Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

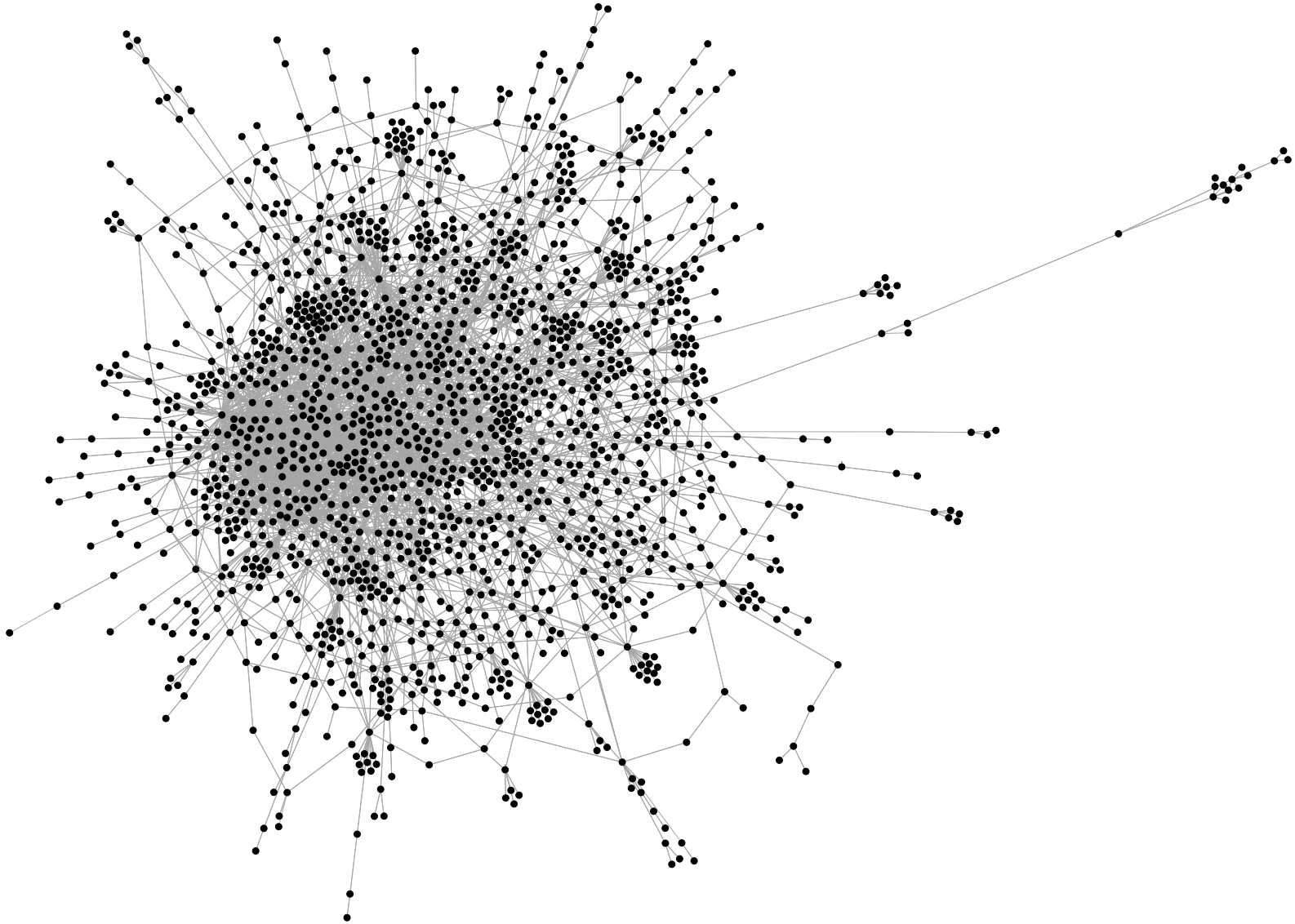
Lecture 3 — Network projection and propagation-based centrality

Today

- Recap
- Network projection
- Propagation-based centrality
- Course project explanation
- Example presentation

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 — $\lvert V \rvert$	n
Number of edges — $\lvert E \rvert$	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

