

Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

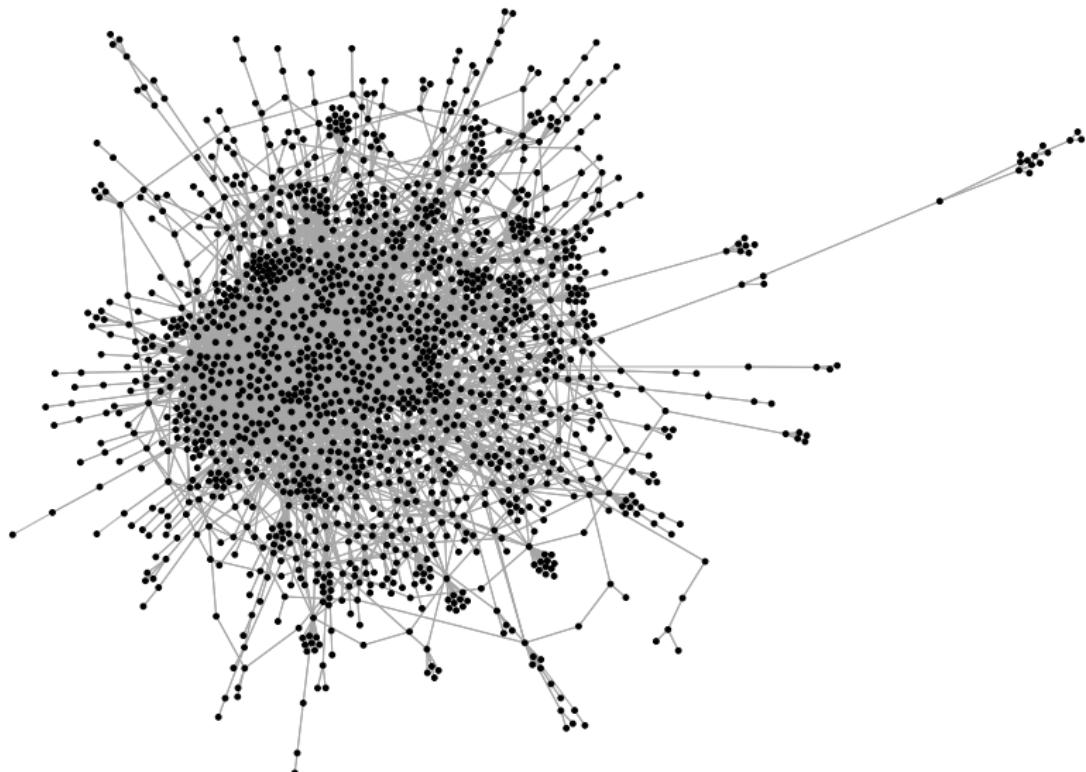
Lecture 3 — Network projection and community structure

So ...

- You are registered for the course (deregistration deadline passed)
- If necessary, you are (in the process of) catchup on missing background knowledge using the “Reading material” on the website
- You understand how Gephi works
- You know how to use Python NetworkX
- You have made serious progress with Assignment 1

Recap

Networks



Notation

Concept

- Network (graph)
- Nodes (objects, vertices, ...)
- Links (ties, relationships, ...)
 - Directed — $E \subseteq V \times V$ — "links"
 - Undirected — "edges"
- Number of nodes — $|V|$
- Number of edges — $|E|$
- Degree of node u
- Distance from node u to v

Symbol

- $G = (V, E)$
- V
- E
- n
- m
- $\deg(u)$
- $d(u, v)$

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

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 (Erdős, Kevin Bacon), protein interaction networks, information networks (Wikipedia), webgraphs, financial networks (Bitcoin) ...

Advanced concepts

- Assortativity, homophily
- Reciprocity
- Power law exponent
- Planar graphs
- Complete graphs
- Subgraphs
- Trees
- Spanning trees
- Diameter, eccentricity
- Bridges
- Graph traversal: DFS, BFS

Centrality measures

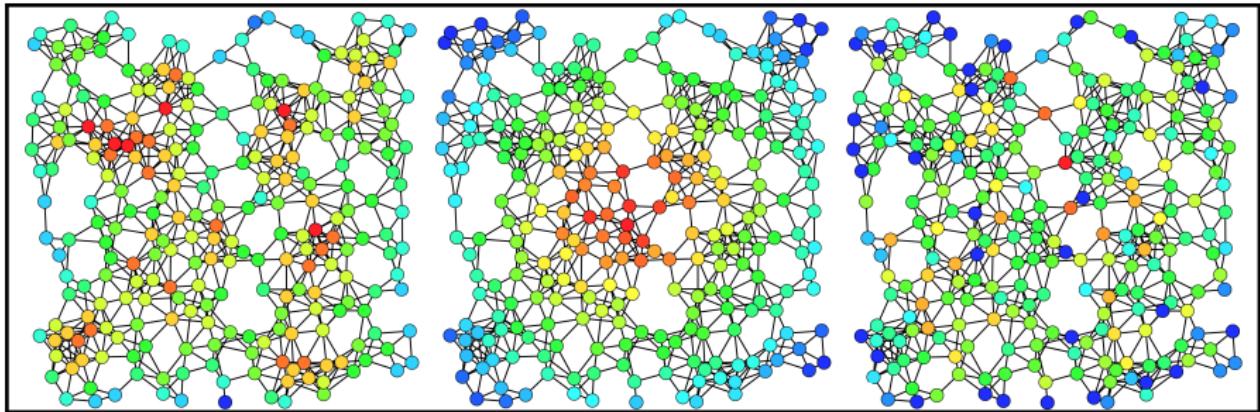


Figure: Degree, closeness and betweenness centrality

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

Today

- Network projection
- Case: Criminal networks
- Finding dense subgraphs
- Community detection
- Case: Corporate networks
- Advanced community detection
- Example presentation (leftover from last week)

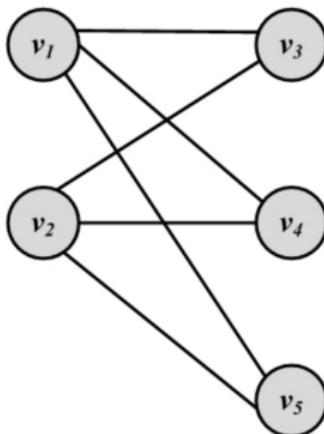


Network projection

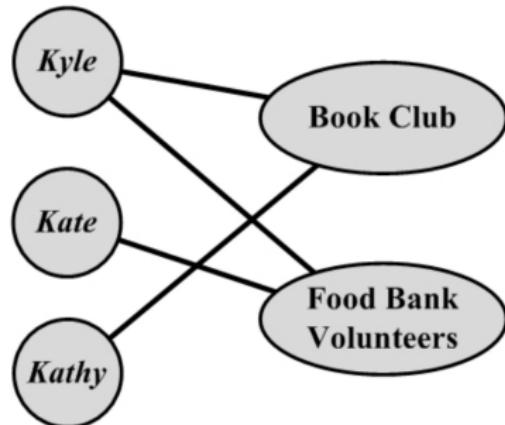
Bipartite graphs

- In bipartite graphs the set 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 **heterogeneous networks** (as opposed to respectively one-mode networks and homogeneous 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

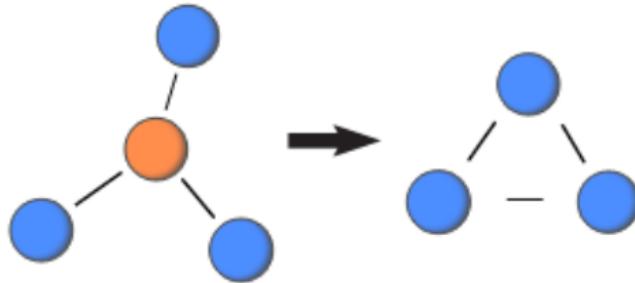
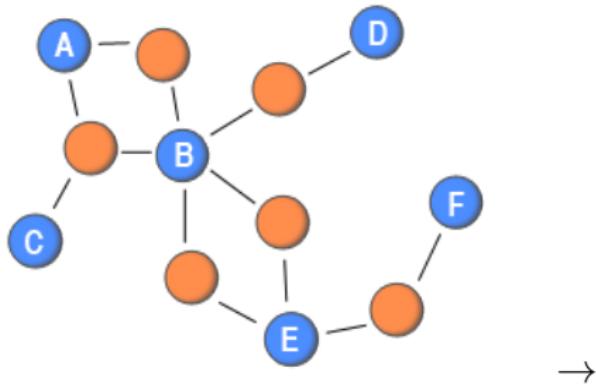


