Universiteit Leiden

Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

Lecture 4 — Network structure of the web

So . . .

- You should be almost done with Assignment 1.
  Some hints:
  - Remember to concisely state how you obtained your answer
  - Algorithms always have an input and an output that corresponds to what is asked in the question, as well as a name
  - For Question 2.5 and 2.6, state whether you considered a directed or undirected variant of the measure. You are allowed (and encouraged) to use a form of approximation or sampling
- Accepting presentation date choices in lecture break; remainder via e-mail next week. By Oct 8, everyone should know their presentation date.

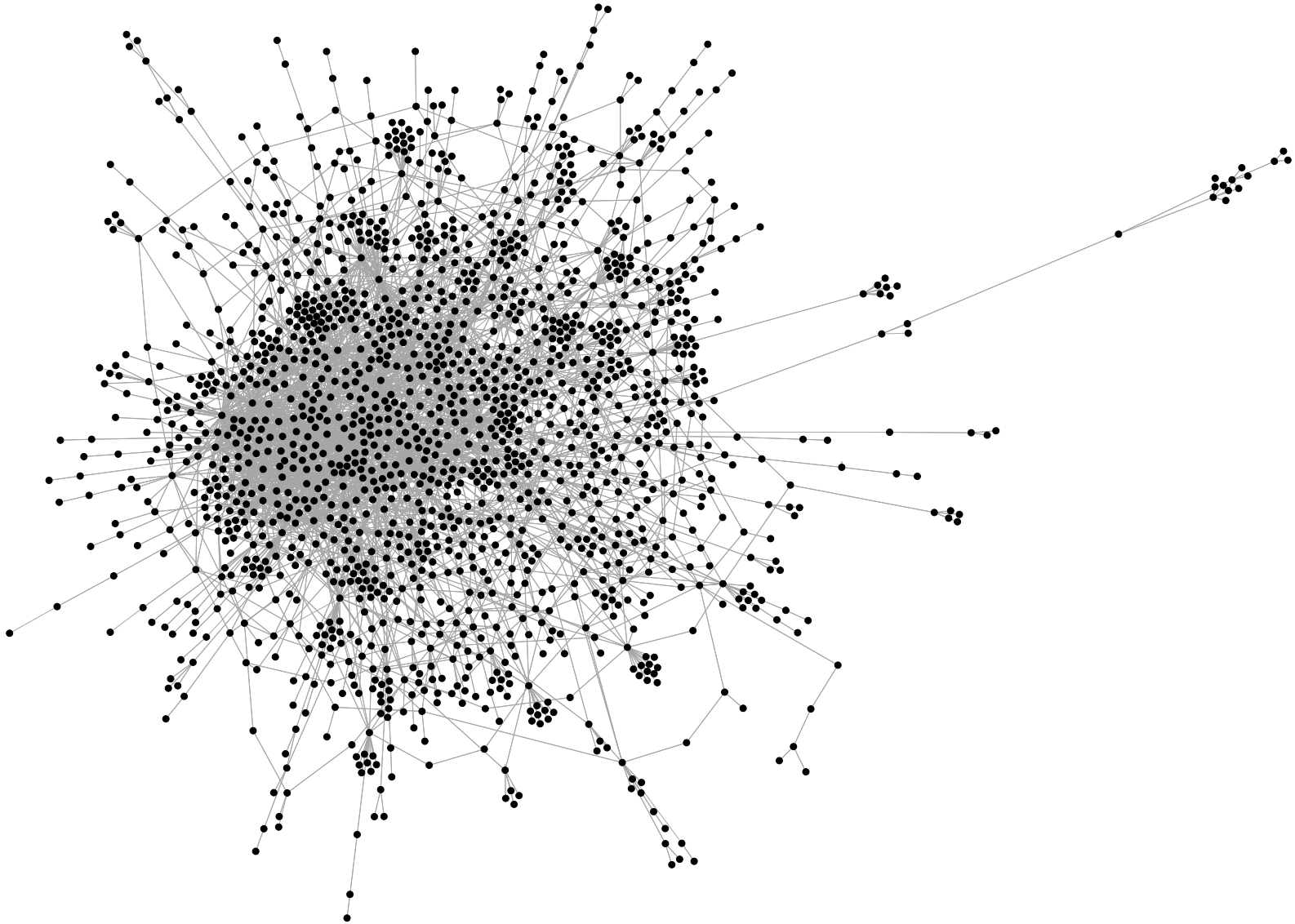
Today

- Recap
- Community detection — interpretation and visualization
- Propagation-based centrality
- Structure of the web
- Traversing Wikipedia
- Division of remaining students over teams and topics
- Planning of first weeks of presentations