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

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

Centrality measures

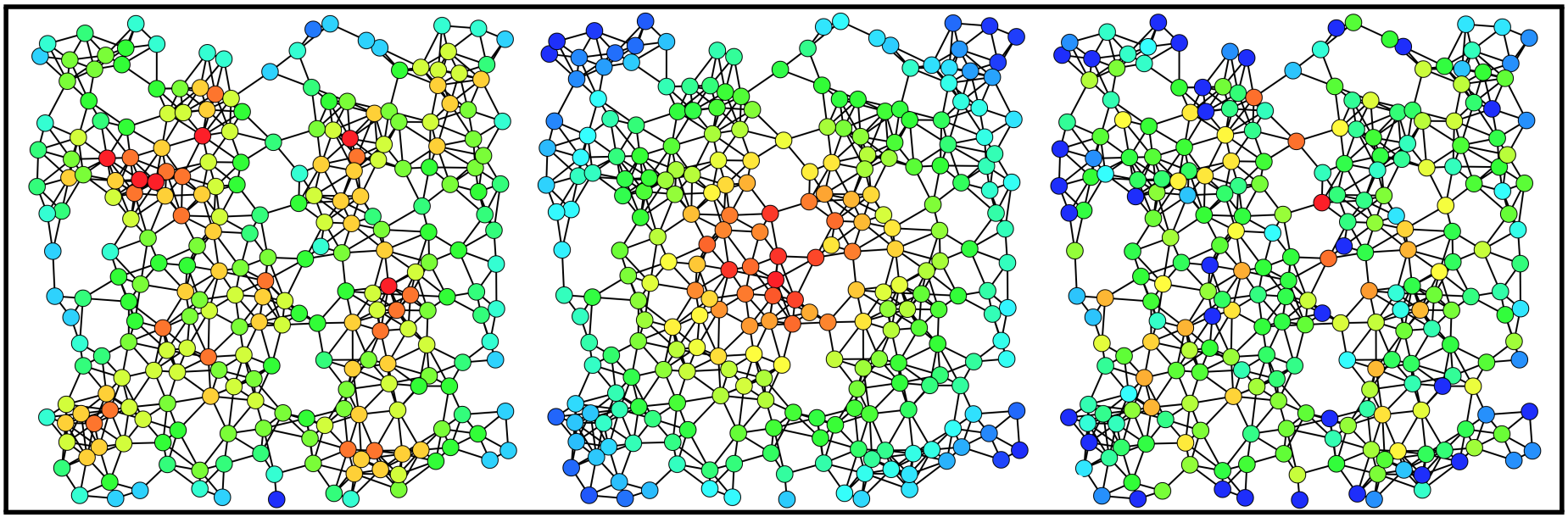


Figure: Degree, closeness and betweenness centrality

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

Community detection

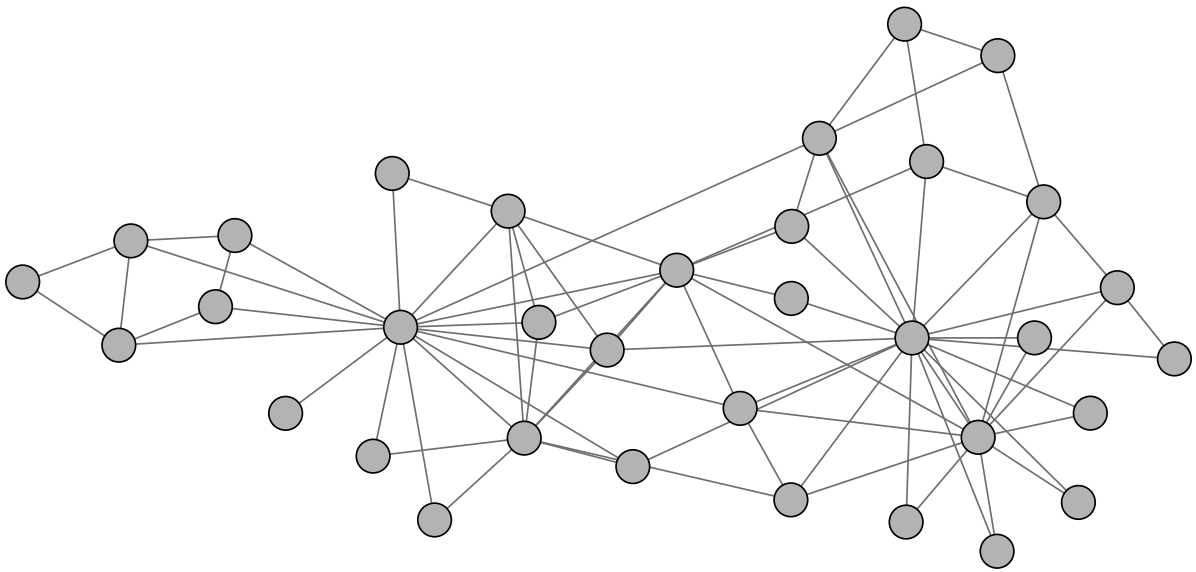


Figure: Communities: node subsets connected more strongly with each other

Community detection

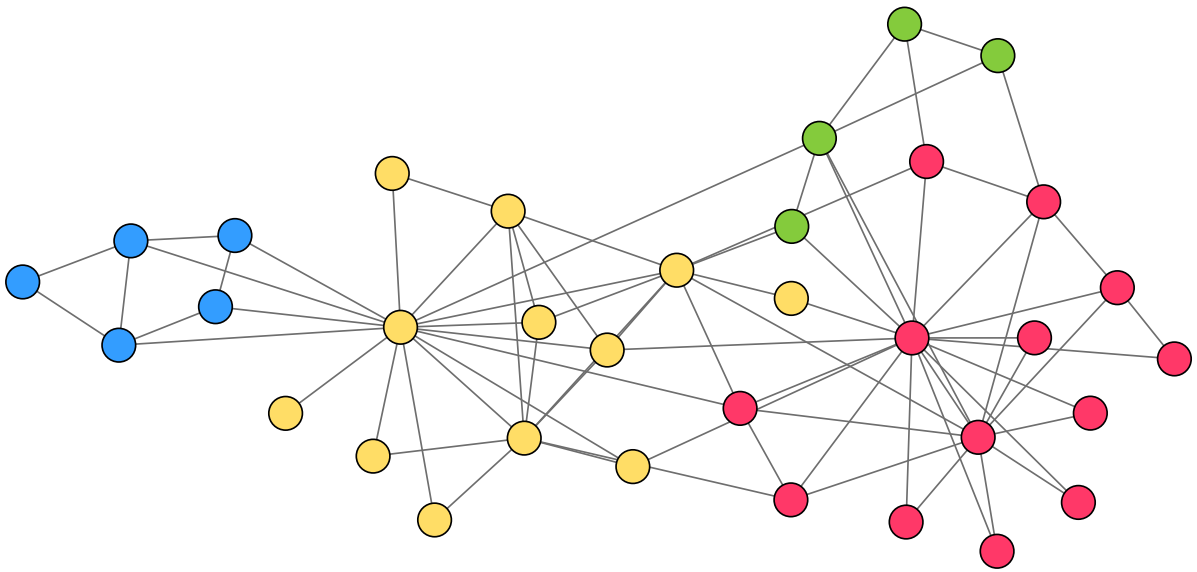


Figure: Communities: node subsets connected more strongly with each other

One approach: Modularity maximization

- **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)

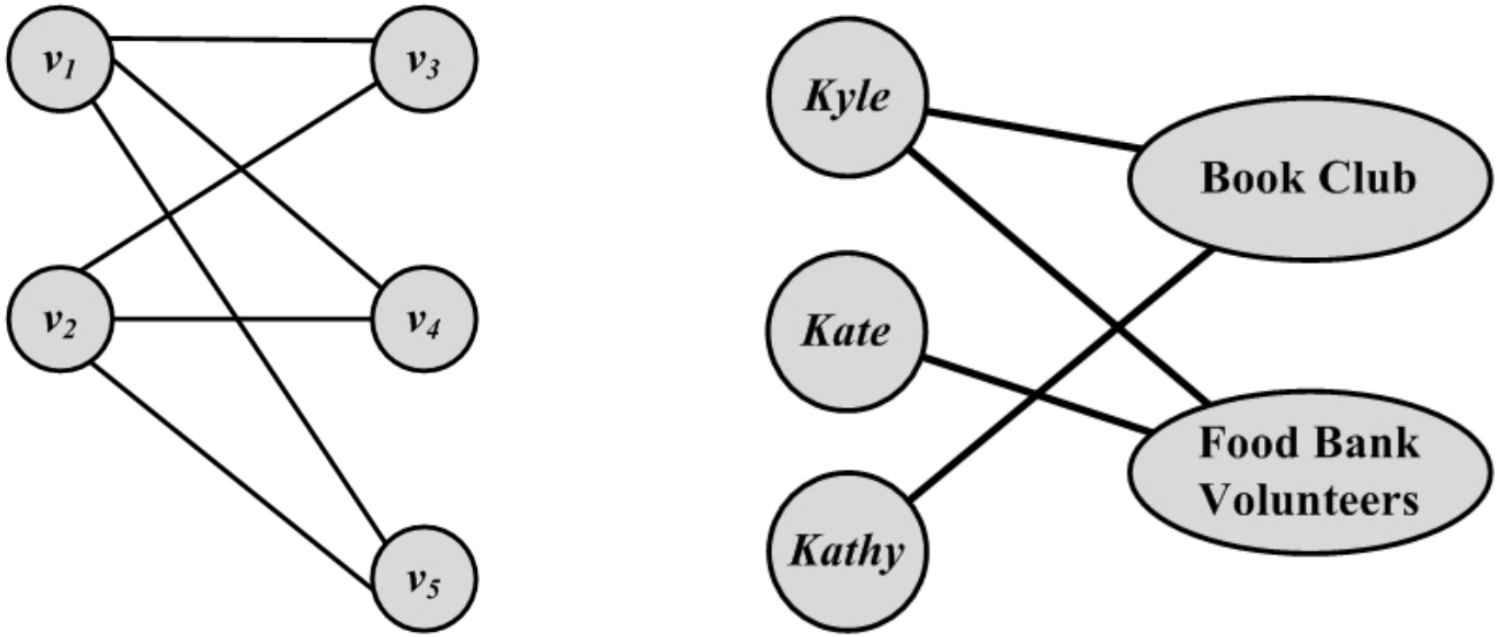
V.D. Blondel, J-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks in *Journal of Statistical Mechanics: Theory and Experiment* 10: P10008, 2008.

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

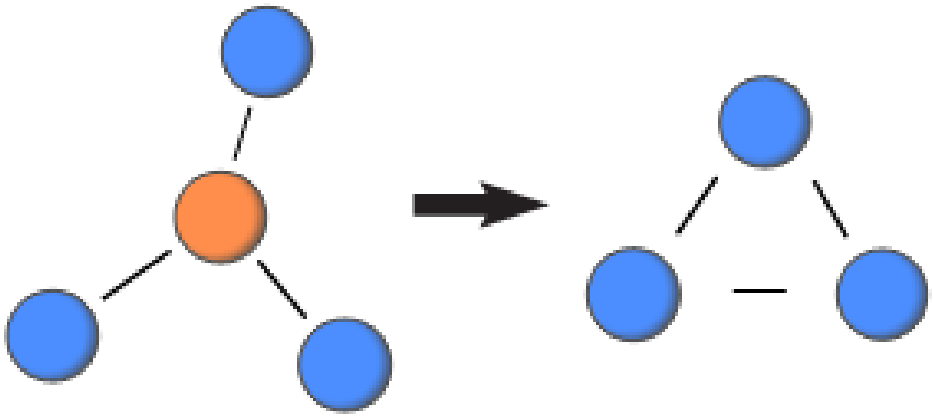
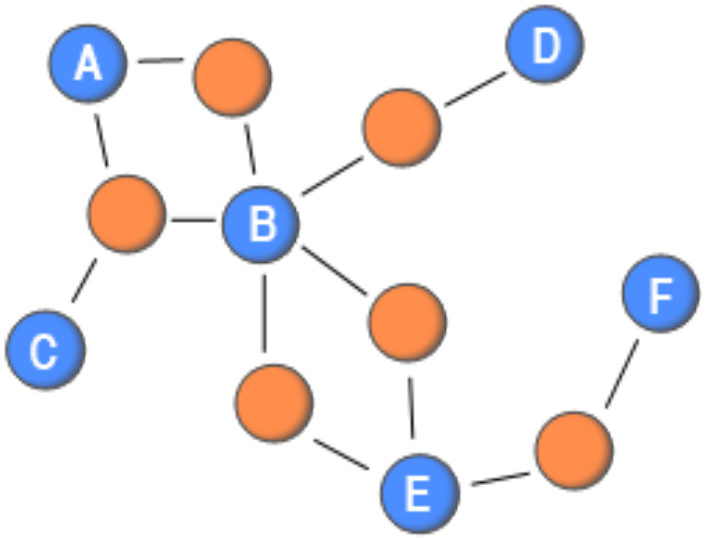


Image: `http://toreopsahl.com/`

Weighted projection



$\rightarrow$

Weighted projection

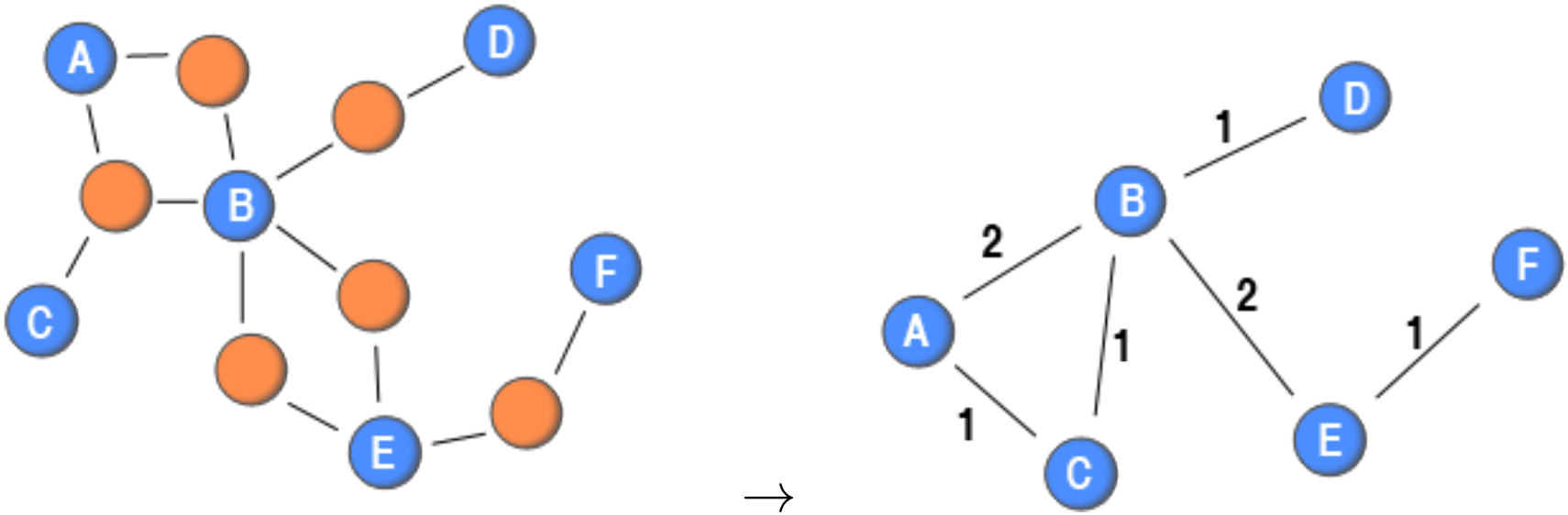
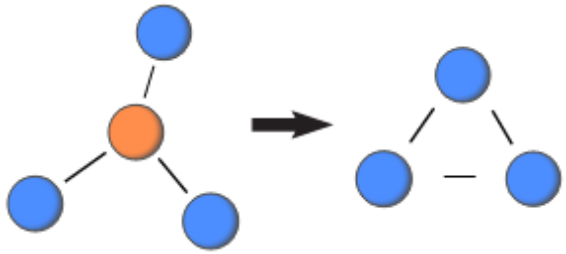