Image: <http://toreopsahl.com/>

Weighted projection



Weighted projection

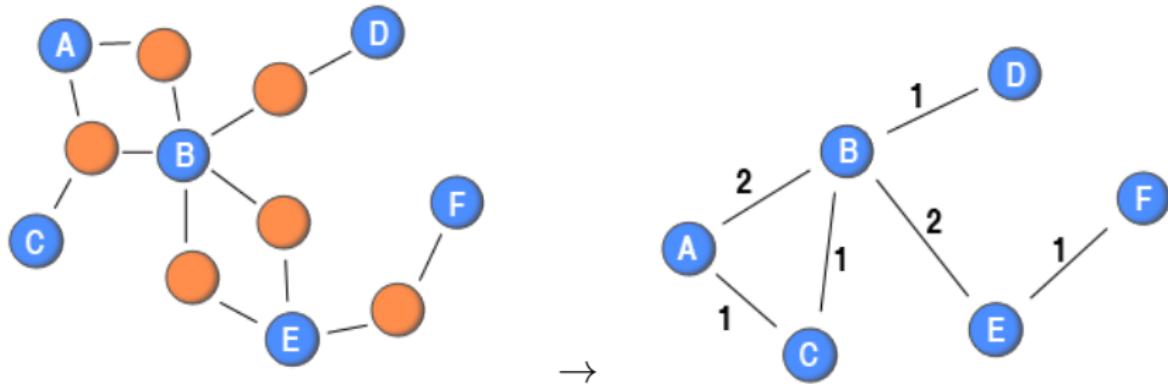
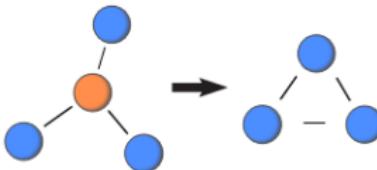


Image: <http://toreopsahl.com>

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

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

Clubs

- ADO Den Haag
- Ajax
- Feyenoord
- FC Utrecht
- PSV

[Meer...](#)

Relaties

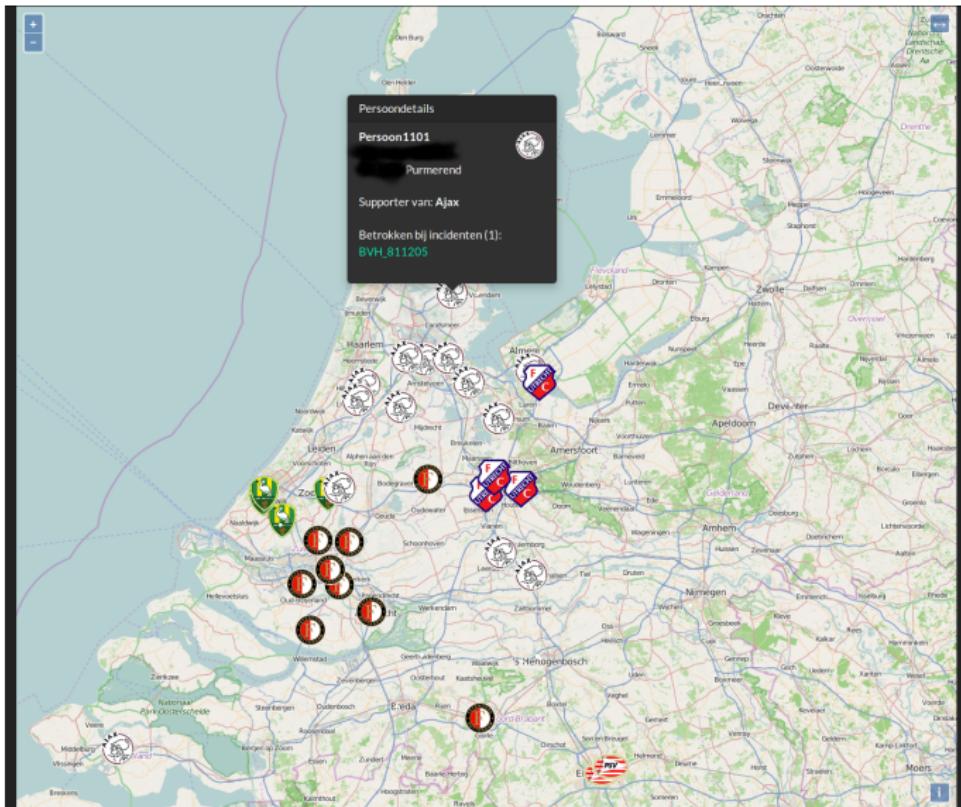
- Links VVS
- Links BVH

[Meer...](#)

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



Dit app 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

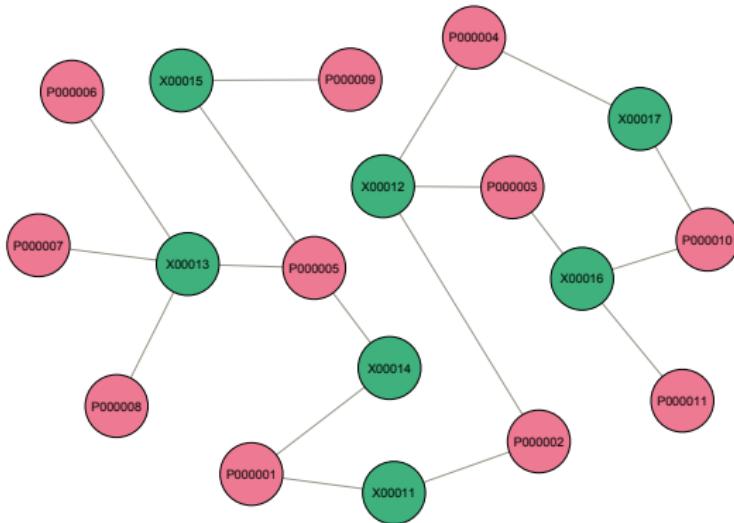
Table: Data on suspects involved in incidents