Universiteit Leiden

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

Universiteit Leiden

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

Universiteit Leiden

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

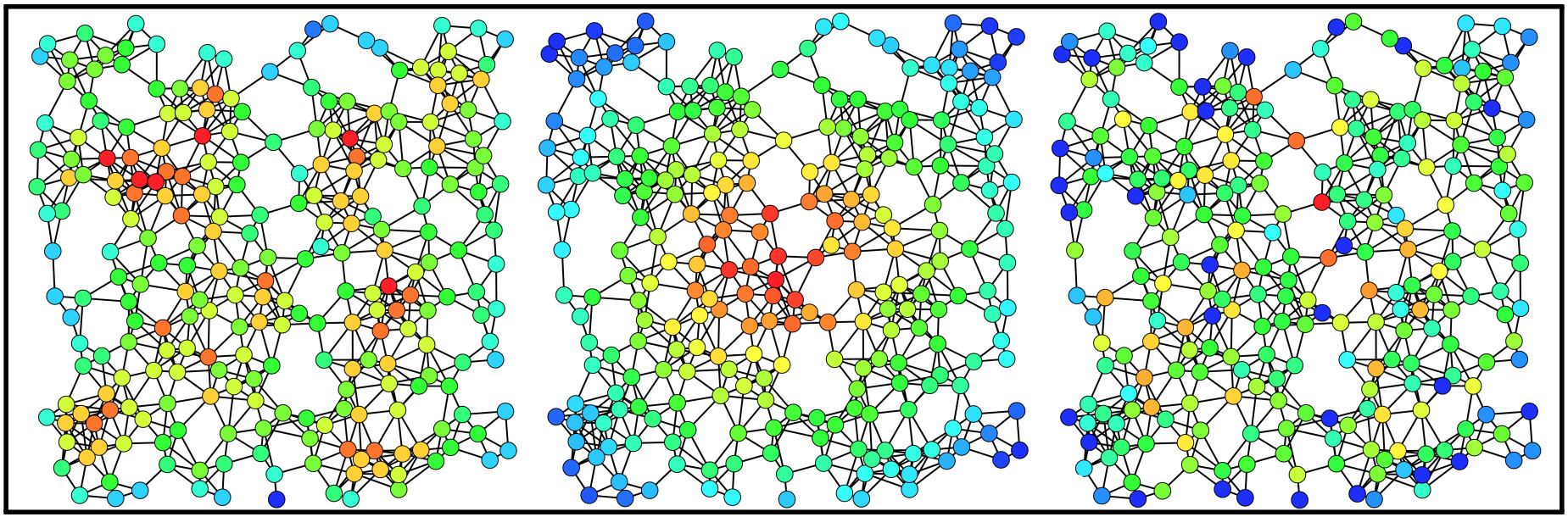


Figure: Degree, closeness and betweenness centrality

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

Community detection

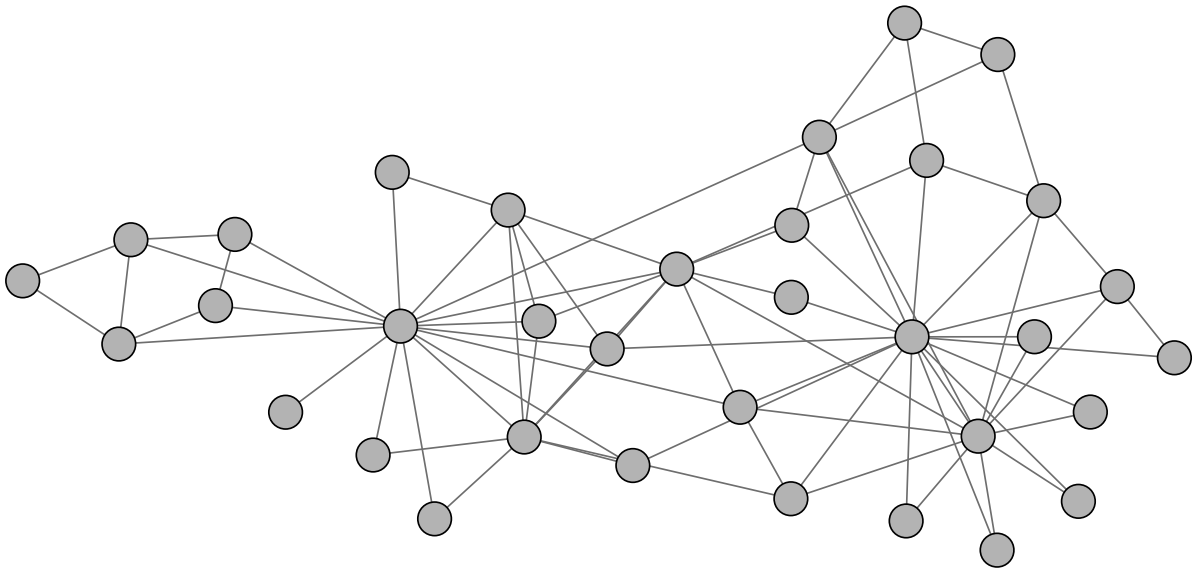


Figure: Communities: node subsets connected more strongly with each other

Community detection

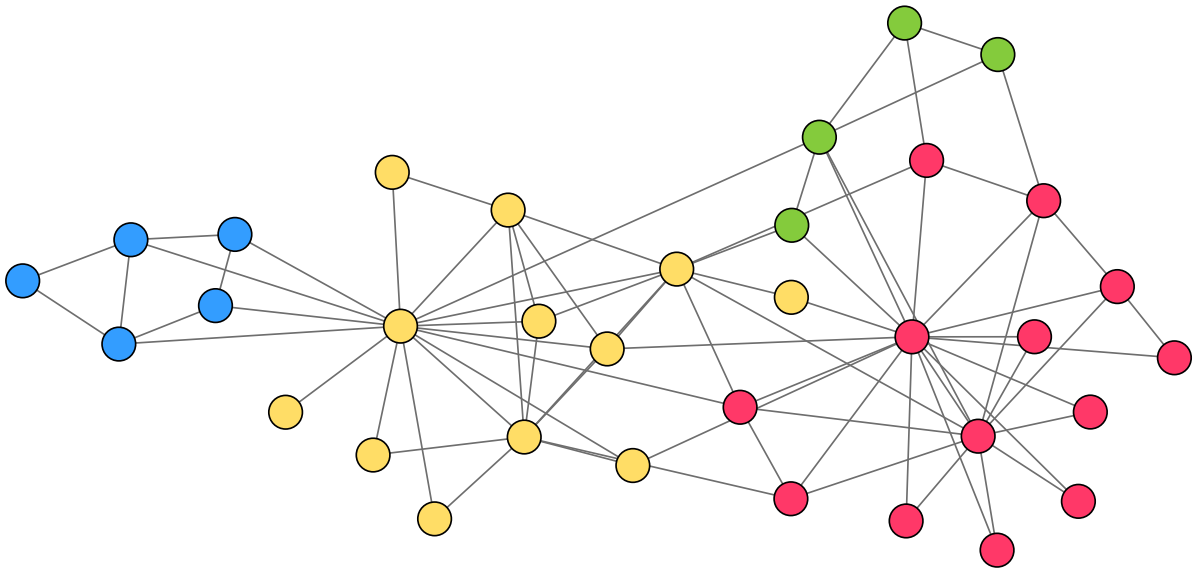


Figure: Communities: node subsets connected more strongly with each other

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.