Image: `http://toreopsahl.com`

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

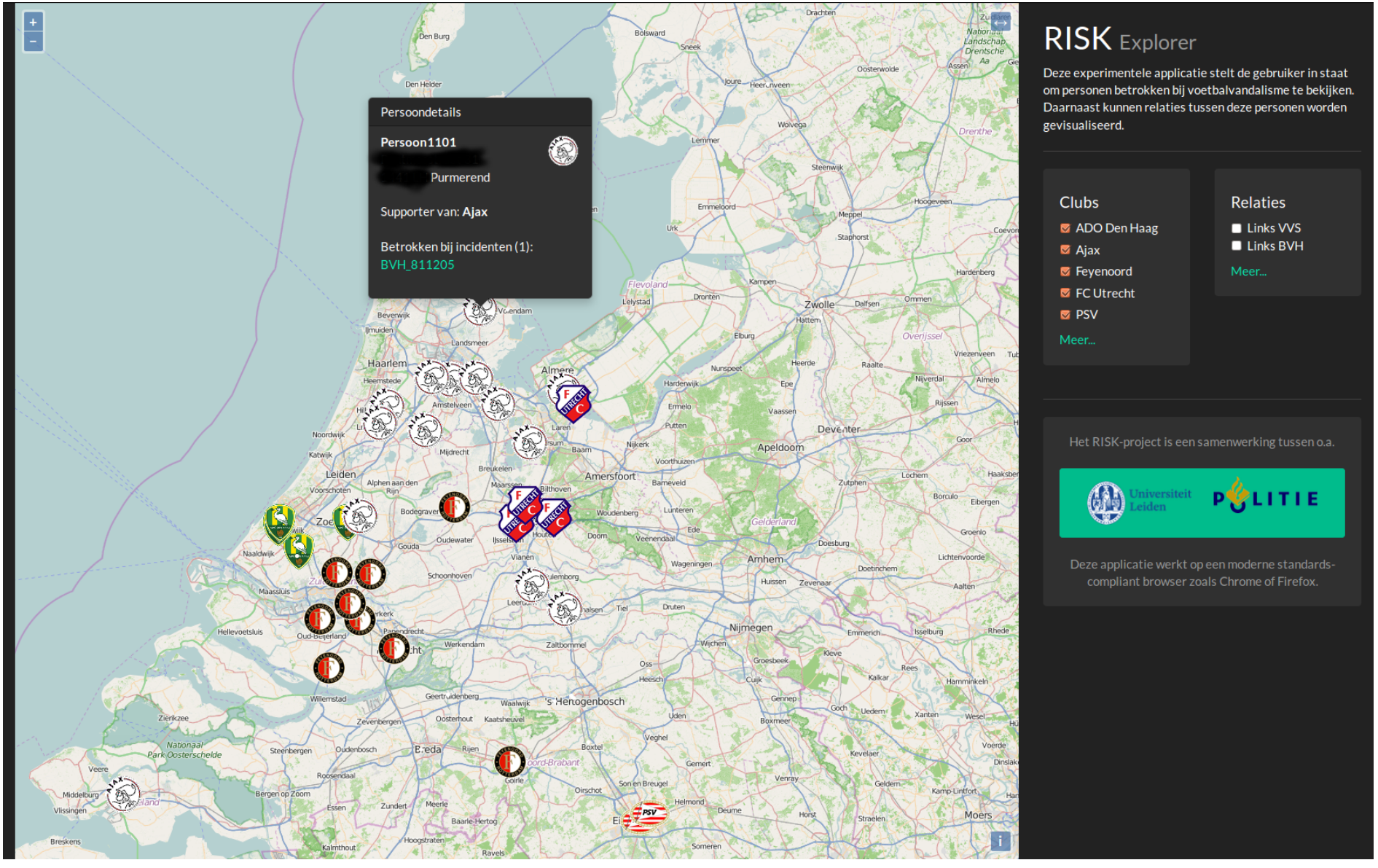
Network analysis on (almost) any dataset

- 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

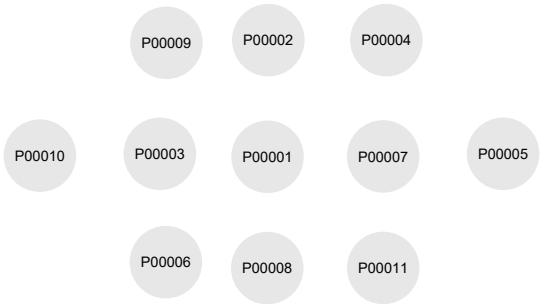
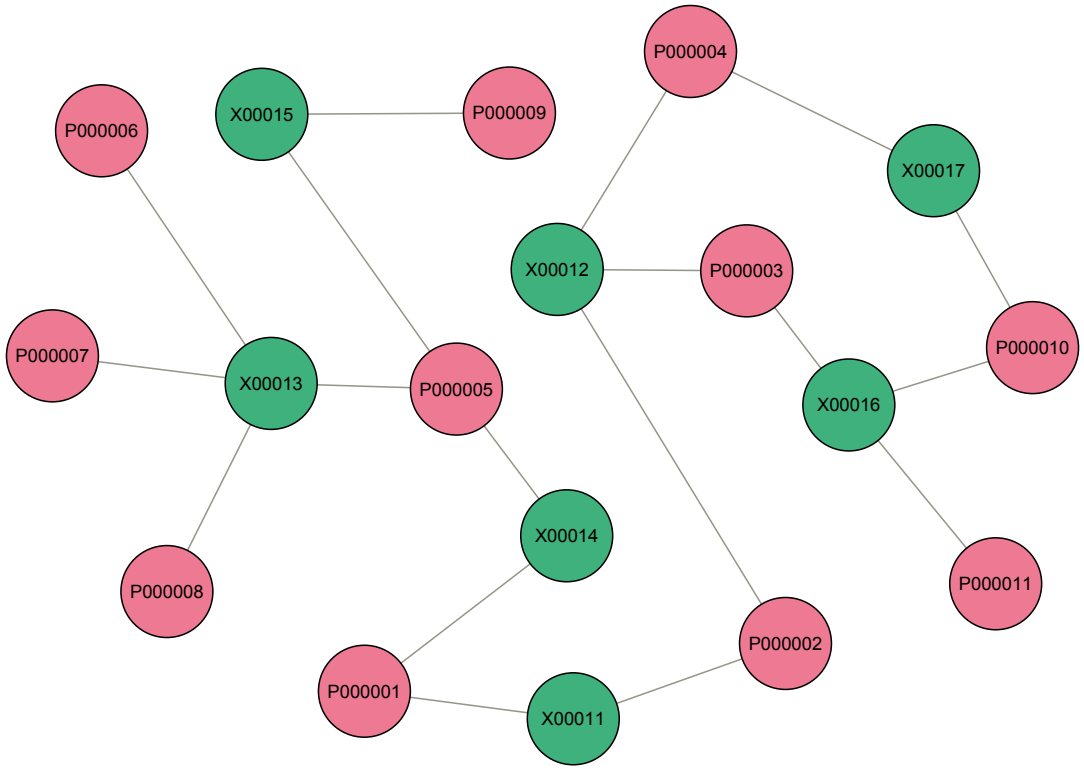