Network formation



Figure: Suspects are nodes

Two-mode criminal network



Network formation



Figure: Edges are based on common involvement as a suspect

Network visualization

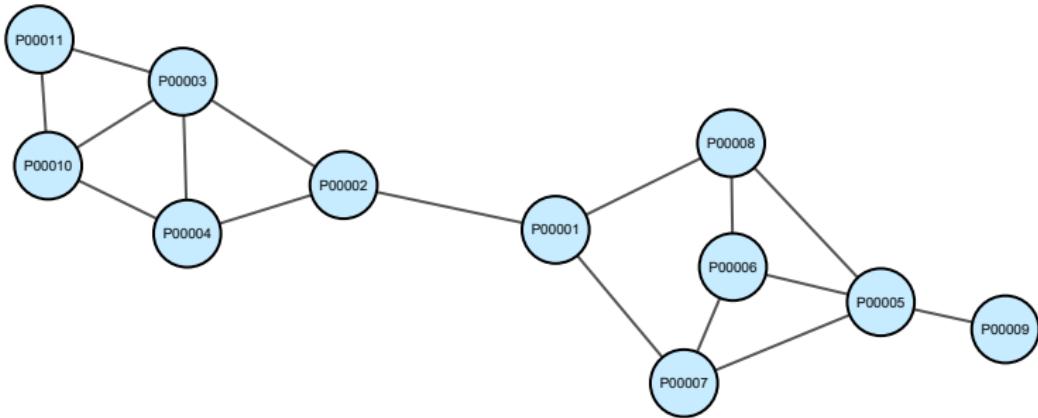


Figure: Force-directed visualization algorithm reveals structure

Network analysis: Centrality

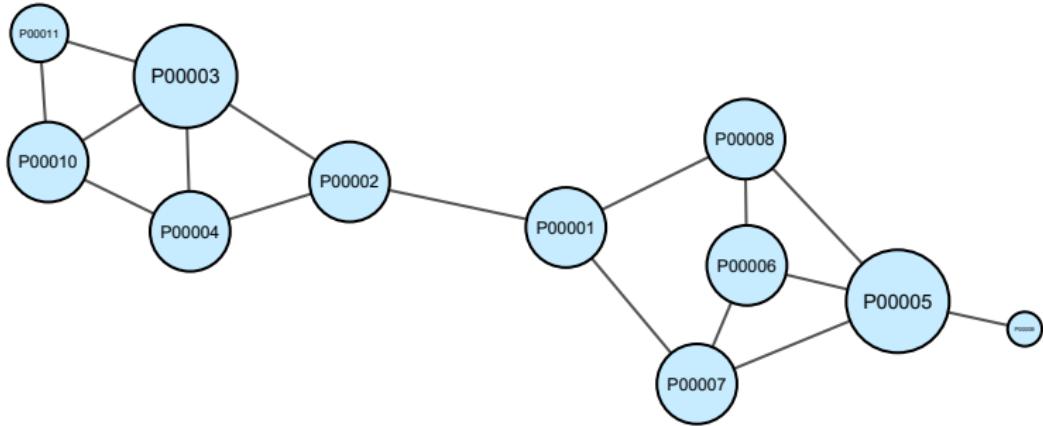


Figure: Degree centrality finds locally important nodes

Network analysis: Centrality

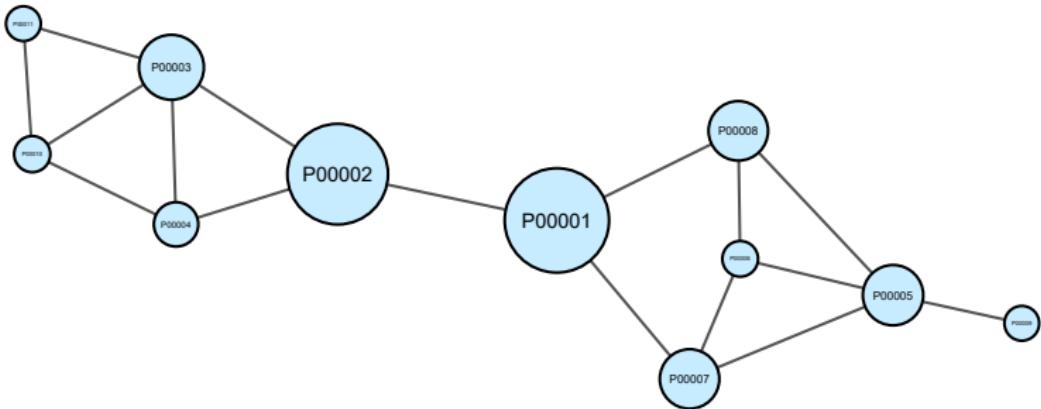


Figure: Betweenness centrality reveals globally important nodes

Network analysis: Community detection

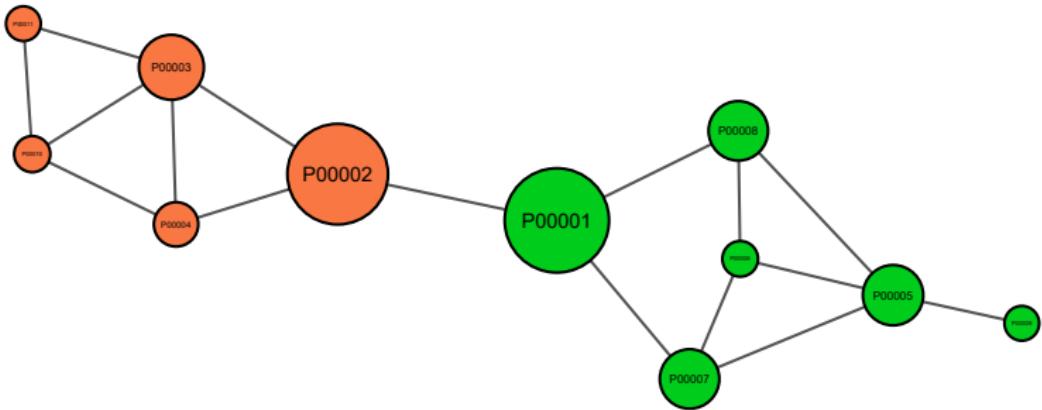
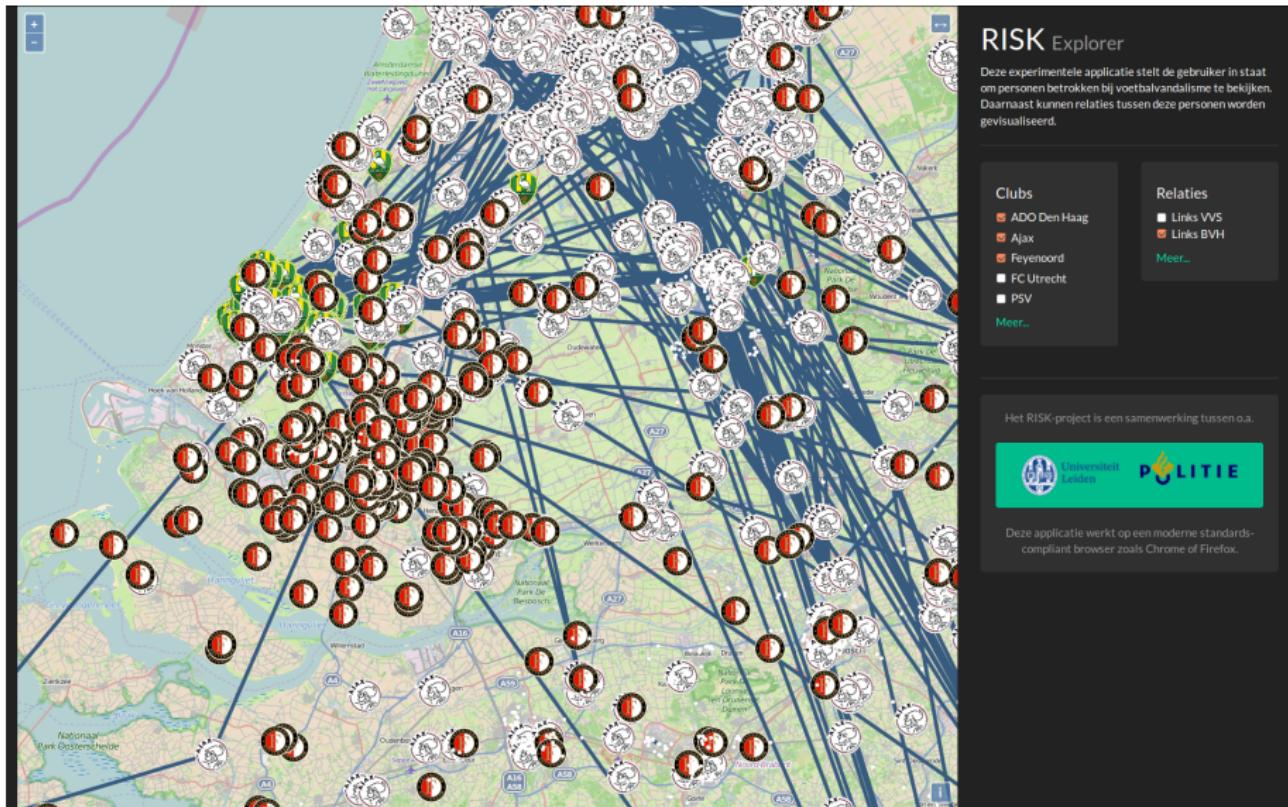
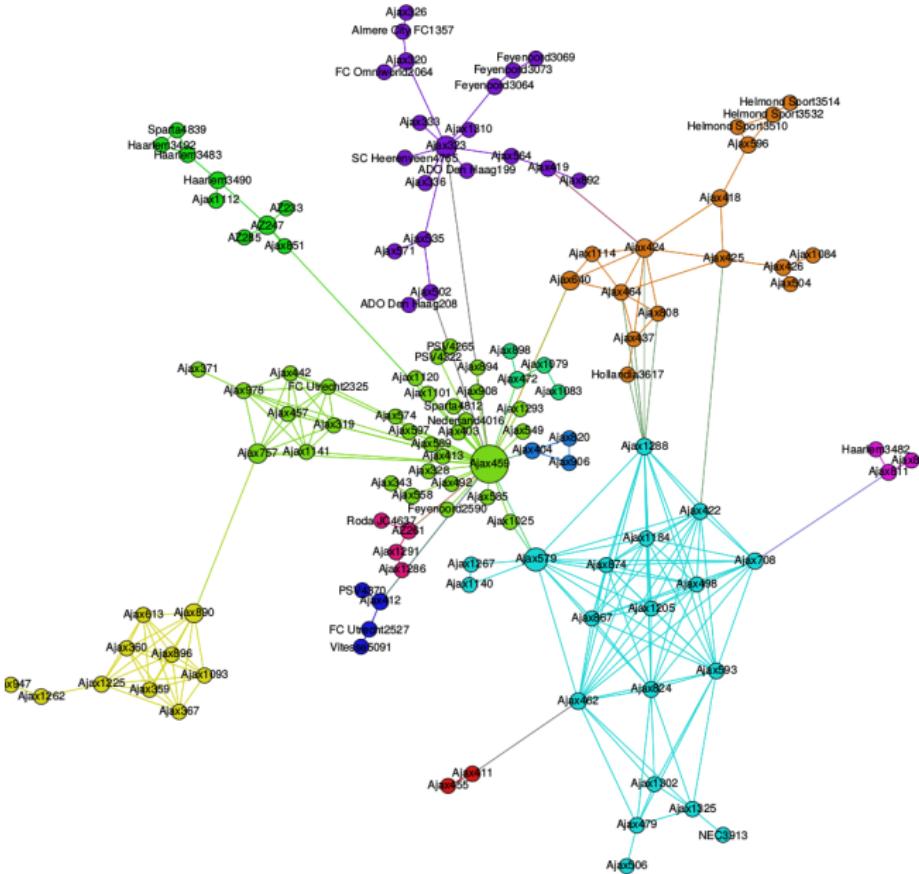