Evaluating communities and partitions

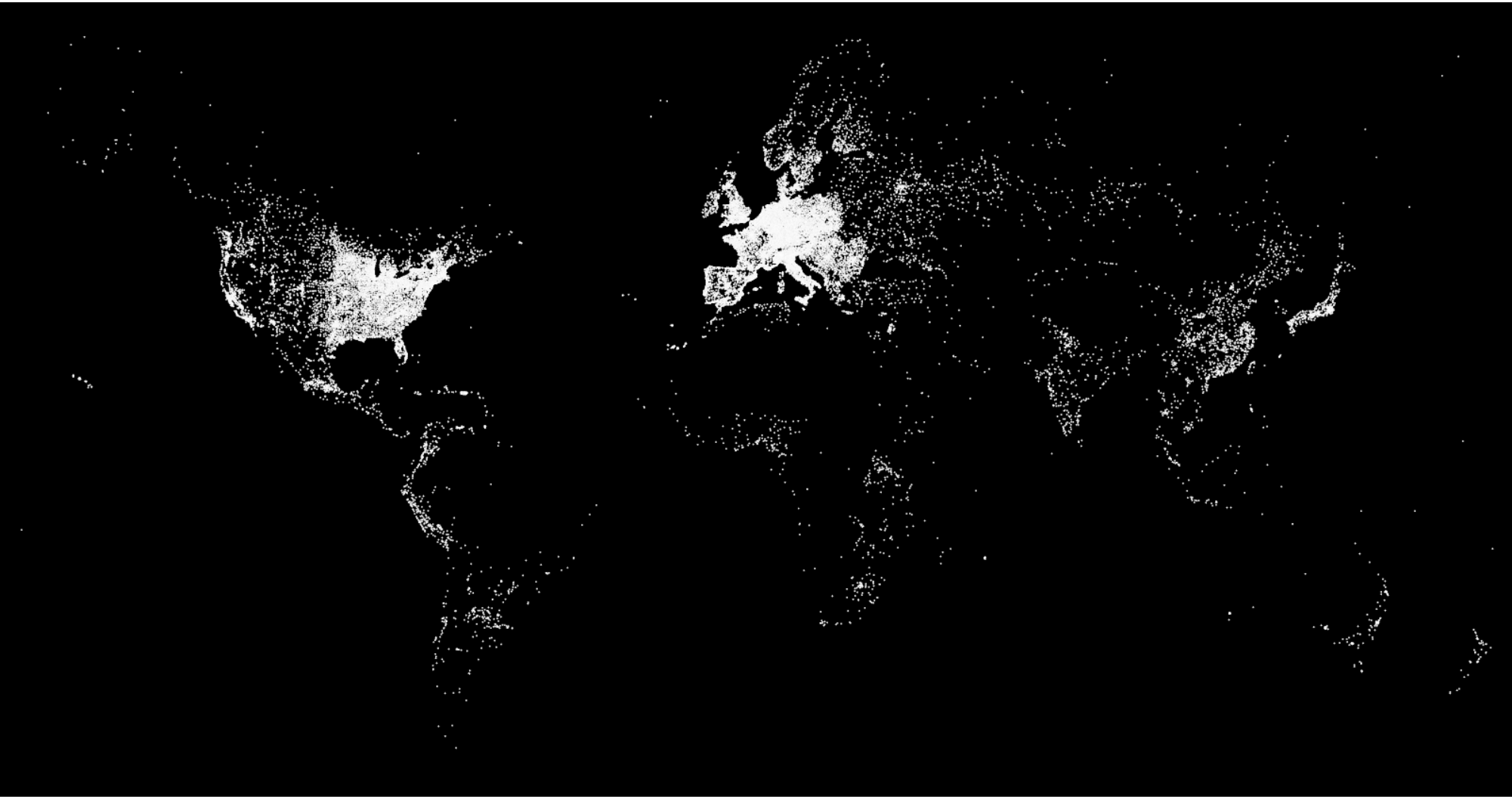
Universiteit Leiden

- **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 city networks

- Nodes are cities
- Edges between cities are based on firms sharing directors
- Weights on edges denote the number of connections
- Each city has an associated country
- Provides insight in geographical orientation of global economy