Figure: Suspects are nodes

Two-mode criminal network



Network formation

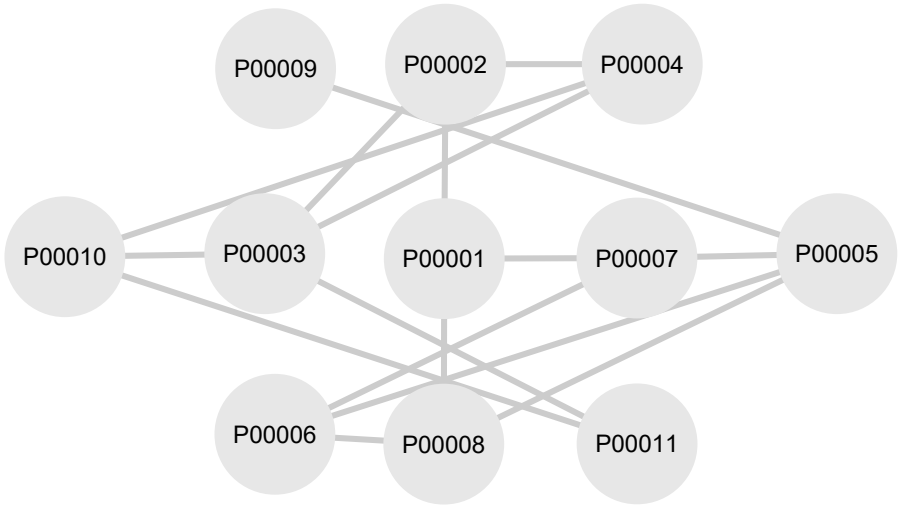


Figure: Edges are based on common involvement as a suspect

Network visualization

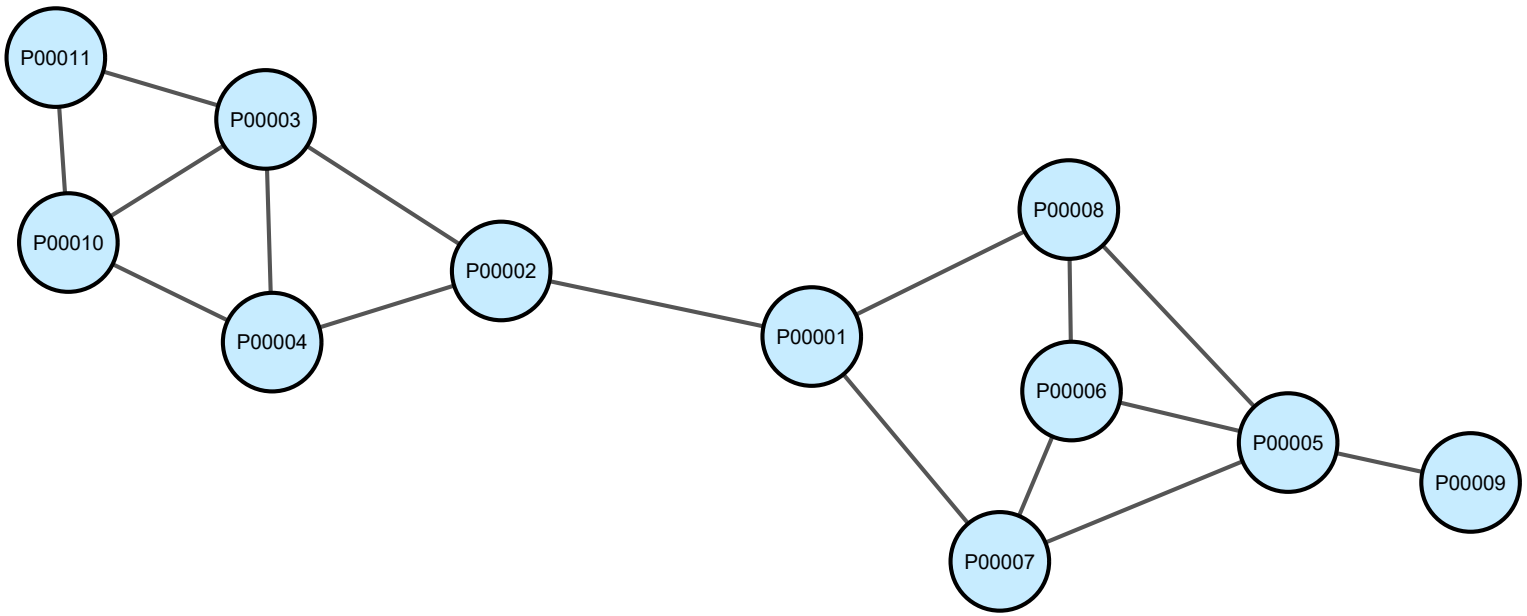


Figure: Force-directed visualization algorithm reveals structure

Network analysis: Centrality

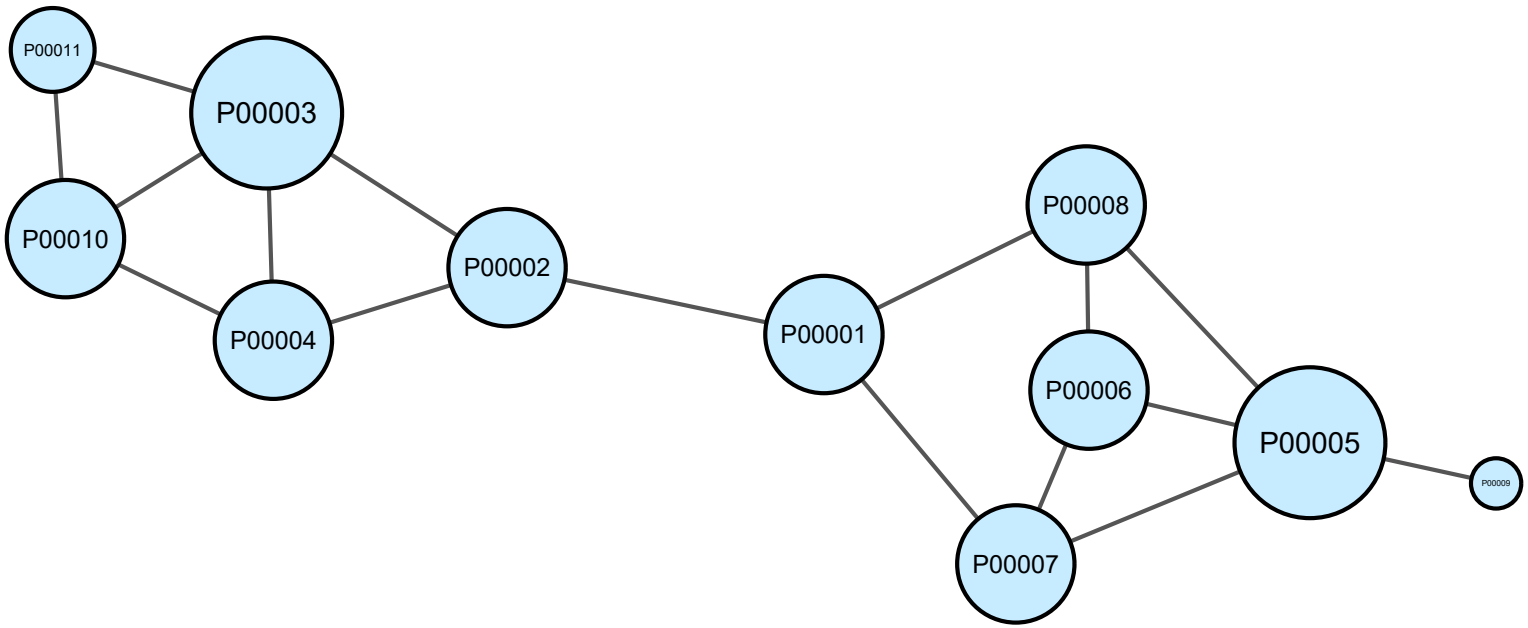


Figure: Degree centrality finds locally important nodes