Figure: Community detection finds groups of tightly connected nodes

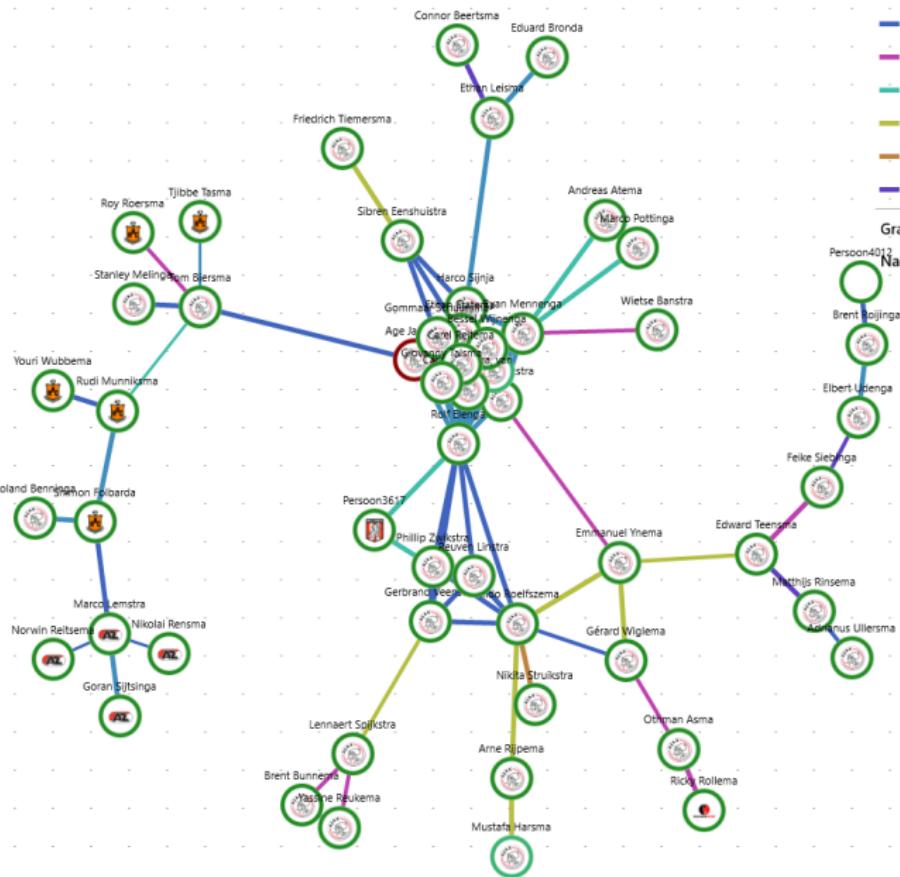




Legenda

- BEDREIGING
- EENVOUDIGE MISHANDELING
- RUIZIE/TWIST (ZONDER...)
- ZWARTE MISHANDELING
- STRAATROOF
- OPENLIJKE GEWELDPLEG...
- GEKWAL. DIEFSTAL IN/...

Graafdiepte: 1
Personen 4012
Namen weergeven



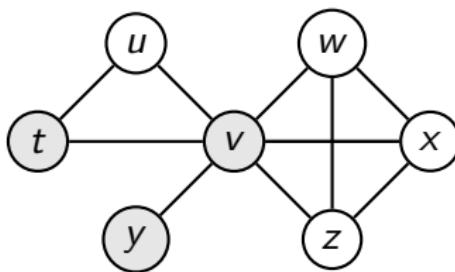


Dense subgraphs and cliques

Subgraphs and cliques

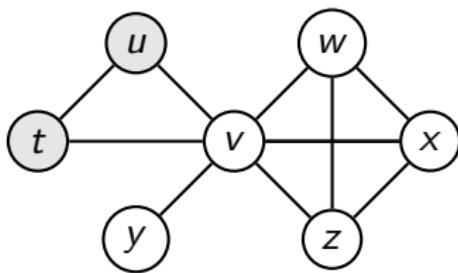
- Subgraph $G' = (V', E')$ of a graph $G = (V, E)$ with $V' \subseteq V$ and $E' \subseteq E$
- Complete graph: graph with all edges present
- **Clique**: complete subgraph
- Maximal clique: complete subgraph of maximal size (cannot be extended with another node)
- Maximum clique: largest possible clique in the graph
- **Clique problem**: find the maximum clique(s) of a graph
(one of Karp's 21 NP-complete problems introduced in 1972)

Subgraphs and cliques



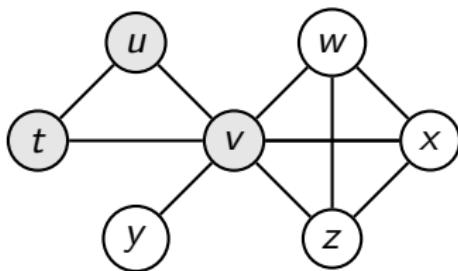
- The subgraph with $V' = \{t, v, y\}$ is not complete

Subgraphs and cliques



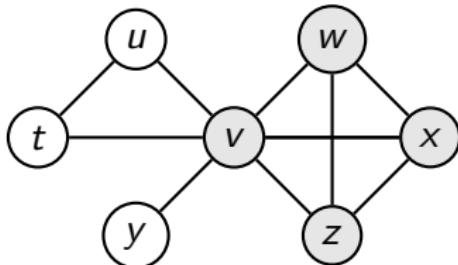
- The subgraph with $V' = \{t, v, y\}$ is not complete
- **The subgraph with $V'' = \{t, u\}$ is complete and thus a clique of size 2**

Subgraphs and cliques



- The subgraph with $V' = \{t, v, y\}$ is not complete
- The subgraph with $V'' = \{t, u\}$ is complete and thus a clique of size 2
- **The subgraph with $V''' = \{t, u, v\}$ is a maximal clique of size 3**

Subgraphs and cliques



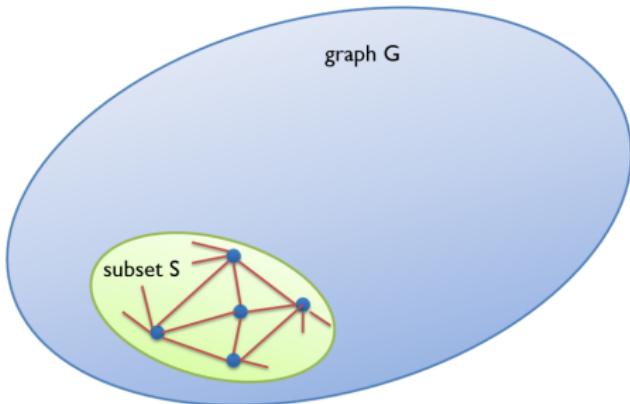
- The subgraph with $V' = \{t, v, y\}$ is not complete
- The subgraph with $V'' = \{t, u\}$ is complete and thus a clique of size 2
- The subgraph with $V''' = \{t, u, v\}$ is a maximal clique of size 3
- **The subgraph with $V'''' = \{v, w, x, z\}$ is a maximum clique of size 4**

Dense subgraphs

- Density: $m/n(n - 1)$ (value $\in [0; 1]$)
- Complete graphs have density 1
- Density (alternative): m/n (value $\in [0; n - 1]$)
- Dense subgraph: subgraph with a high density (almost a clique)
- **Densest subgraph problem:** find a densest subgraph in an undirected graph (NP-hard; reduction from the clique problem)
- Applications in community detection
- Many (exponential) algorithms have been proposed

A.V. Goldberg, "Finding a maximum density subgraph", Technical report, 1984.