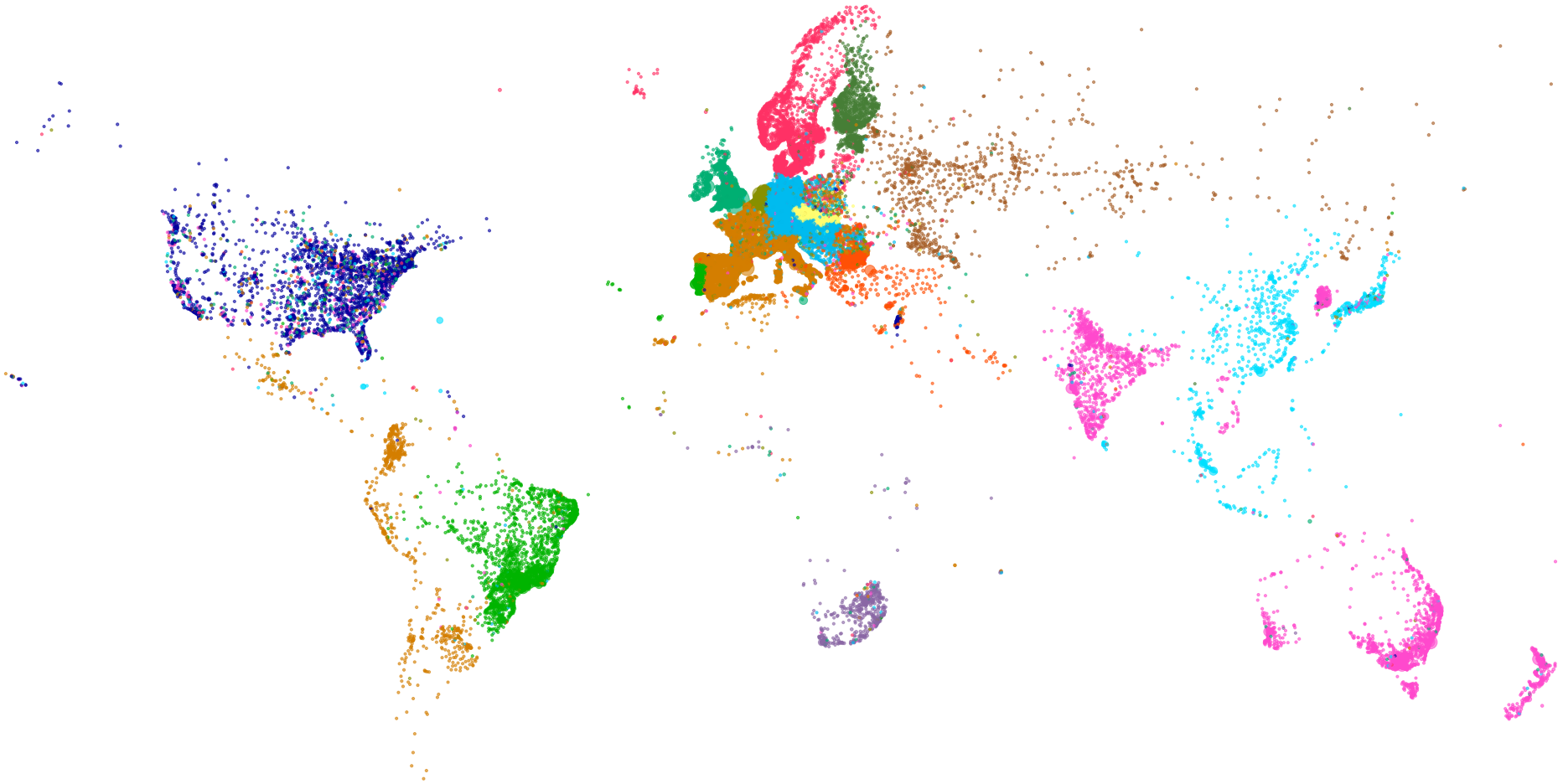
Corporations



Corporate network

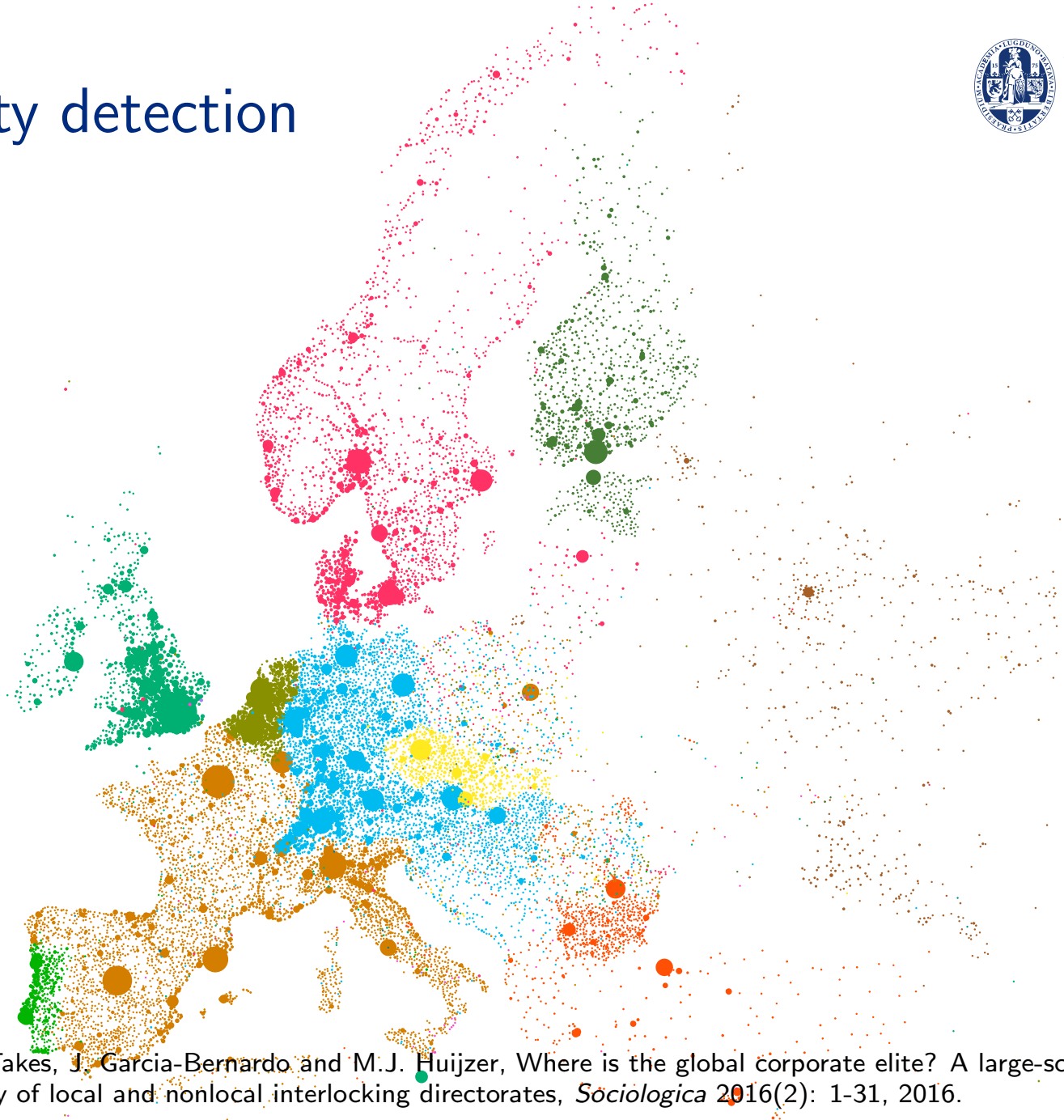
F.W. Takes and E.M. Heemskerk, Centrality in the Global Network of Corporate Control, *Social Network Analysis and Mining* 6(1): 1-18, 2016.

Community detection



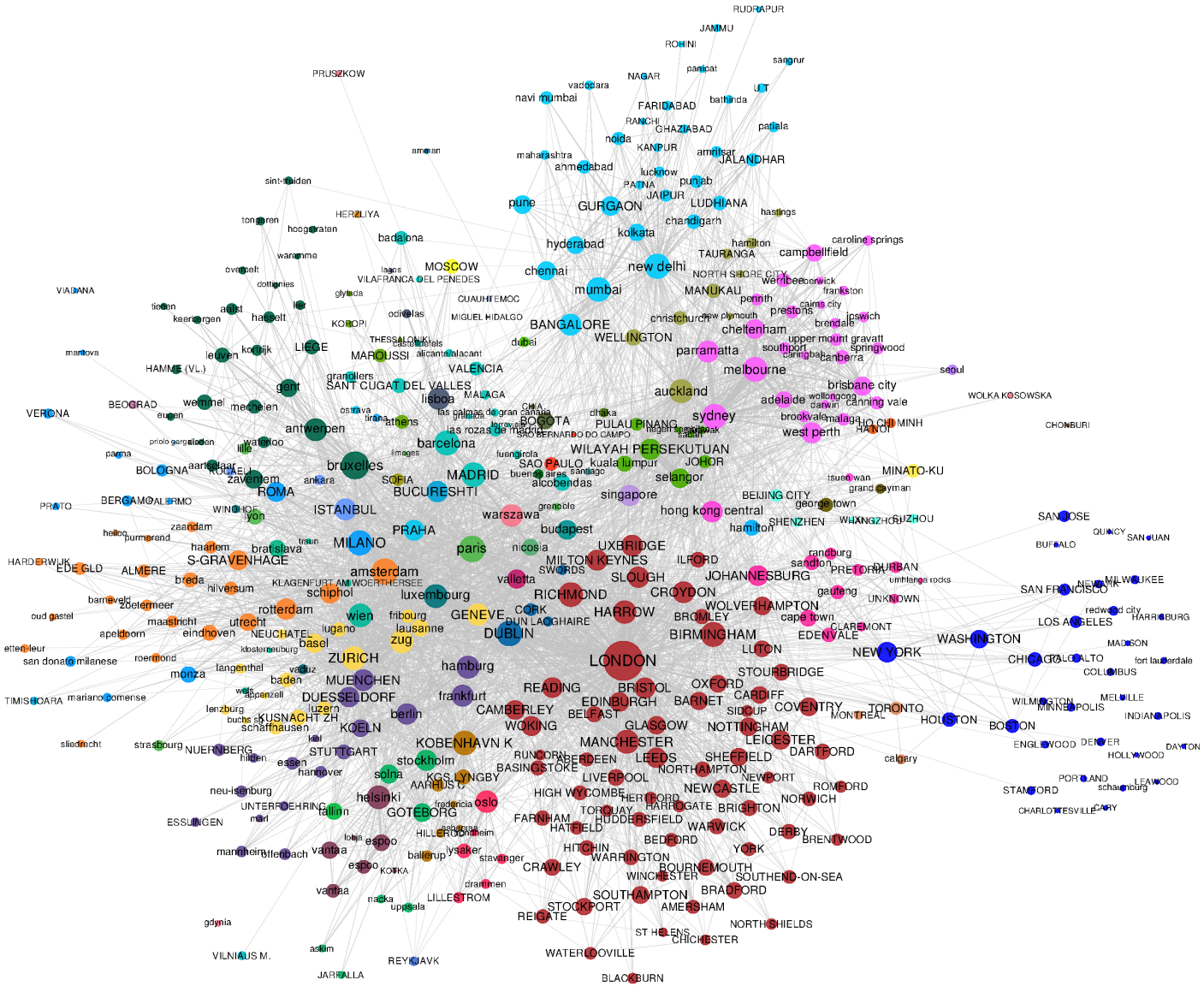
E.M. Heemskerk and F.W. Takes, The Corporate Elite Community Structure of Global Capitalism, *New Political Economy* 21(1): 90-118, 2016.

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.