Network analysis: Centrality

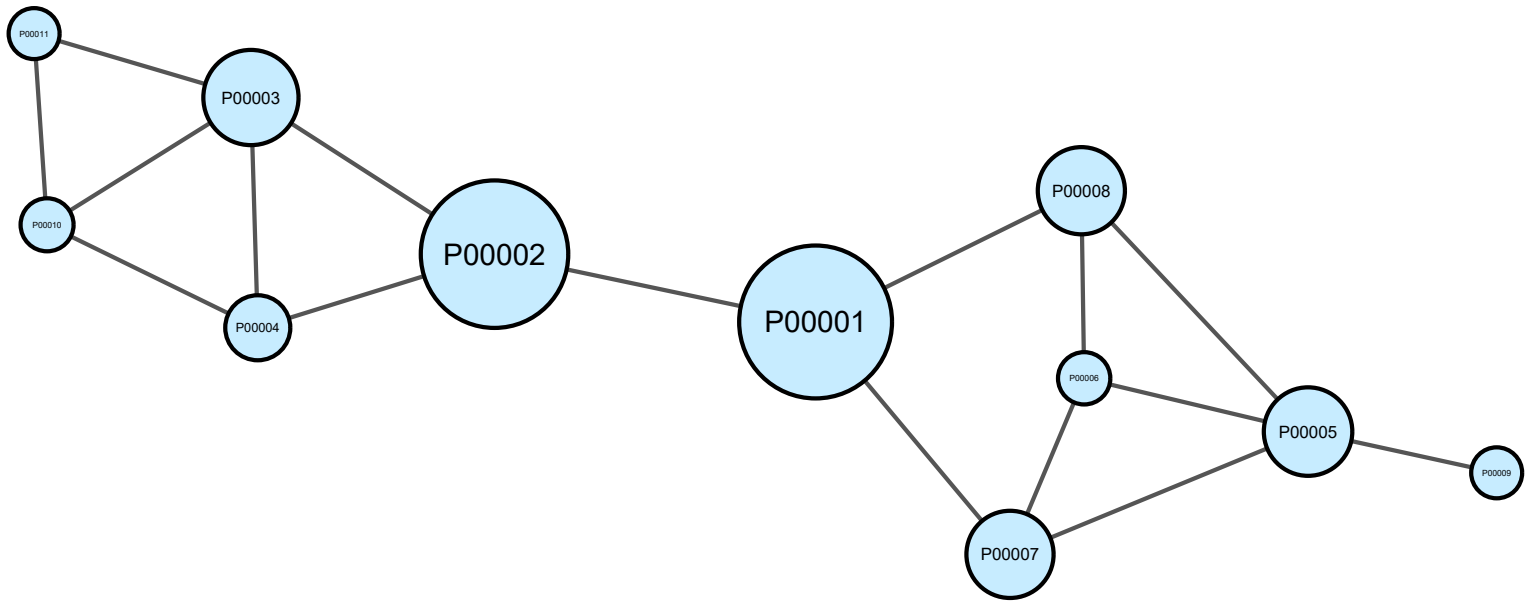


Figure: Betweenness centrality reveals globally important nodes

Network analysis: Community detection

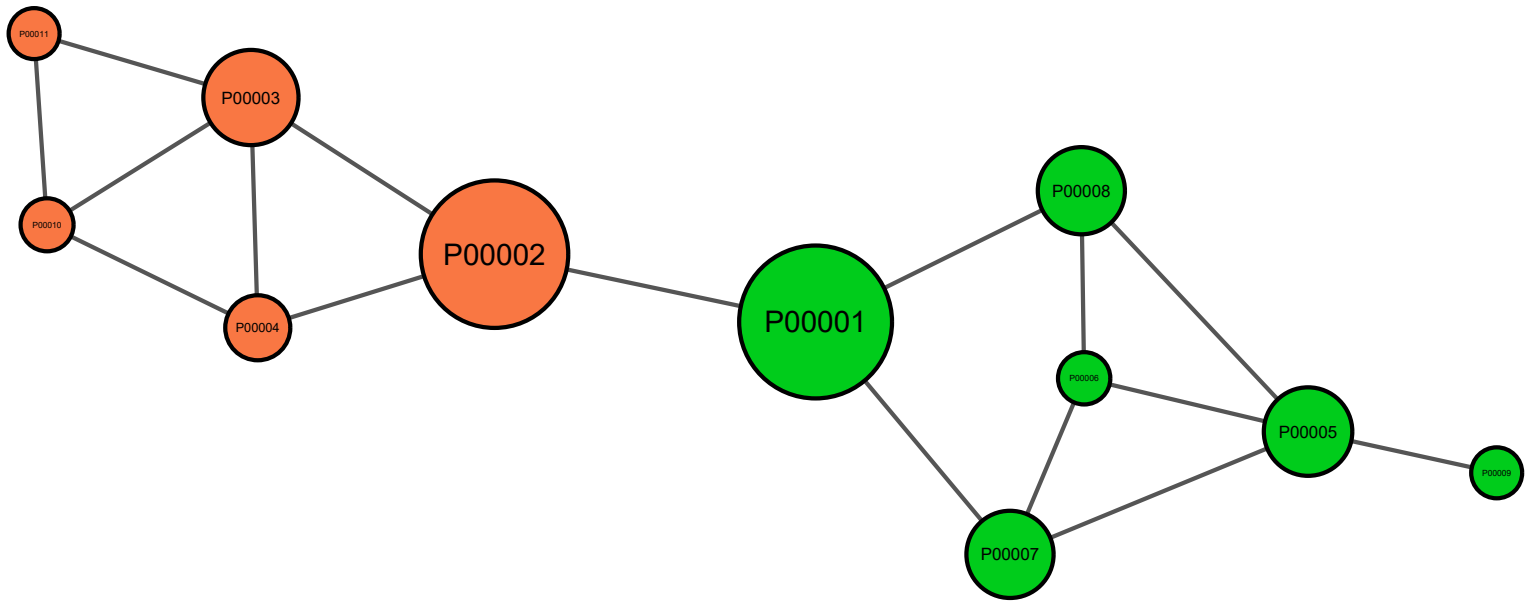
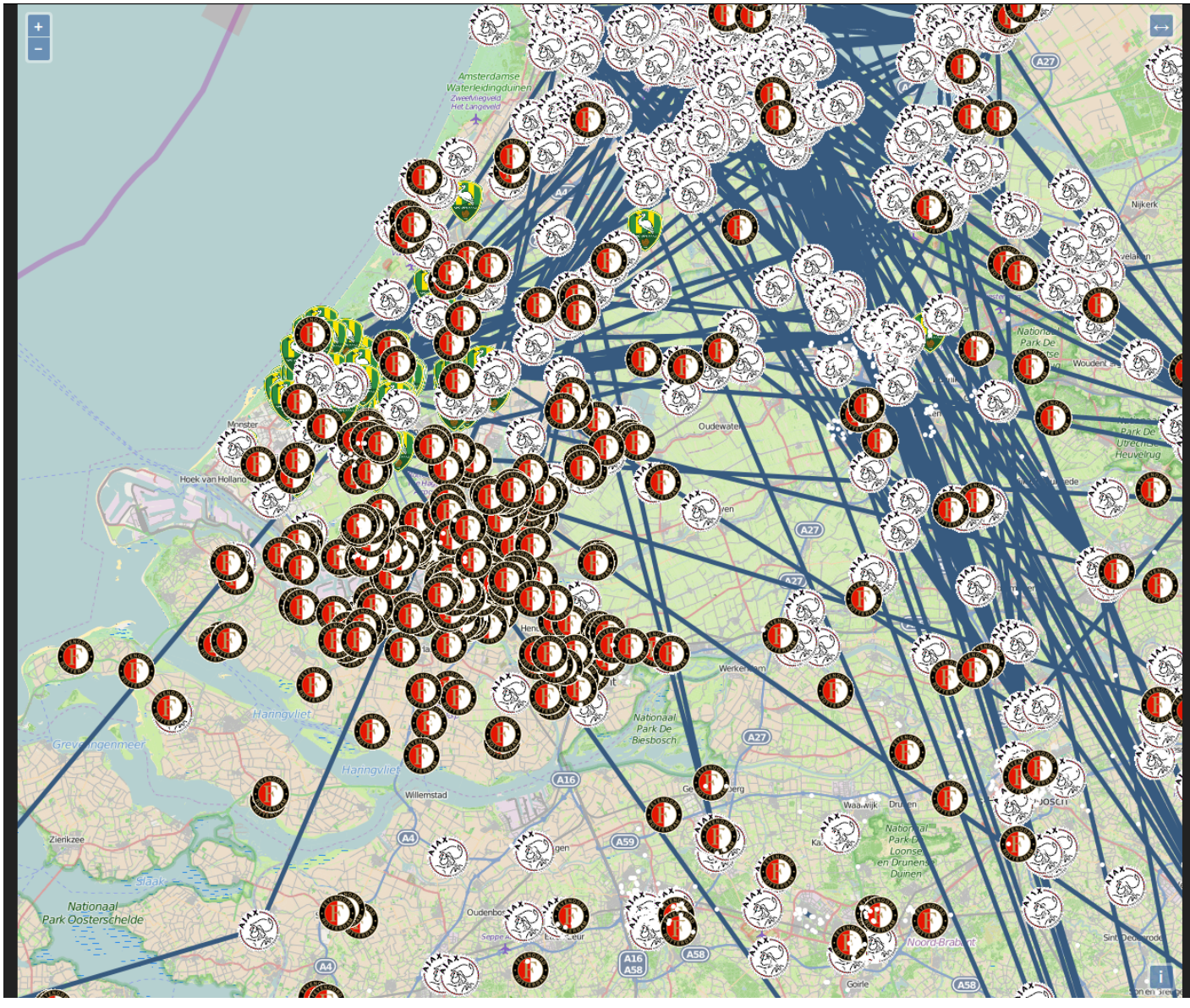


Figure: Community detection finds groups of tightly connected nodes



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

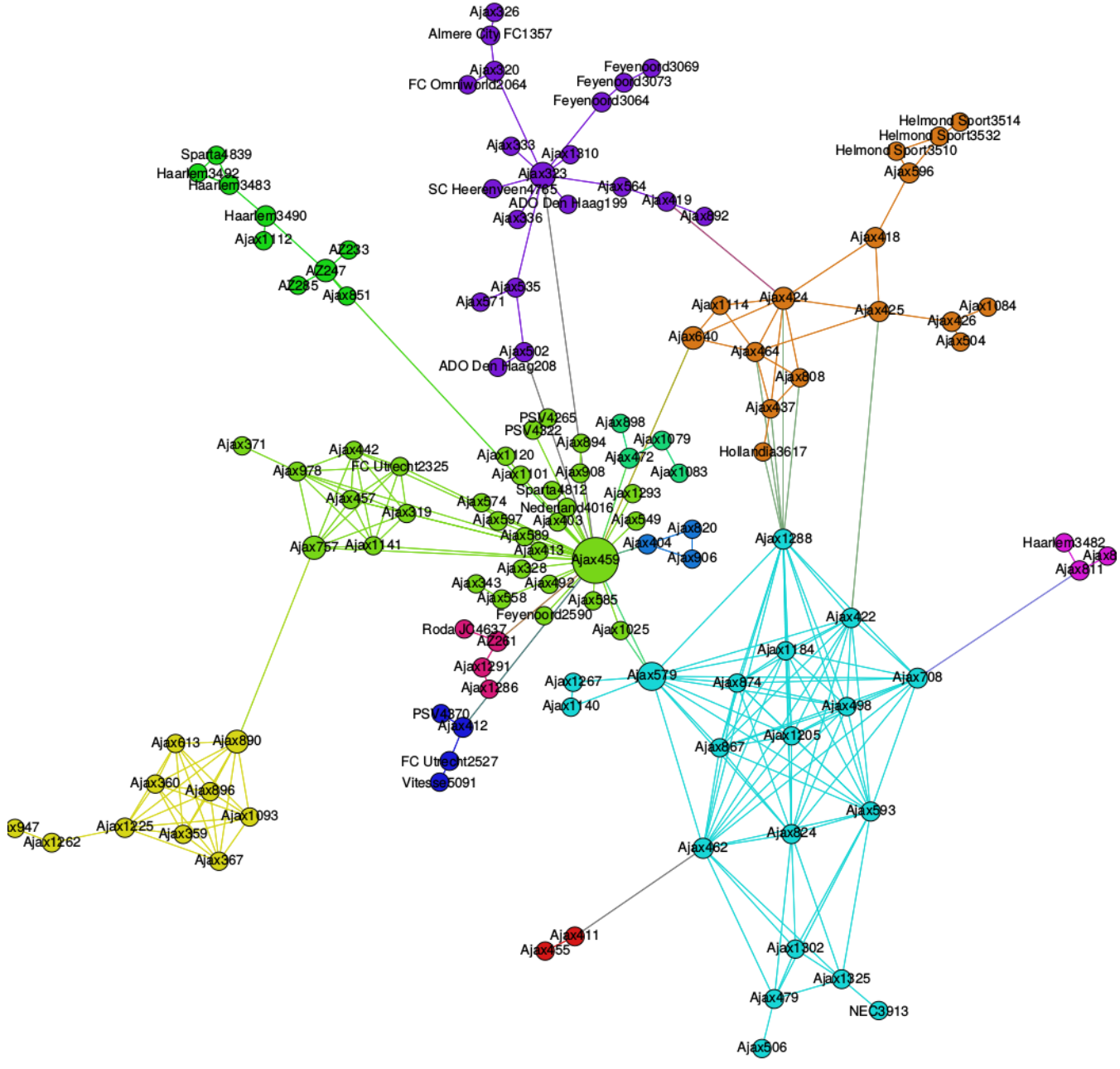
Meer...

