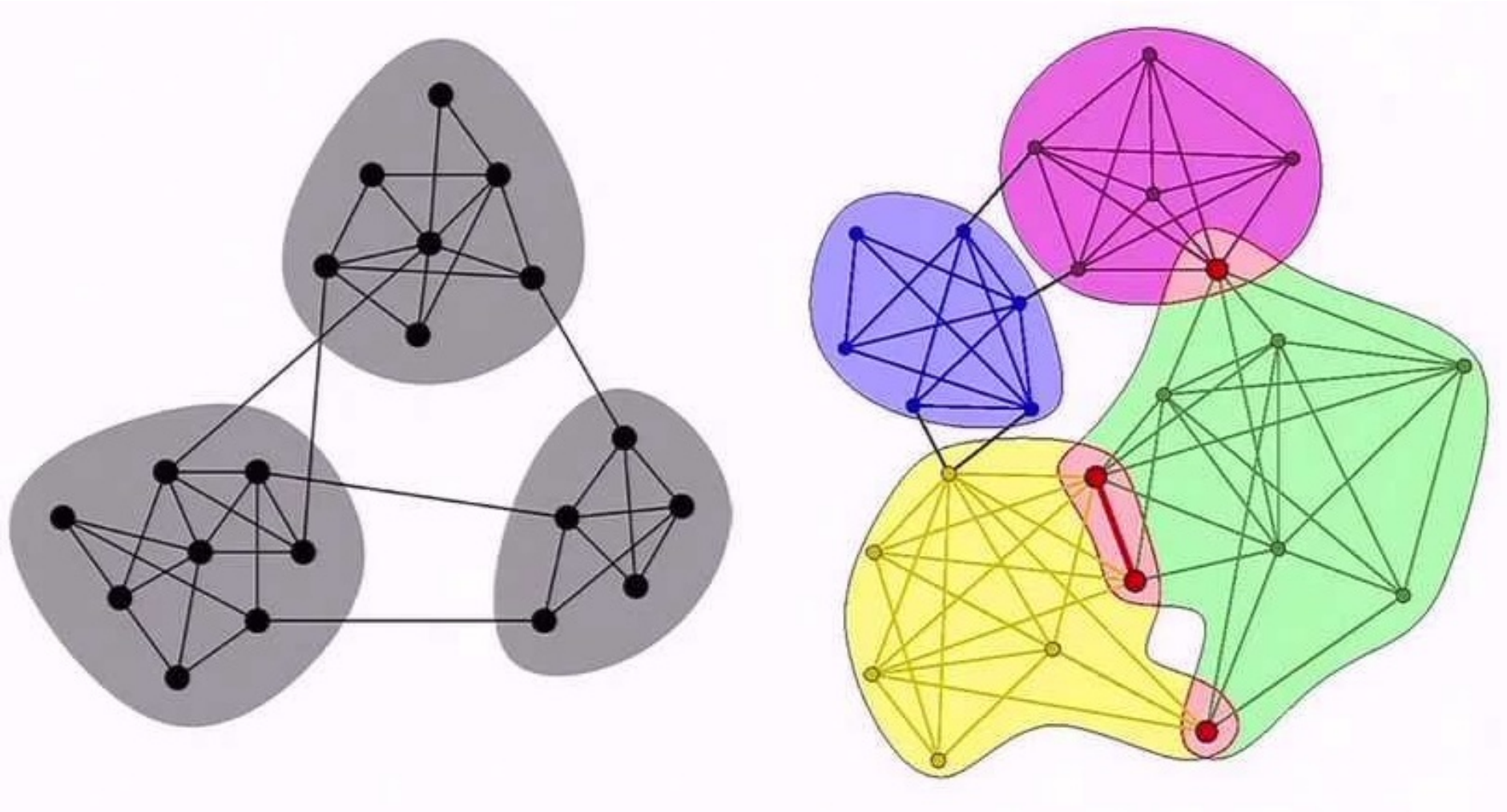
Figure: Communities: node subsets connected more strongly with each other

Modularity

- **Community** (alternative definition): subset of nodes for which the fraction of links inside the community is higher than expected
- **Modularity**: numerical value Q indicating the quality of a given division of a network into communities. Higher value of Q means more links within communities (and fewer between)
- Resolution parameter r indicating how "tough" the algorithm should look for communities
- Algorithms optimize (maximize) the modularity score Q given some r (using local search, heuristics, hill climbing, genetic algorithms or other optimization techniques)

V.D. Blondel, J-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks in *Journal of Statistical Mechanics: Theory and Experiment* 10: P10008, 2008.

Partitions vs. communities



J. Leskovec, Affiliation Network Models for Densely Overlapping Communities, MMDS 2012.

Upcoming lab session & Homework for next week

- **Today:** stick around if you are already certain that you will take the course, and want to find a teammate for the project already
- Lab session: Introduction to NetworkX
- Work on Assignment 1

- "Homework": Make serious progress with Assignment 1
- Make choice of participation in course explicit. Un-enroll if you wish to drop the course