Nodes colored by country (sample)



Community composition

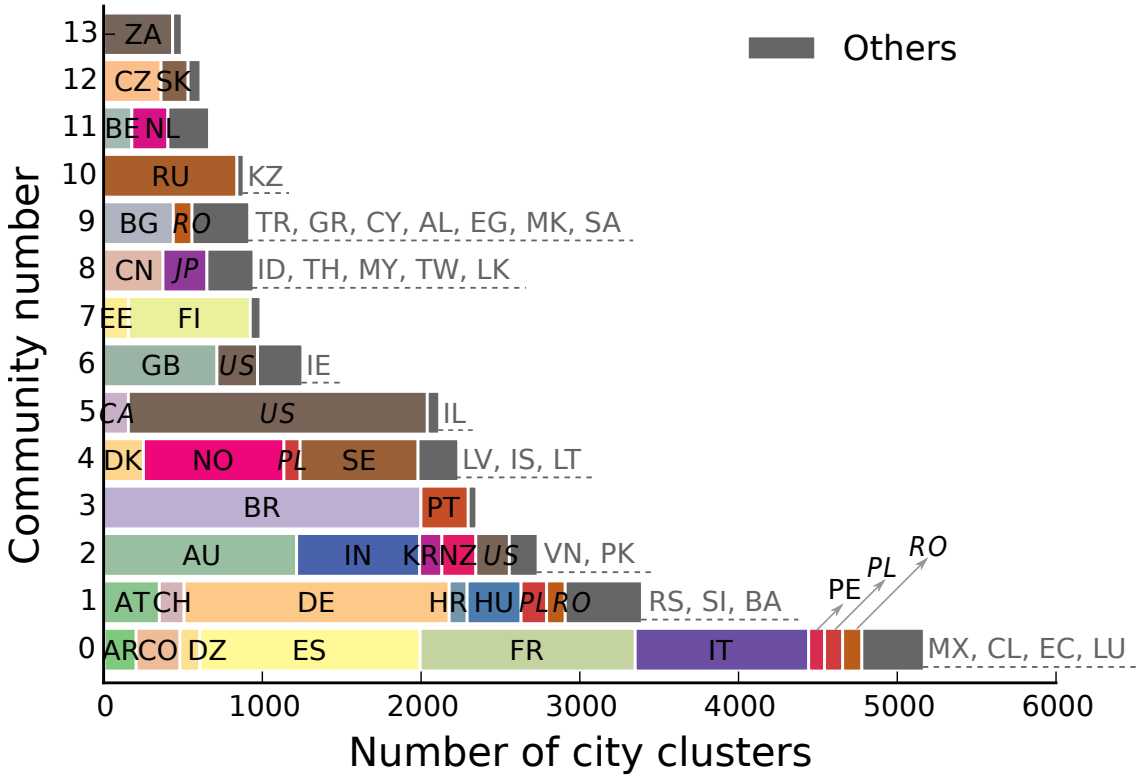


Figure: Country involved in each community

Co-citation network

- Nodes are scientific publications
- Edges indicate that papers cite the same previous work
- Each node has an associated scientific field
- Network provides insight in how scientific fields interact