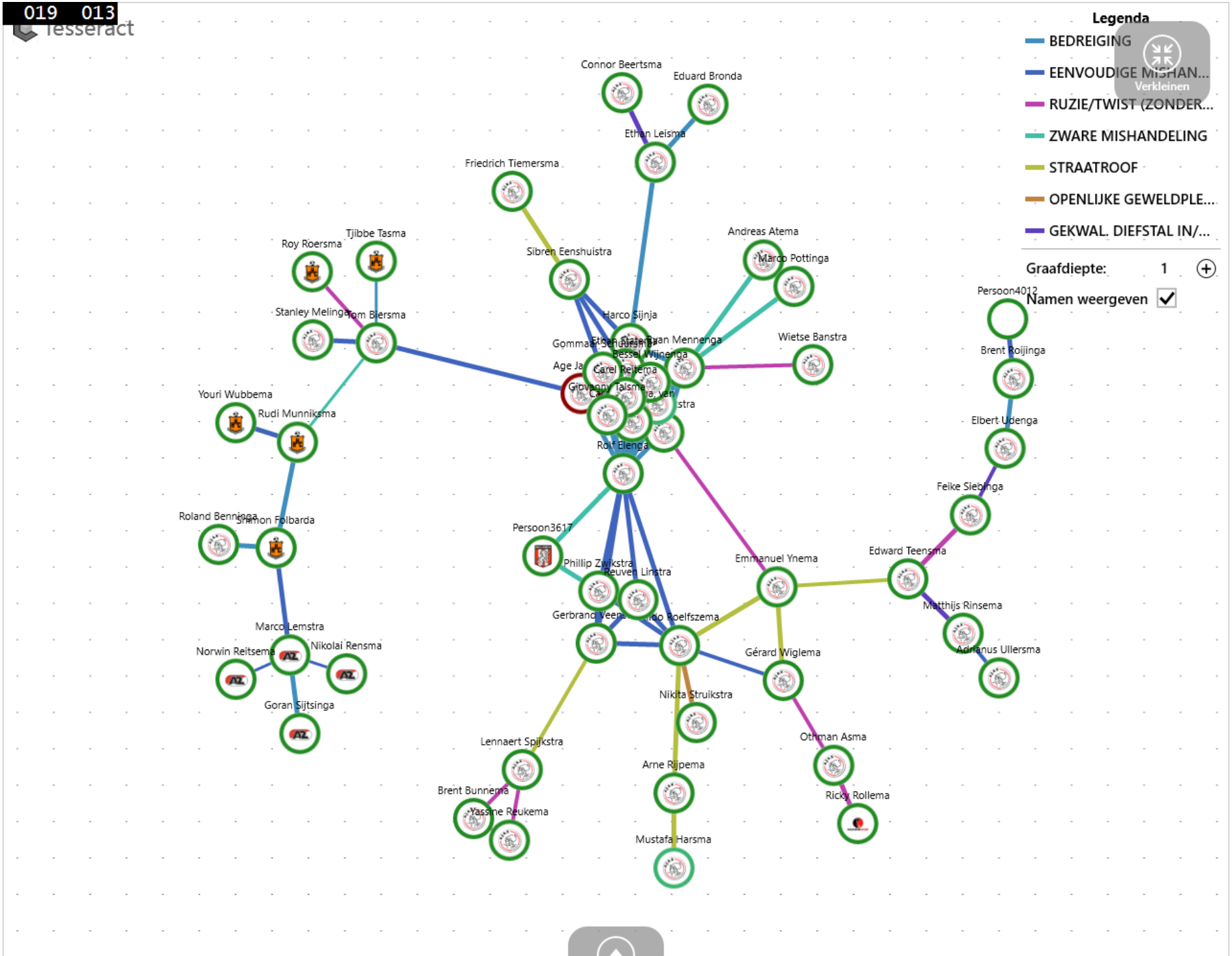
Het RISK-project is een samenwerking tussen o.a.

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

HITS

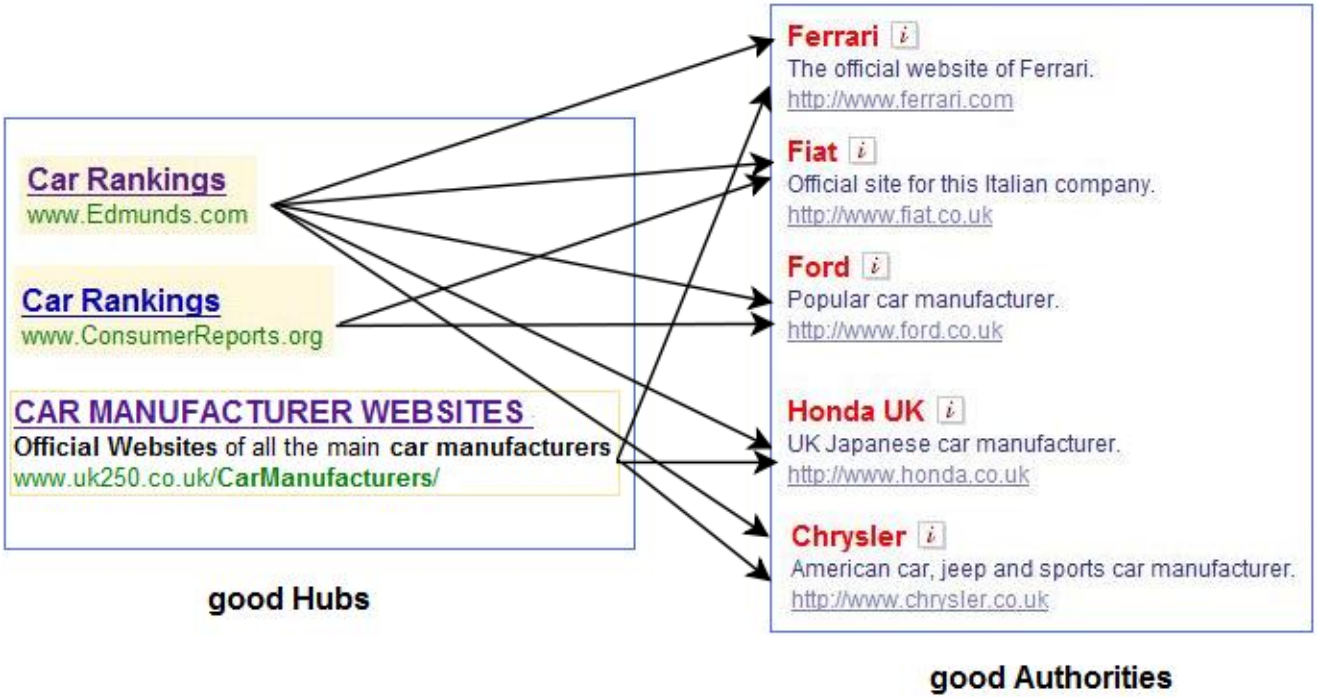
Centrality measures

- Distance/path-based measures:
  - Degree centrality $O(n)$
  - Closeness centrality $O(mn)$
  - Betweenness centrality $O(mn)$
  - Eccentricity centrality $O(mn)$
- **Propagation-based** measures:
  - Hyperlink Induced Topic Search (HITS)
  - PageRank

Hyperlink Induced Topic Search

- A link to a page is a "vote" for that page
- But how important is the page casting the vote?
- **Hyperlink Induced Topic Search (HITS)**
- **Hubs**: pages that link to good authorities
- **Authorities**: contain useful information and are therefore linked from many good hubs
- Jon Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46(5): 604–632, 1999.

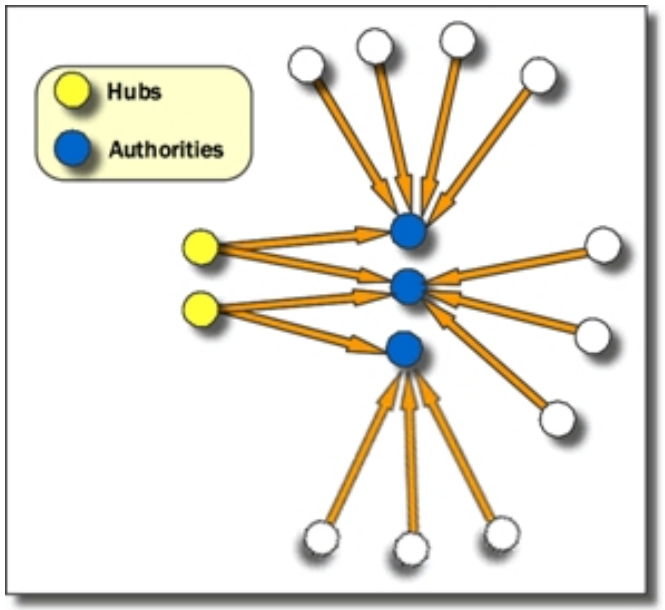
Hyperlink Induced Topic Search



Query: Top automobile makers

http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

Hyperlink Induced Topic Search



Leskovec, Stanford CS224W (`http://cs224w.stanford.edu`)

Hubs and authorities

- A "good webpage" is either a hub or an authority
- Each page $v \in V$ has two scores:
  - **Hub score** $h(v)$
  - **Authority score** $a(v)$
- Iterative algorithm
- Rules/definitions are somewhat "recursive"
- **Propagation model** that updates state at time $t+1$ based on t

HITS algorithm

- For all nodes $v \in V$, at $t = 0$ initialize $a^0(v) = h^0(v) = 1/\sqrt{n}$
- Repeat:
  1. $t = t + 1$
  2. Update the authority scores, so for all nodes $v \in V$: $a^{t+1}(v) = \sum_{v \in N'(v)} h^t(v)$
  3. Update the hub scores, so for all nodes $v \in V$: $h^{t+1}(v) = \sum_{v \in N(v)} a^t(v)$
  4. Normalize both scores so that $\sum_{v \in V} (a^{t+1}(v))^2 = \sum_{v \in V} (h^{t+1}(v))^2 = 1$
- Until scores converge:
  $\sum_{v \in V} (a^{t+1}(v) - a^t(v))^2 < \epsilon$ and
  $\sum_{v \in V} (h^{t+1}(v) - h^t(v))^2 < \epsilon$
  For some small ϵ.