Densest subgraph



- S is a subgraph of G
- S is not a clique, but it is dense (8 out of 10 edges present)

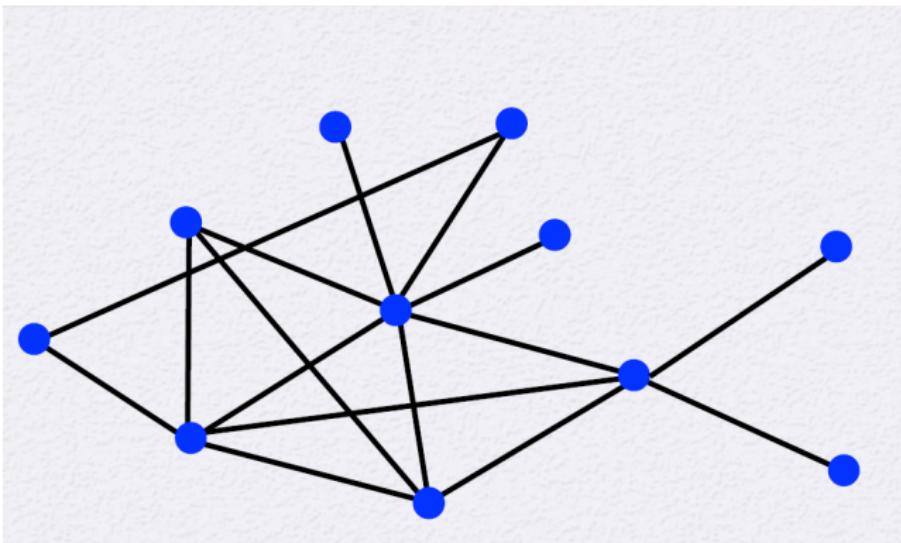
Image: <http://www.cs.princeton.edu/~zdvir/apx11slides/>

Iterative algorithm

- Density (alternative): m/n (value $\in [0; n - 1]$)
- Simple iterative algorithm to find densest subgraph by Charikar et al.:
 - 1 Compute the average degree of the graph
 - 2 Delete all nodes whose degree is below the average
 - 3 Keep track of the density at each step
 - 4 Go to step 1 if nodes were deleted in this iteration
 - 5 Output the densest graph seen over all iterations

Charikar et al., "Greedy approximation algorithms for finding dense components in a graph",
in LNCS 1913, pp. 84–95, 2000.

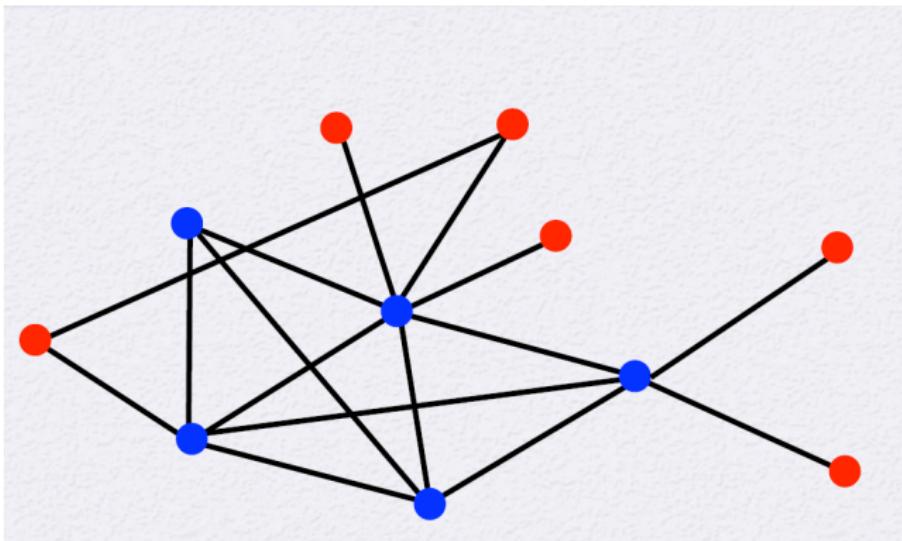
Iterative algorithm



- Iteration 1: Current density $16/11 = 1.45$, avg. degree = 2.9
- Best density (iteration) = ...

Source: B. Bahmani et al., Densest subgraphs in streaming and MapReduce, in VLDB 5(5), pp. 454–465, 2012.

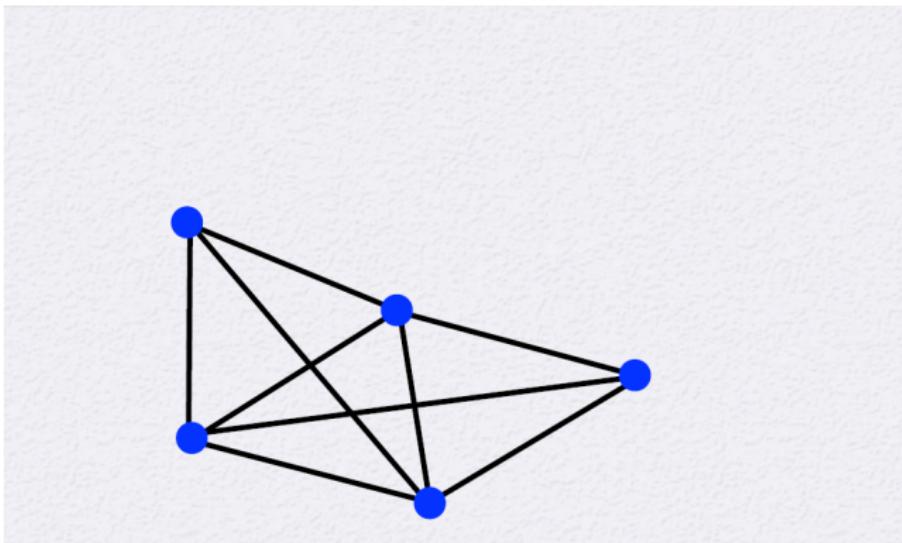
Iterative algorithm



- Iteration 1: Current density $16/11 = 1.45$, avg. degree = 2.9
- **Best density (iteration) = 1.45 (1)**

Source: B. Bahmani et al., Densest subgraphs in streaming and MapReduce, in VLDB 5(5), pp. 454–465, 2012.

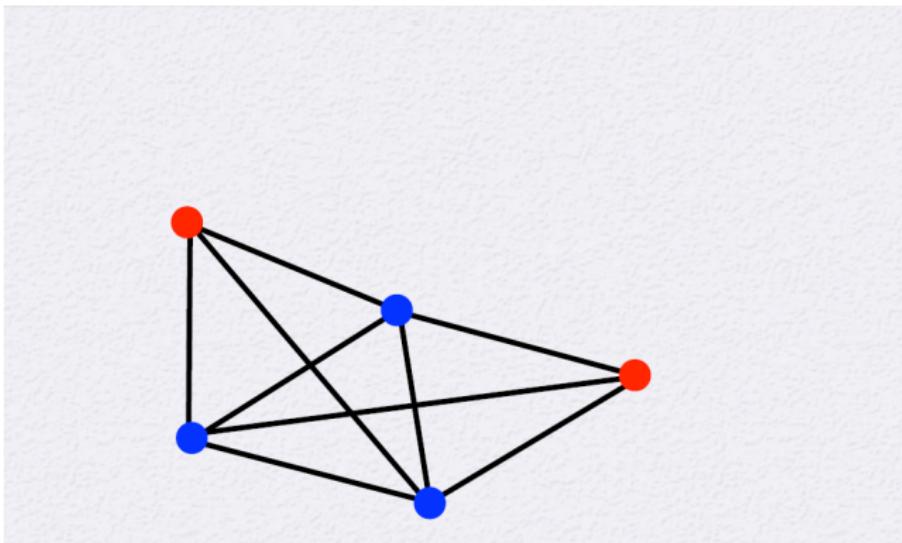
Iterative algorithm



- **Iteration 2: Current density $9/5 = 1.8$, avg. degree = 3.6**
- Best density (iteration) = 1.45 (1)

Source: B. Bahmani et al., Densest subgraphs in streaming and MapReduce, in VLDB 5(5), pp. 454–465, 2012.

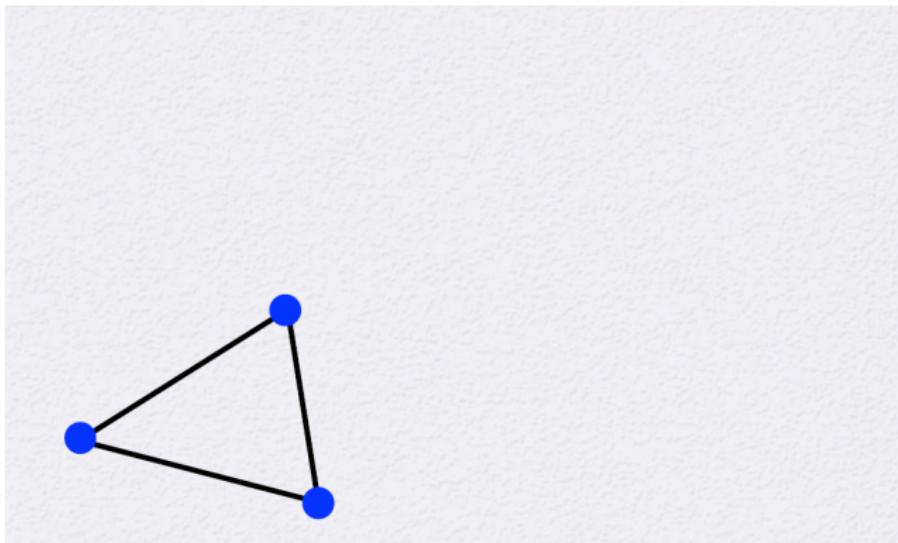
Iterative algorithm



- Iteration 2: Current density $9/5 = 1.8$, avg. degree = 3.6
- **Best density (iteration) = 1.8 (2)**

Source: B. Bahmani et al., Densest subgraphs in streaming and MapReduce, in VLDB 5(5), pp. 454–465, 2012.