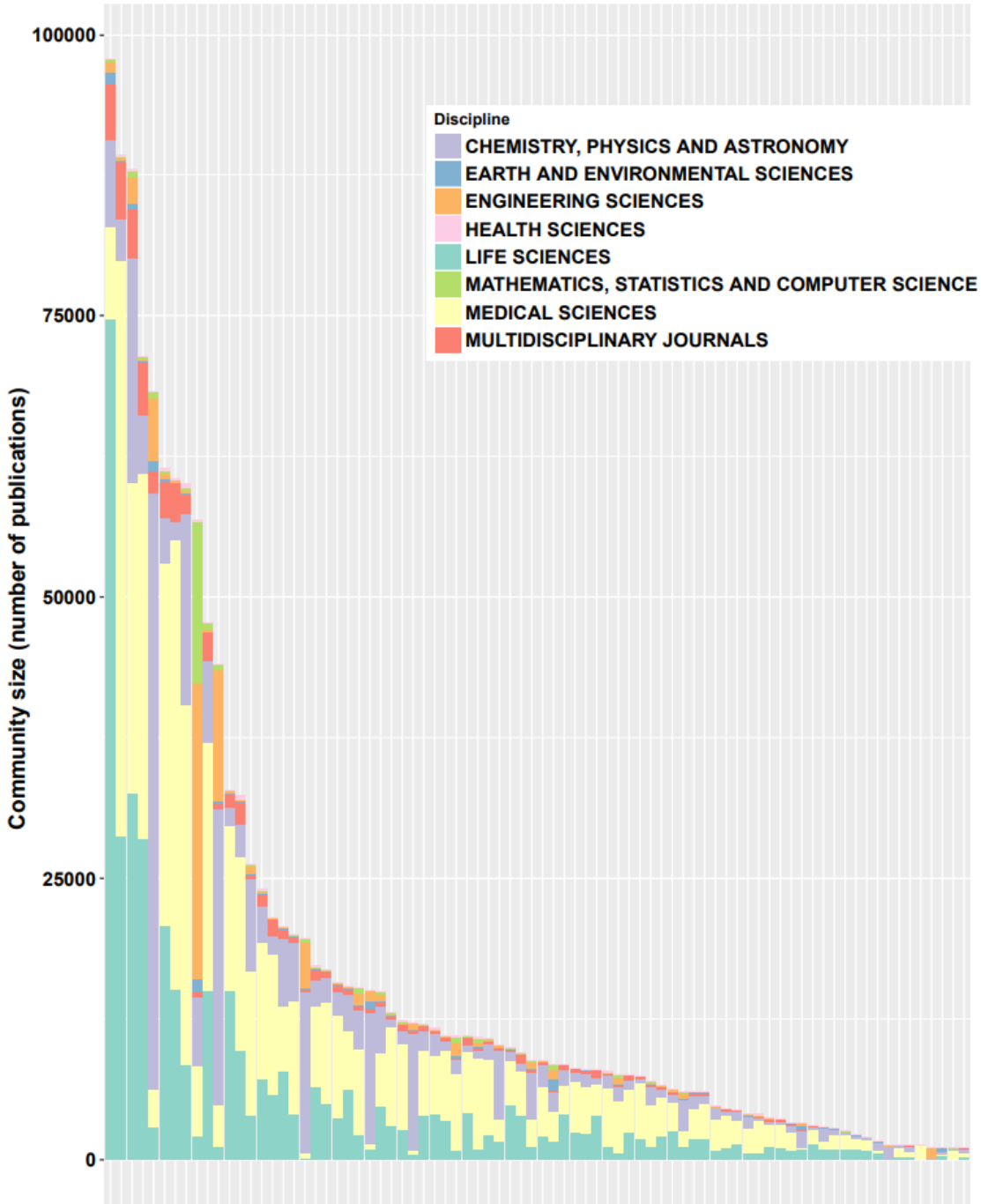
Co-citation network

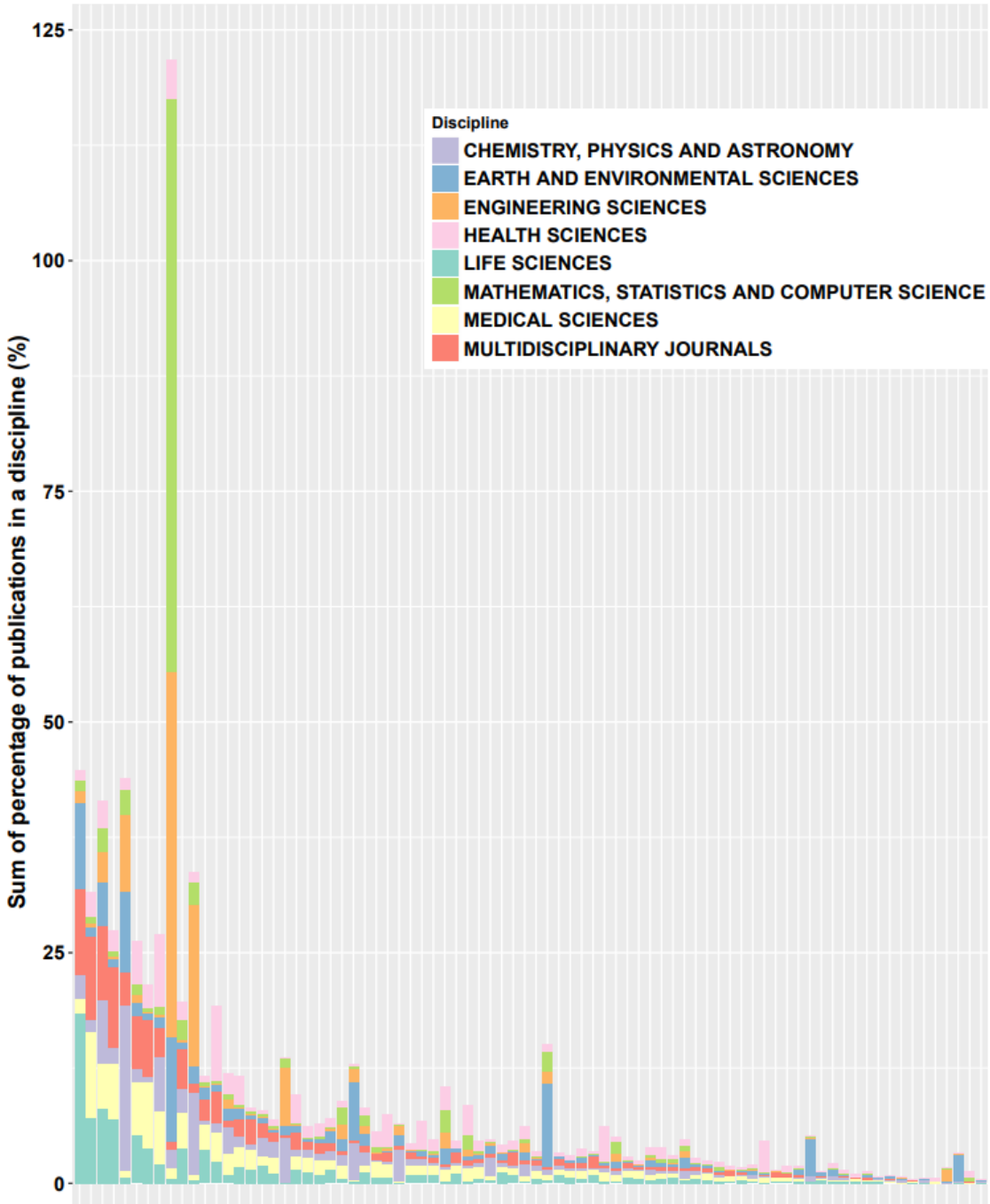
- Nodes are scientific publications
- Edges indicate that papers cite the same previous work
- Each node has an associated scientific field
- Network provides insight in how scientific fields interact
- Size: 1.6 million nodes and 44 million edges
- 99% in giant component, scale-free, small-world

Co-citation network

Discipline	Number of publications
MEDICAL SCIENCES	550672.68
LIFE SCIENCES	403633.85
CHEMISTRY, PHYSICS AND ASTRONOMY	293971.77
ENGINEERING SCIENCES	66186.33
MULTIDISCIPLINARY JOURNALS	55394.00
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE	23192.52
EARTH AND ENVIRONMENTAL SCIENCES	10596.43
HEALTH SCIENCES	5043.42

Figure: Categories of publications (weighting applied if multiple apply)


Discipline
CHEMISTRY, PHYSICS AND ASTRONOMY
EARTH AND ENVIRONMENTAL SCIENCES
ENGINEERING SCIENCES
HEALTH SCIENCES
LIFE SCIENCES
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
MEDICAL SCIENCES
MULTIDISCIPLINARY JOURNALS
Community size (number of publications)
100000
75000
50000
25000
0


125
100
75
50
25
0
Sum of percentage of publications in a discipline (%)
Discipline
CHEMISTRY, PHYSICS AND ASTRONOMY
EARTH AND ENVIRONMENTAL SCIENCES
ENGINEERING SCIENCES
HEALTH SCIENCES
LIFE SCIENCES
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
MEDICAL SCIENCES
MULTIDISCIPLINARY JOURNALS