HITS algorithm (easy mode)

- For all nodes v, **initialize** the hub and authority scores equally
- Repeat:
  1. $t = t + 1$
  2. **Update the authority score** of all nodes v to the sum of the hub scores of the nodes pointing to v
  3. **Update the hub score** of all nodes v to the sum of the authority scores of the nodes to which v points
  4. **Normalize** both scores so that they sum to 1
- Until values **converge**: between iteration t and $t + 1$ the values of both scores differ less than ϵ

HITS complexity

- Space: 2 lists of size n for hub and authority scores, so $O(n)$
- Time: Update and normalize n values in each iteration based on their neighborhoods of average size (m/n), so $O(n \cdot (m/n)) = O(m)$
- Usually 100 iterations for convergence, so $100 \cdot m$
- Compare this to betweenness or closeness centrality which takes $O(mn)$ time ...

Winner takes it all

PageRank

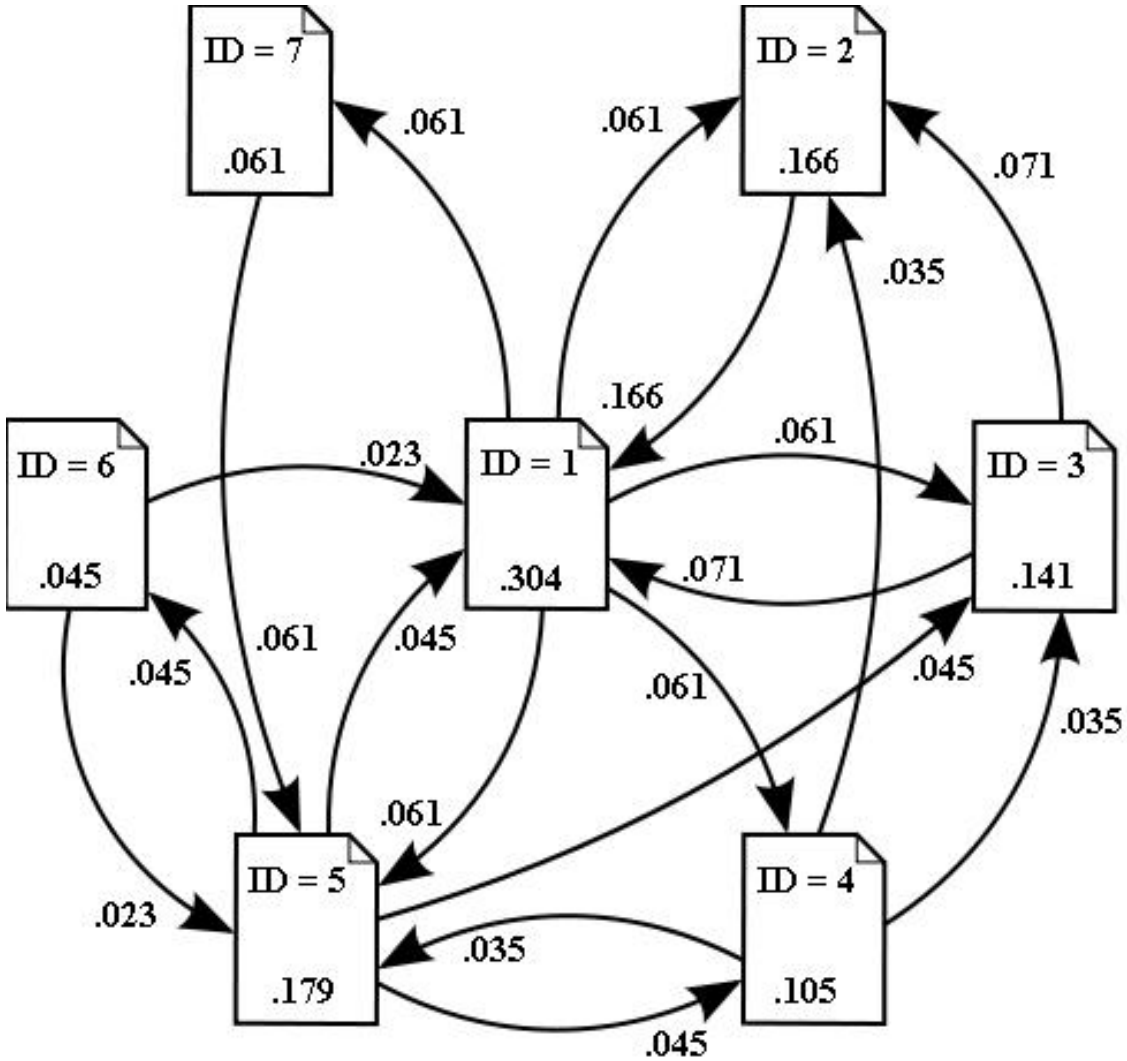
PageRank

- A link to a page is a "vote" for that page
- But how important is the page casting the vote?
- PageRank answer: that just depends on how many other pages vote for that page
- **PageRank**: number from 0 (low) to 10 (high) that indicates the importance of a page
- Similar to eigenvector centrality
- 1998: Page and Brin founded Google Inc.
- Larry Page and Sergey Brin, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford Infolabs, 1999.

Towards PageRank

- Assume that $outdeg(v)$ is the outdegree of node $v \in V$
- Each page has its own importance $PR(v)$
- Each page v casts equal votes of size $\frac{PR(v)}{outdeg(v)}$ for all other pages $w \in N(v)$ that it links to
  (in practice, `rel="nofollow"` prevents this)
- The amount of importance $PR(v)$ that a page receives depends on the pages that link to it: $PR(v) = \sum_{w \in N'(v)} \frac{PR(w)}{outdeg(w)}$
- Again recursive
- Does it converge?

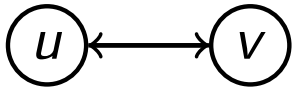
Towards PageRank example



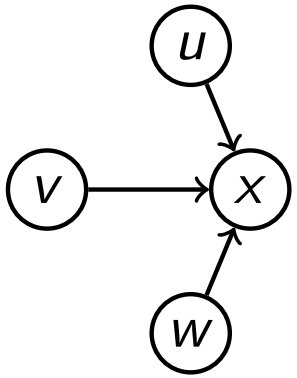
Challenges

$$PR(v) = \sum_{w \in N'(v)} \frac{PR(w)}{outdeg(w)}$$

- **Spider traps**: links back and forth:



- **Dead ends**: pages that do not have outgoing links



Towards PageRank

- **Random Surfer** model
- Idea: a user browsing the web either
  - clicks a link on the current page, or
  - opens an arbitrary other page
- With probability p, follow a link to a neighbor
- With probability $1 - p$, jump to a random node
- In practice: $p = 0.85$ and thus $1 - p = 0.15$ ("follow five links and jump")