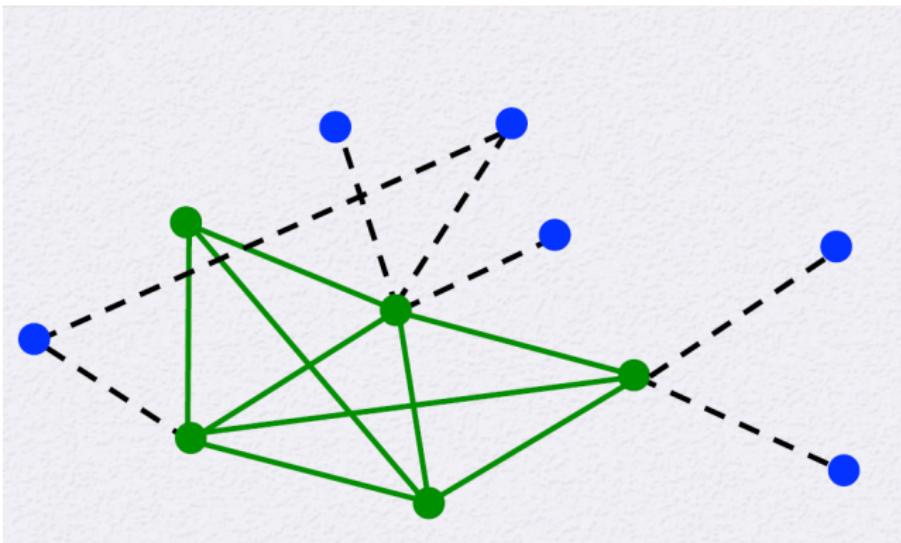
Iterative algorithm



- **Iteration 3: Current density $3/3 = 1$, avg. degree = 2**
- Best density (iteration) = 1.8 (2) (unchanged)

Source: B. Bahmani et al., Densest subgraphs in streaming and MapReduce, in VLDB 5(5), pp. 454–465, 2012.

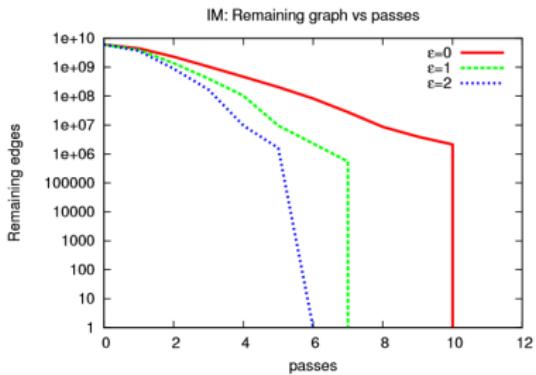
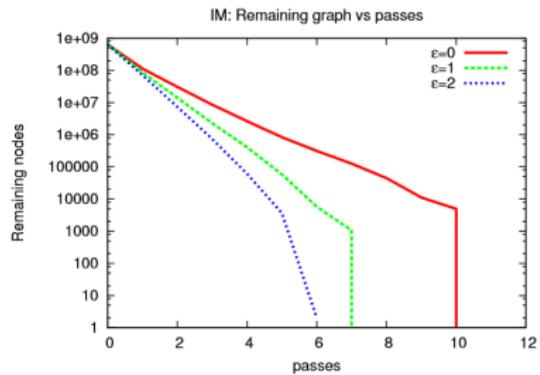
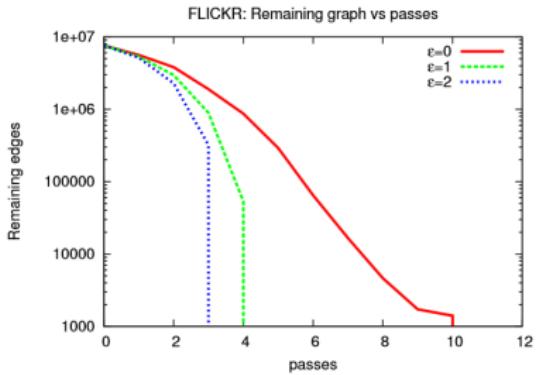
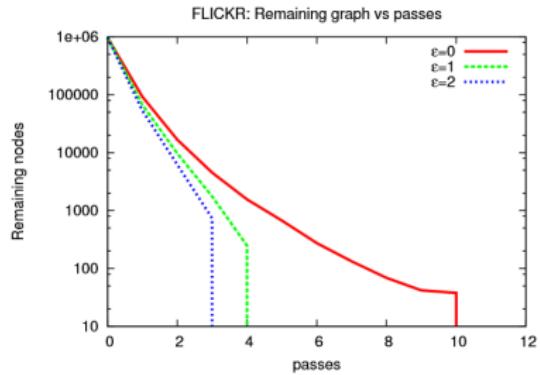
Iterative algorithm



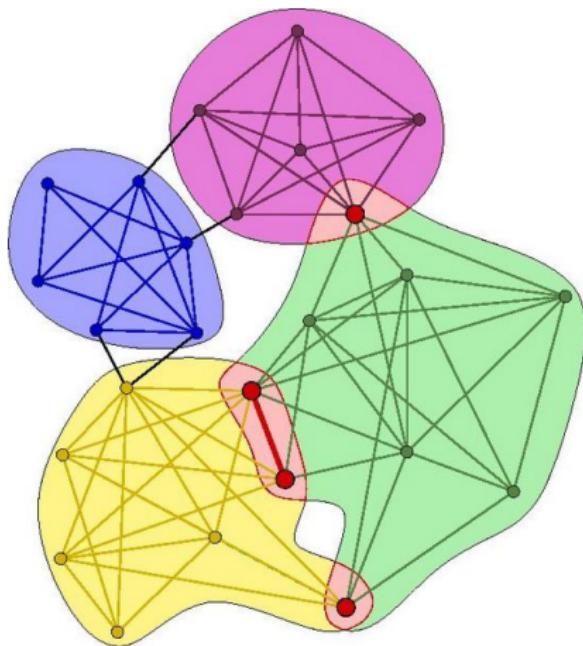
- No node deleted in previous iteration, algorithm terminates
- Best density (iteration) = **1.8 (2)**

Source: B. Bahmani et al., Densest subgraphs in streaming and MapReduce, in VLDB 5(5), pp. 454–465, 2012.

Iterative algorithm performance



Dense subgraphs for community detection



Source: J. Leskovec, Community detection workshop (1), 2014.



Community detection

Related: clustering

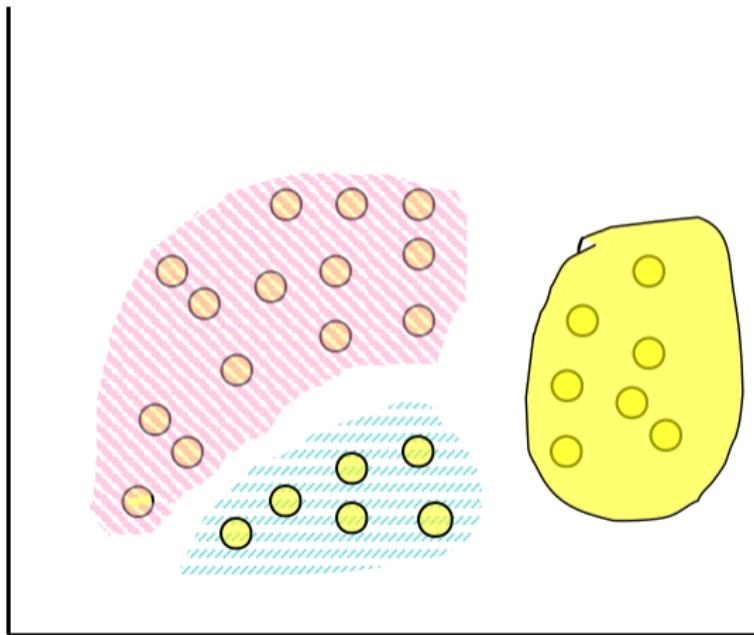


Image: KDnuggets - Clustering, 2009.