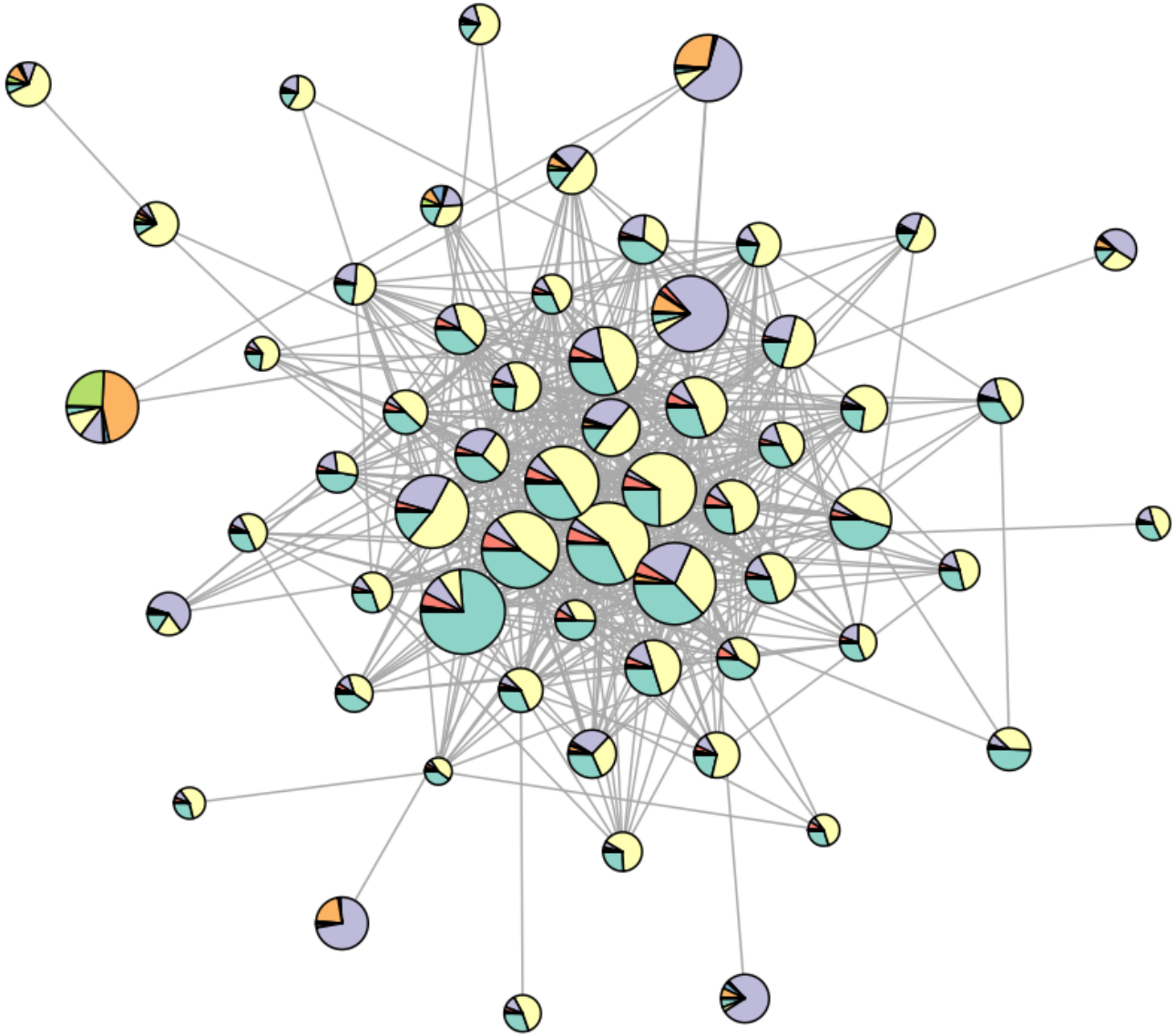


Figure: Community composition and connections

Universiteit Leiden

The web

World wide web

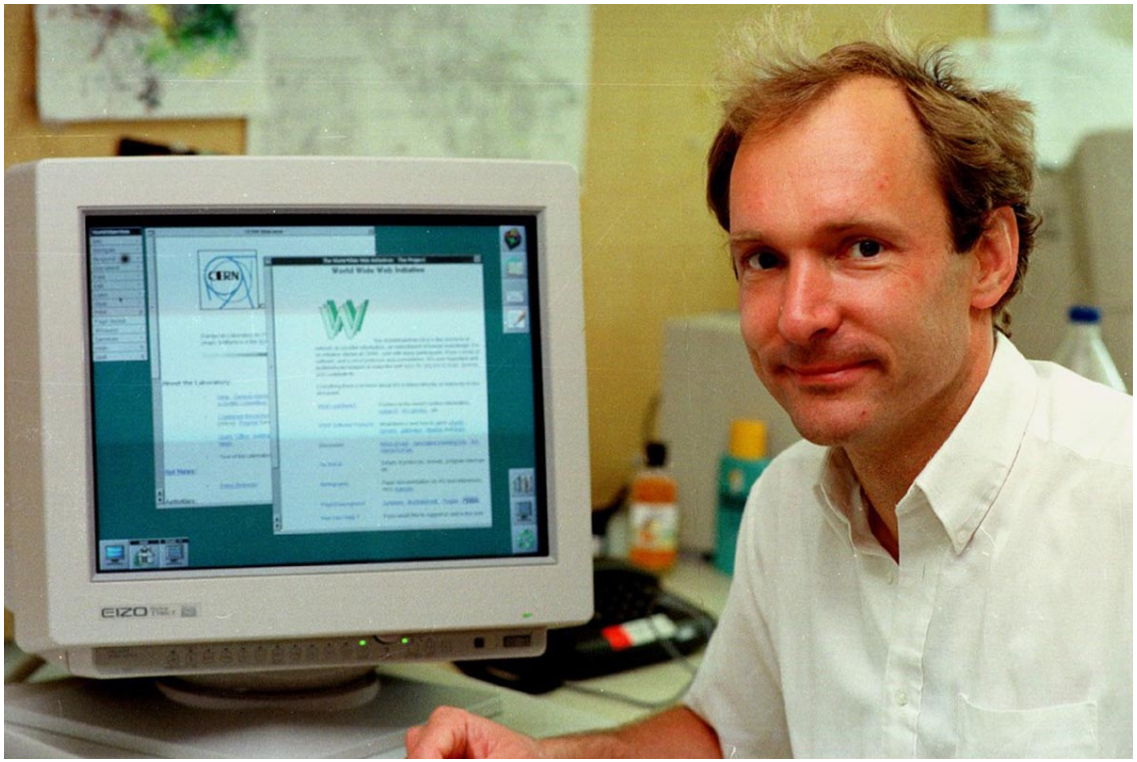


Figure: Tim Burners-Lee

Turing award

Figure: ACM Turing Award 2016

World wide web

- Around since 1990
- Chaos of webpages: how to order?

World wide web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories

World wide web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services

World wide web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services
- Search engines based on term frequency

World wide web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services
- Search engines based on term frequency
  - Keyword stuffing

World wide web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services
- Search engines based on term frequency
  - Keyword stuffing
- Search engines based on "smart" webpage ranking
  - So how?

Universiteit Leiden

HITS

Centrality measures

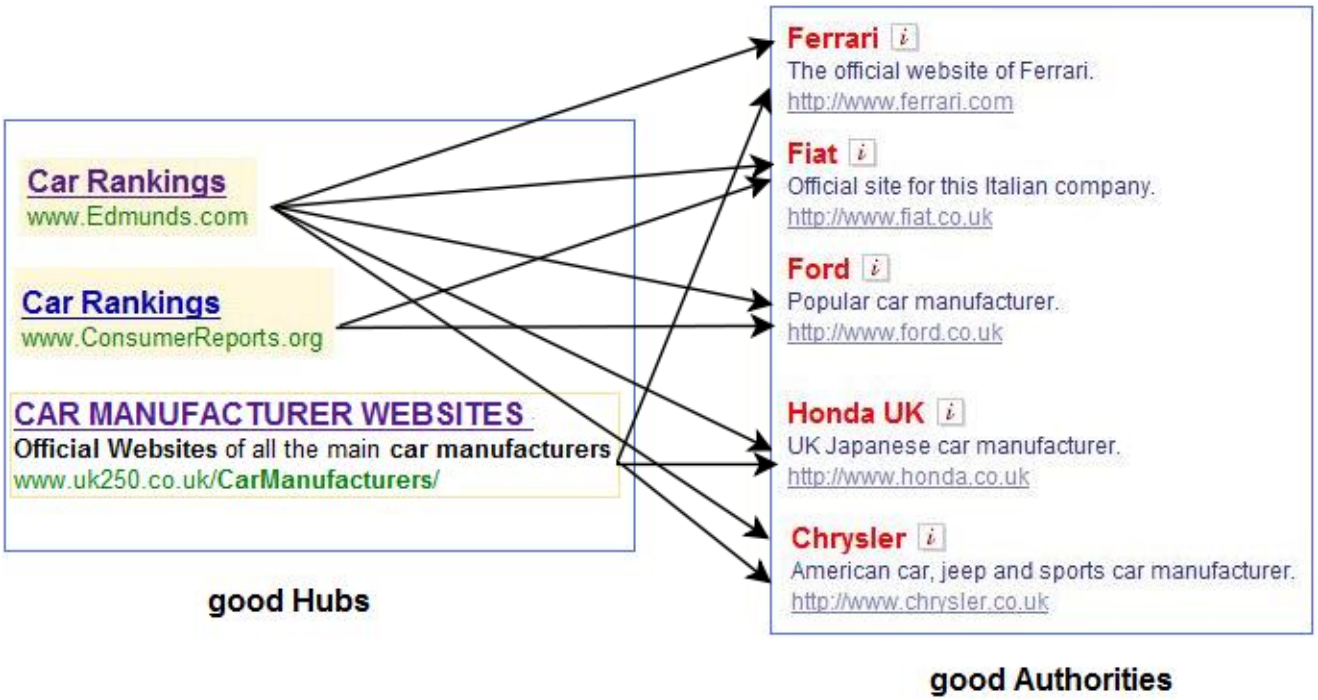
- Distance/path-based measures:
  - Degree centrality $O(n)$
  - Closeness centrality $O(mn)$
  - Betweenness centrality $O(mn)$
  - Eccentricity centrality $O(mn)$
- **Propagation-based** measures:
  - Hyperlink Induced Topic Search (HITS)
  - PageRank