PageRank algorithm

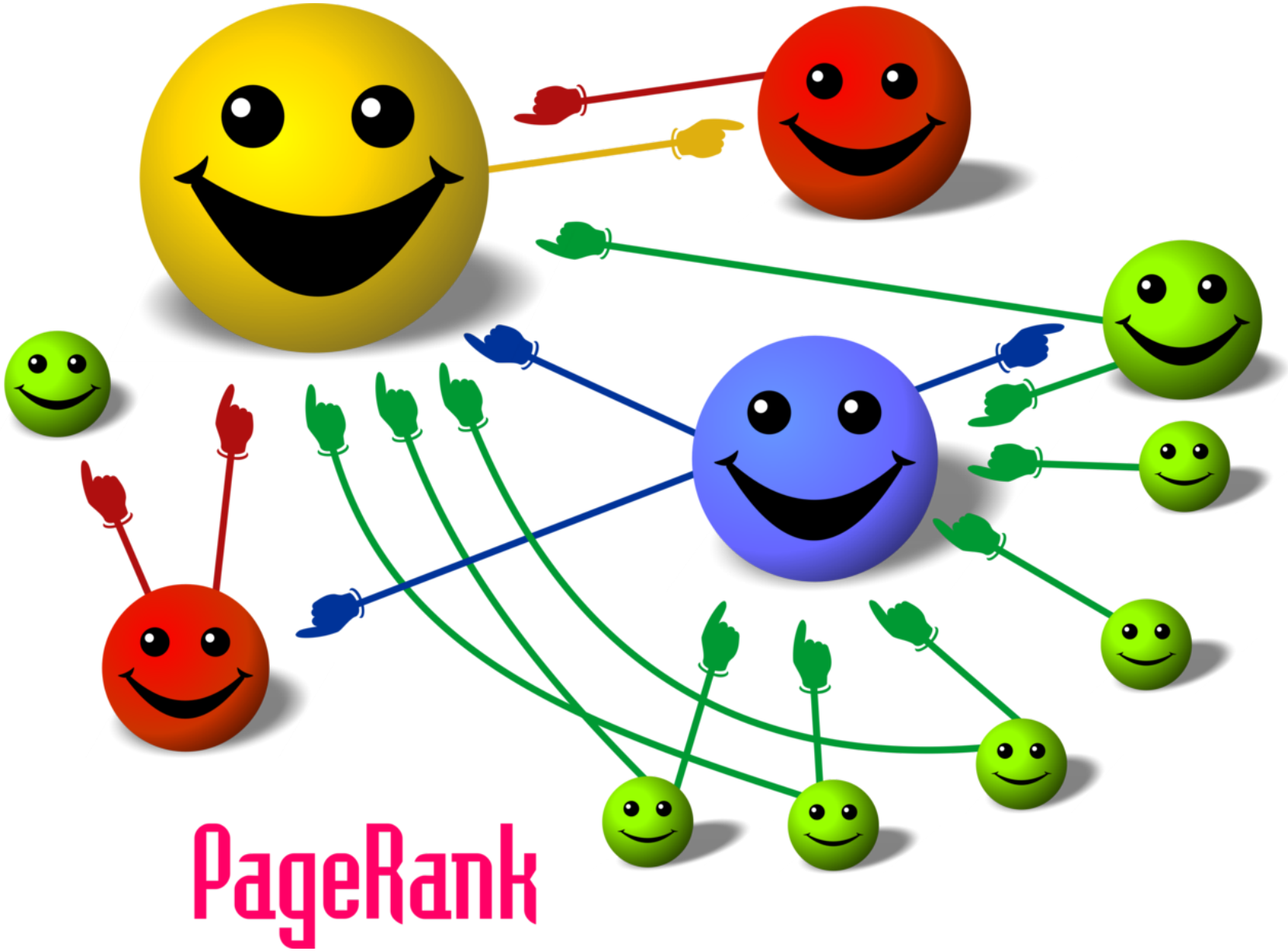
- For all nodes $v \in V$, initialize $PR^0(v) = (1/n)$
- $t = 0$
- Repeat:
  1. $t = t + 1$
  2. $PR^t(v) = \frac{1-p}{n} + p \cdot \sum_{w \in N'(v)} \frac{PR^{t-1}(w)}{outdeg(w)}$
  3. Normalize so that $\sum_{v \in V} PR(v) = 1$
     (just divide each value by the sum of all values)
- Until scores converge:
  $\sum_{v \in V} |PR^t(v) - PR^{t-1}(v)| < \epsilon$
  (for some small value of ϵ)

PageRank centrality

- PageRank $C_{PR}(v)$, which is the value of $PR(v)$ after iteratively and simultaneously applying:

$$PR(v) = \frac{1-p}{n} + p\left(\sum_{w \in N'(v)} \frac{PR(w)}{outdeg(w)}\right)$$

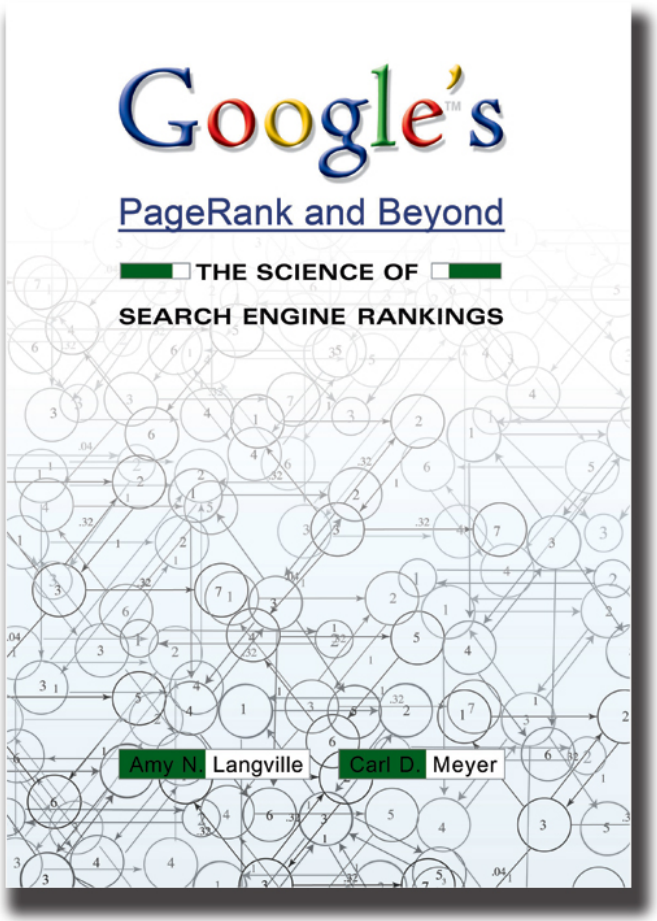
  for each of the nodes $v \in V$ and then normalizing the values so that they sum to 1, where $PR(v)$ is initialized to $1/n$ and $N'(v)$ is the set of nodes that links to node v and $p = 0.85$
- 100 iterations is usually enough
- Time $100 \cdot m$, similar to HITS.


PageRank

And more ...

- Jump to relevant pages with higher probability
- Choose a relevant neighbor with higher probability
- **Relevance** based on keywords, previous visits, geo aspects, ...
- Computation and definition using matrices
- Many other PageRank variants ...
- Personalized PageRank

PageRank and beyond



Actual Google (Page)Rank

- PageRank $PR(v)$
- Relevant keywords
- User's search history
- Local aspects
- "Rewards and punishments"

PageRank hunters

Elaine Washburn aan info

details weergeven -10 Allen beantwoorden

Hi Frank,

Please see proposal for [redacted] below:

We represent several industries that might interest you:

- Online gaming: you would receive 150 USD per year
- Finance, telecommunications, tourism or health: you would receive 100 USD per year

The advert will be text, not a visual banner. It will appear on a single page of your website. We aim to complete payment via secure payment partners Paypal or Moneybookers within 48 hours of the advert going live on your site.

Also, please read our terms and conditions: www.moredigital.com/terms.pdf.

Please let me know which industry you prefer, we'll then let you know which client fits your site best and draft an advert!

Best regards,
Elaine

SEO


WHAT THE HELL IS SEO?
CLICK CLICK, GOOGLE SEARCH...
GOOGLE TOP SEARCH RESULT:
SEO - STORA ENSO OYJ IS A PAPER AND BOARD MANUFACTURER IN EUROPE..
AND I THOUGHT THAT WAS SOMETHING ABOUT BEING ON TOP OF GOOGLE...
SEOCOMIC
by JOZZUA - ISULONG-SEOPH.NET

Movies . . .

Centrality measures

- Distance/path-based measures:
  - Degree centrality $O(n)$
  - Closeness centrality $O(mn)$
  - Betweenness centrality $O(mn)$
  - Eccentricity centrality $O(mn)$
- **Propagation-based** measures:
  - Hyperlink Induced Topic Search (HITS) $O(m)$
  - PageRank $O(m)$

Course project

Course project

- Project on specific SNA subtopic, 60% of your course grade
- Teams of exactly 2 students
- Deliverables:
  - **Presentation** on a topic-related paper. $\leq$ 20 minutes for your talk, $\approx$10 minutes for questions and discussion
  - **Paper** presenting a **contribution** to SNA that goes beyond what is done in the paper you study
  - Short peer review document (during peer review session)
  - Relevant project code and supplementary material
  - Bonus for open-source or open science contributions
- Topics divided over teams based on first come, first serve

Course project goal

- Study a specific SNA subtopic
- Provide a (modest) **contribution** to SNA that goes beyond what is done in the paper you study, e.g.:
  - Comparing similar algorithms from different papers
  - Testing the algorithm(s) on larger datasets
  - Validating algorithms using different metrics
  - Addressing future work posed in the paper
  - Replicate the study using more extensive parameter testing
- Project typically requires:
  1. thinking/design
  2. implementation/programming
  3. experimentation
  4. writing

  (likely partially iteratively)

Presentation

- Present
  - **the assigned paper** on the topic of your project
  - **your contribution** (what you present here depends on progress)
- Convey the main message of the paper in an understandable way
- Show a nice demo, pictures, movies or visualization
- Have a clearly structured presentation
- Briefly discuss your project plans
- Demo presentation will follow
- Discussion with (and engagement of) other students is expected (from both presenters and attendees)

Course project

- Read your paper, understand the main problem
- Do a bit of research on related literature
- Determine which algorithms/techniques/parameters/datasets/ subproblems you are going to compare (i.e., your contribution)
- Program (or obtain code of) the different algorithms and techniques
- Obtain and describe applicable datasets for comparing the algorithms
- Perform and report on experiments to compare the algorithms
- Determine and discuss results
- Write a sensible conclusion

Course project paper

- Scientific paper
- LaTeX
- 6 to 10 pages, two columns
- Images, figures, graphs, diagrams, tables, references, ...
- Between 5 and 9 sections
- Peer review and code review
- Option for "intermediary paper check" before final hand-in

Common pitfalls/excuses

- Only starting to read your paper after Assignment 2 (you are late)
- Starting just before the first paper deadline (you are very late)
- Starting writing only just before first review (your paper will likely be too meager)
- Starting writing code only just before (your algorithms will be slow, you do not have enough time to run your experiments, you claim it is "because everyone is using the servers")
- Copying from the internet
- A presentation without pictures
- Literally reading out every sentence on your slides
- Too much text on your slides
- Not writing the paper in LaTeX

Course project schedule

- See website

Upcoming lab session and "homework"

- Work on Assignment 1
- Course project team formation; stick around if you need a team mate

- Finalize Assignment 1; last week is the last lab session to ask questions
- Next week: consult the list of project topics on course website, and think of what you may want to work on together with your team mate