Community detection

- **Community:** subset of nodes connected more strongly with each other 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

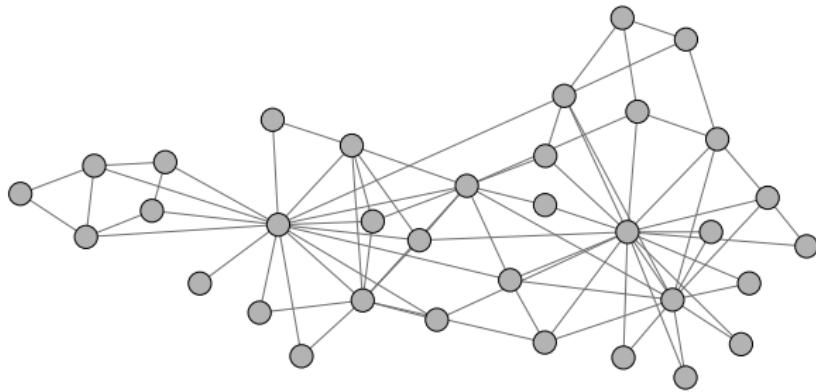


Figure: Communities: node subsets connected more strongly with each other

Community detection

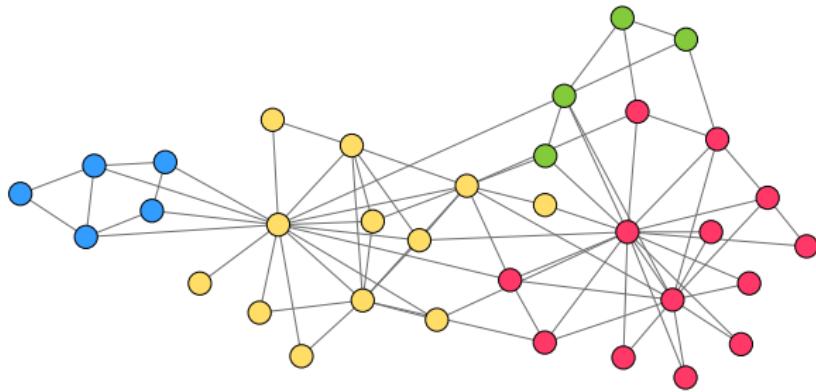


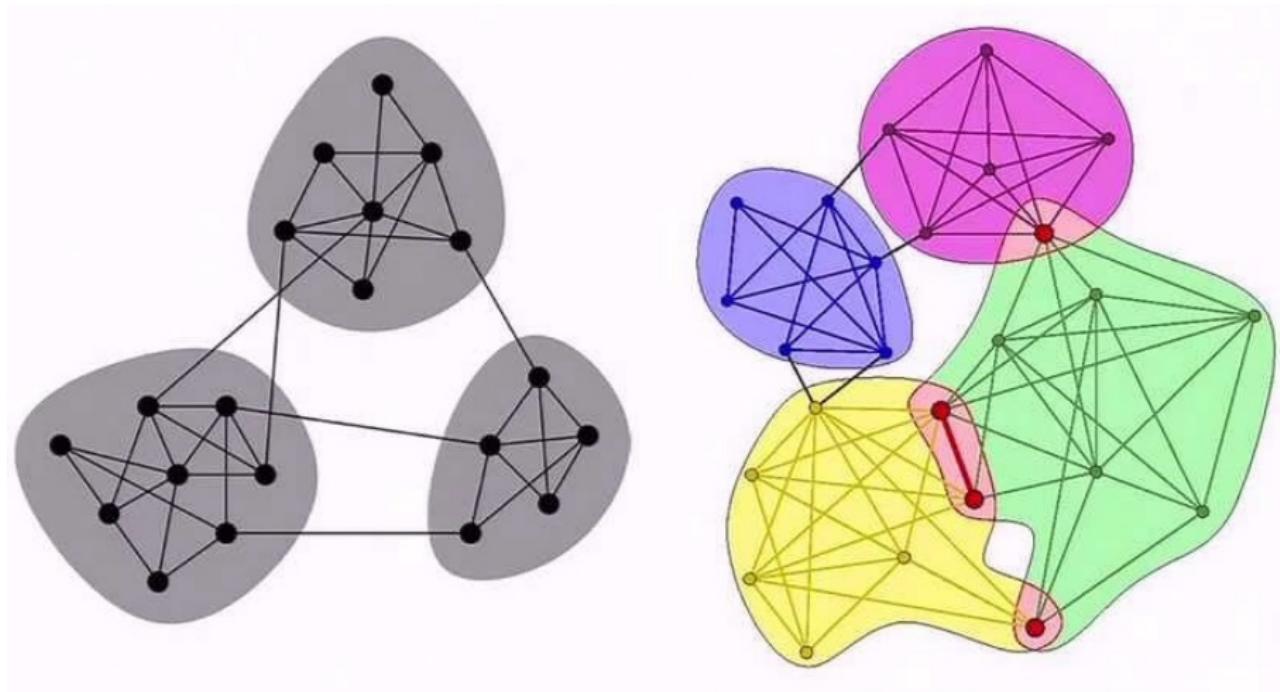
Figure: Communities: node subsets connected more strongly with each other

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.

Yet another network

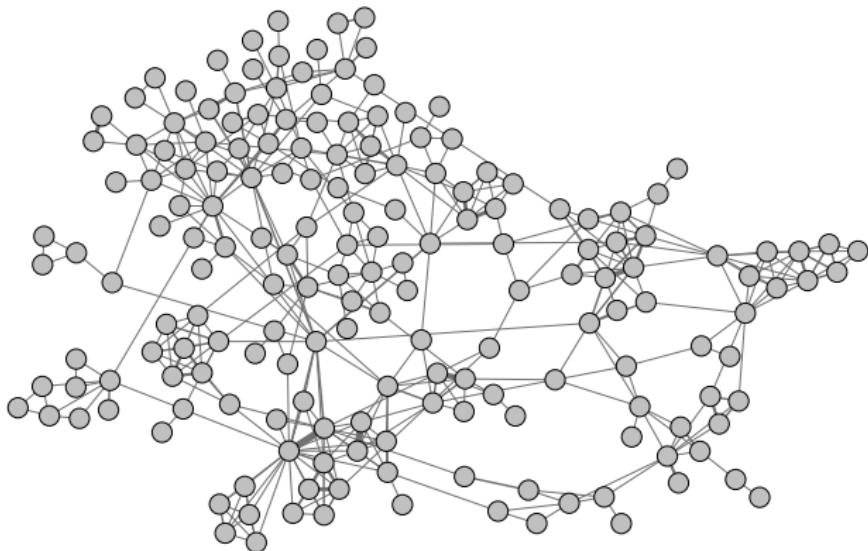


Figure: Do you see communities?

Resolution = 1.0

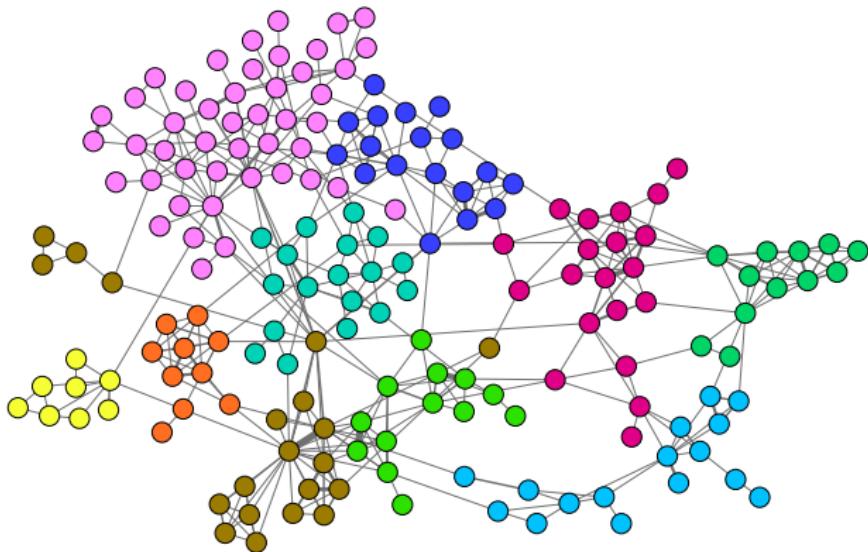


Figure: Modularity = 0.747; 10 communities

Resolution = 2.0

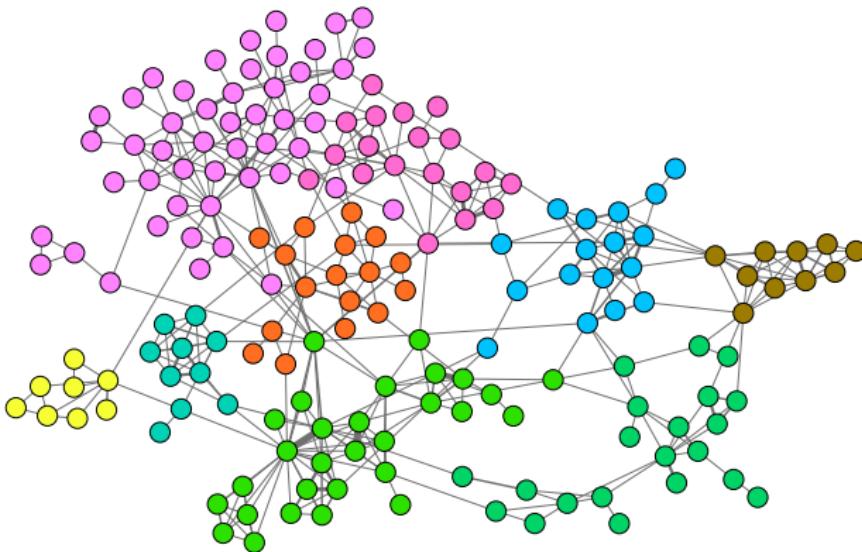


Figure: Modularity = 0.732; 8 communities

Resolution = 4.0

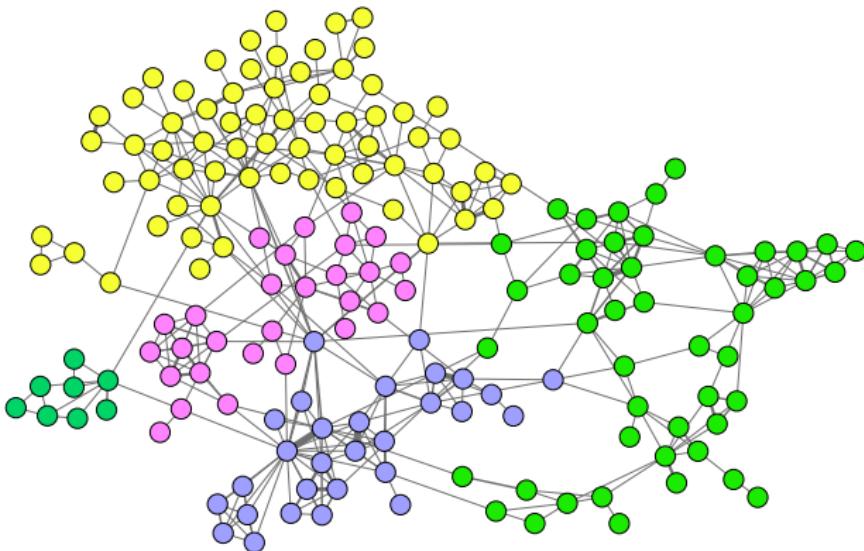


Figure: Modularity = 0.641; 5 communities

Evaluating communities and partitions

- **Communities:** groups of nodes that are more connected amongst each other than with the other nodes of the network
- **Partitions:** non-overlapping communities
- Compare with groups of nodes based on common attributes
- Human interpretation by hand can suffer from subjective bias

Corporate board interlocks

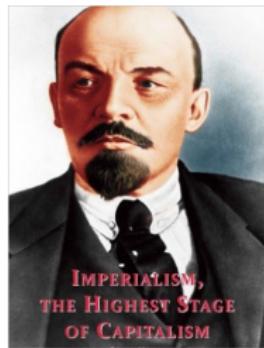
- **Nodes** are organizations/firms/companies/corporations

Corporate board interlocks

- **Nodes** are organizations/firms/companies/corporations
- **Edge** are **board interlocks**: relationships between firms because they share a board member or director

Corporate board interlocks

- **Nodes** are organizations/firms/companies/corporations
- **Edge** are **board interlocks**: relationships between firms because they share a board member or director
- Vladimir I. Lenin, *Imperialism, The Highest Stage of Capitalism*, 1916.
- "... a personal union, so to speak, is established between the banks and the biggest industrial and commercial enterprises, the merging of one with another through the acquisition of shares, through the appointment of bank directors to the Supervisory Boards (or Boards of Directors) of industrial and commercial enterprises, and vice versa."



Board interlocks

■ Causes of interlocks:

- Collusion
- Cooptation and monitoring
- Legitimacy
- Career advancement
- Social cohesion

■ Consequences of interlocks:

- Corporate control
- Economic performance
- Access to resources

Board interlocks

■ Causes of interlocks:

- Collusion
- Cooptation and monitoring
- Legitimacy
- Career advancement
- Social cohesion

■ Consequences of interlocks:

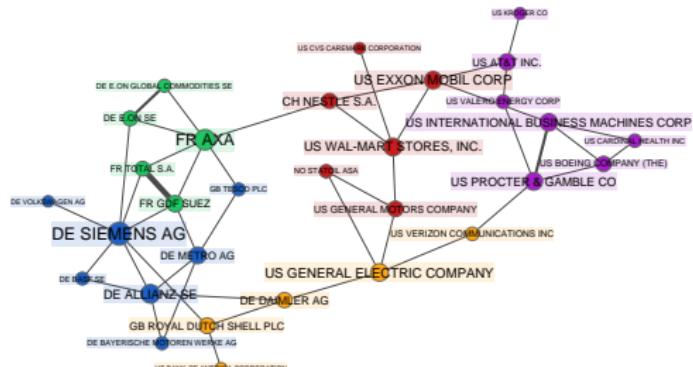
- Corporate control
- Economic performance
- Access to resources



M. Mizruchi, What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates, *Annual review of Sociology* 22: 271–298, 1996.

Corporate networks at city level

- Nodes are cities
- Edges between cities are based on firms sharing directors (interlocks)
- Weights on edges denote the number of connections
- Each city has an associated country
- Provides insight in geographical orientation of global economy, globalization, and existence of local and/or global business groups



Corporations



Board interlock network



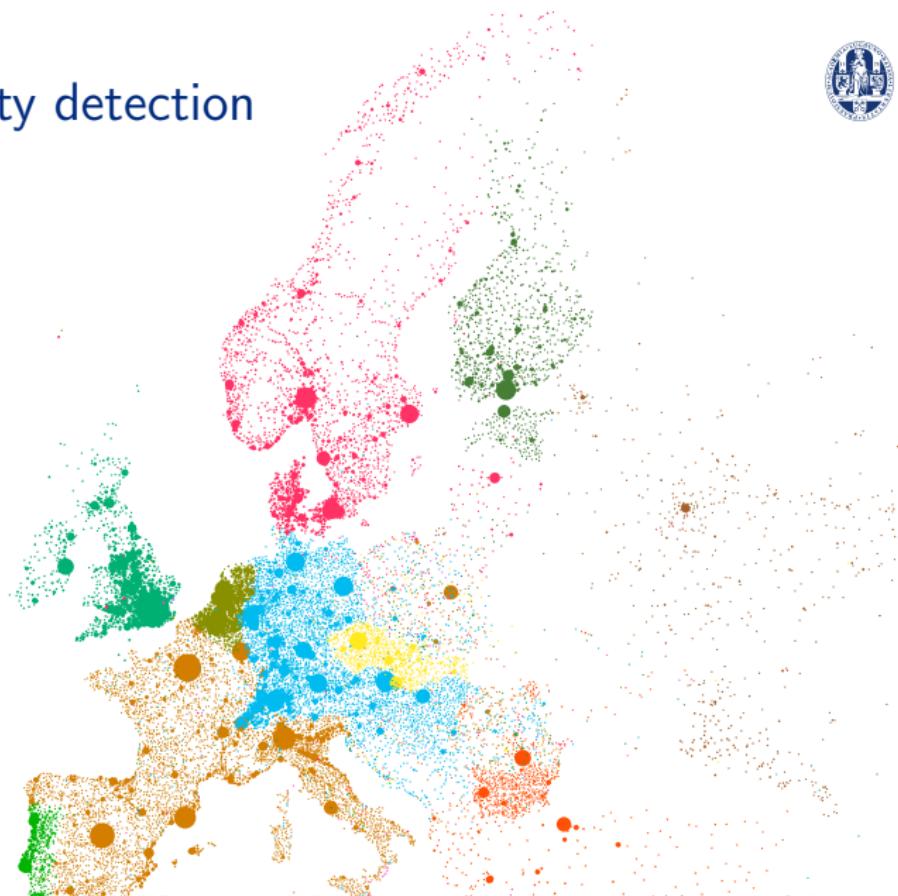
F.W. Takes and E.M. Heemskerk, Centrality in the Global Network of Corporate Control, *Social Network Analysis and Mining* 6(1): 1-18, 2016.

Community detection



E.M. Heemskerk and F.W. Takes, The Corporate Elite Community Structure of Global Capitalism, *New Political Economy* 21(1): 90-118, 2016.

Community detection



E.M. Heemskerk, F.W. Takes, J. Garcia-Bernardo and M.J. Huijzer, Where is the global corporate elite? A large-scale network study of local and nonlocal interlocking directorates, *Sociologica* 2016(2): 1-31, 2016.



Example presentation



Now and the upcoming lab session

- Team formation; stick around today!
- Choose your topic (in Brightspace, under “Groups”) before October 1
- Lab session this week: Finalize Assignment 1
- Ask all your final questions