Hyperlink Induced Topic Search

Universiteit Leiden

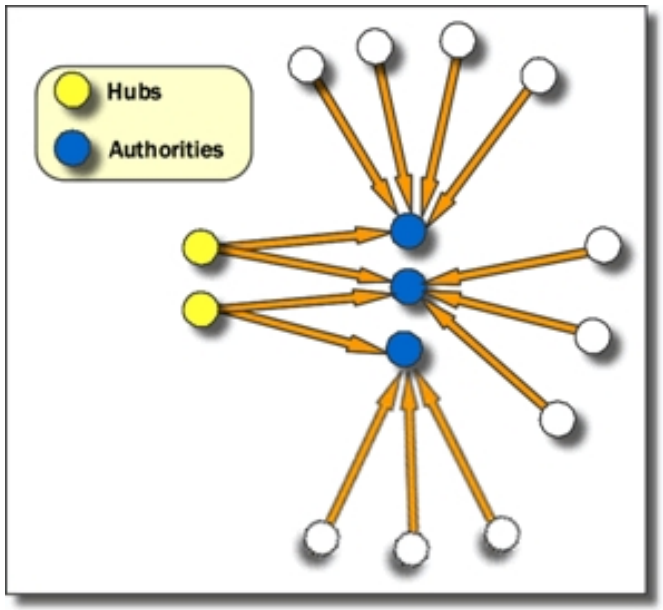
- A link to a page is a "vote" for that page
- But how important is the page casting the vote?
- **Hyperlink Induced Topic Search (HITS)**
- **Hubs**: pages that link to good authorities
- **Authorities**: contain useful information and are therefore linked from many good hubs
- Jon Kleinberg, Authoritative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

Hyperlink Induced Topic Search



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)

Universiteit Leiden

- 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

Universiteit Leiden

PageRank

PageRank

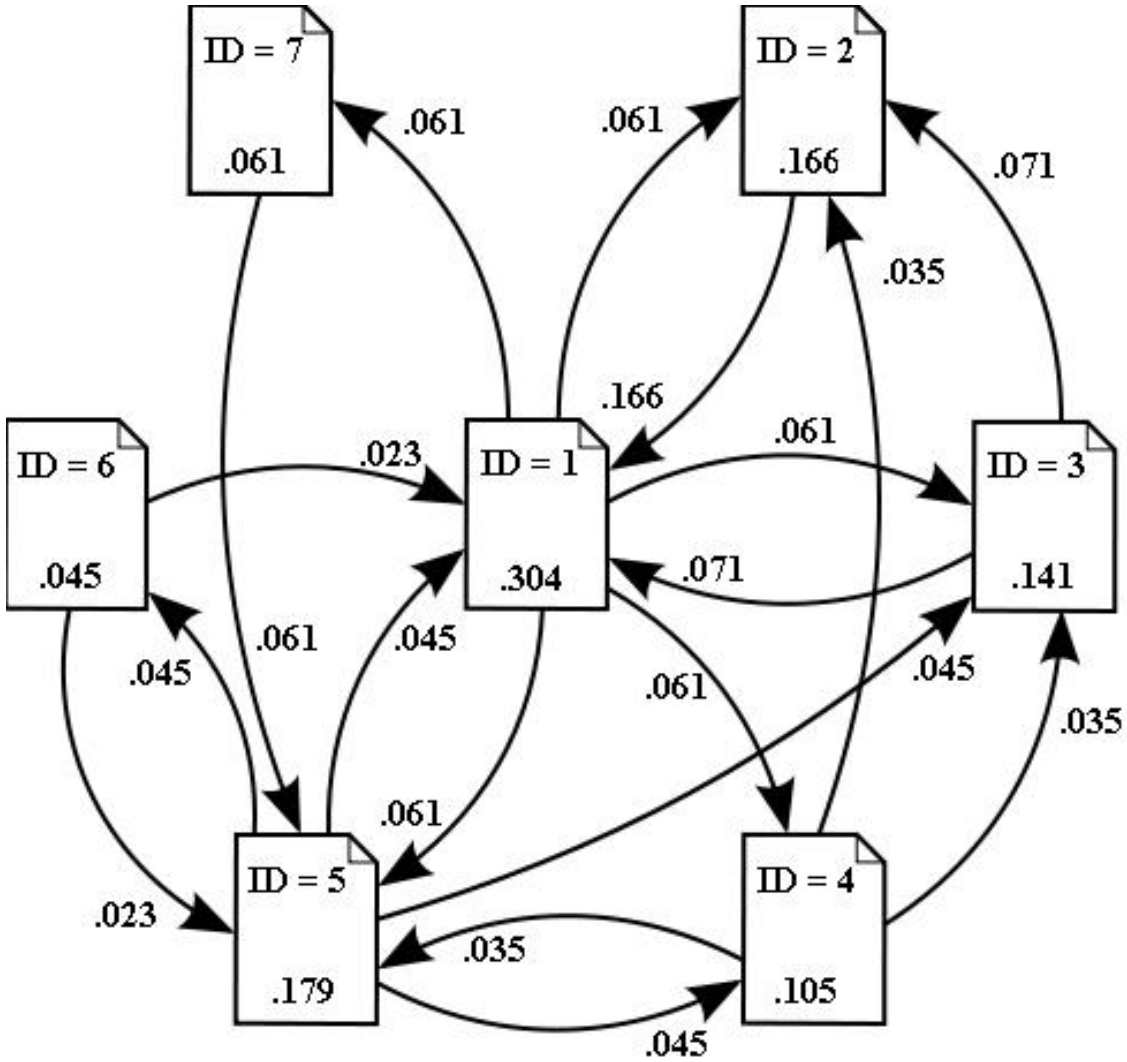
Universiteit Leiden

- 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?

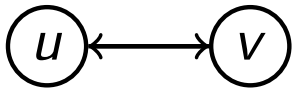
Towards PageRank example



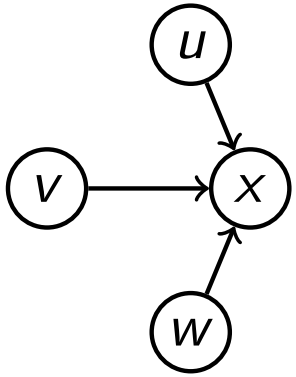
Challenges

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

- **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

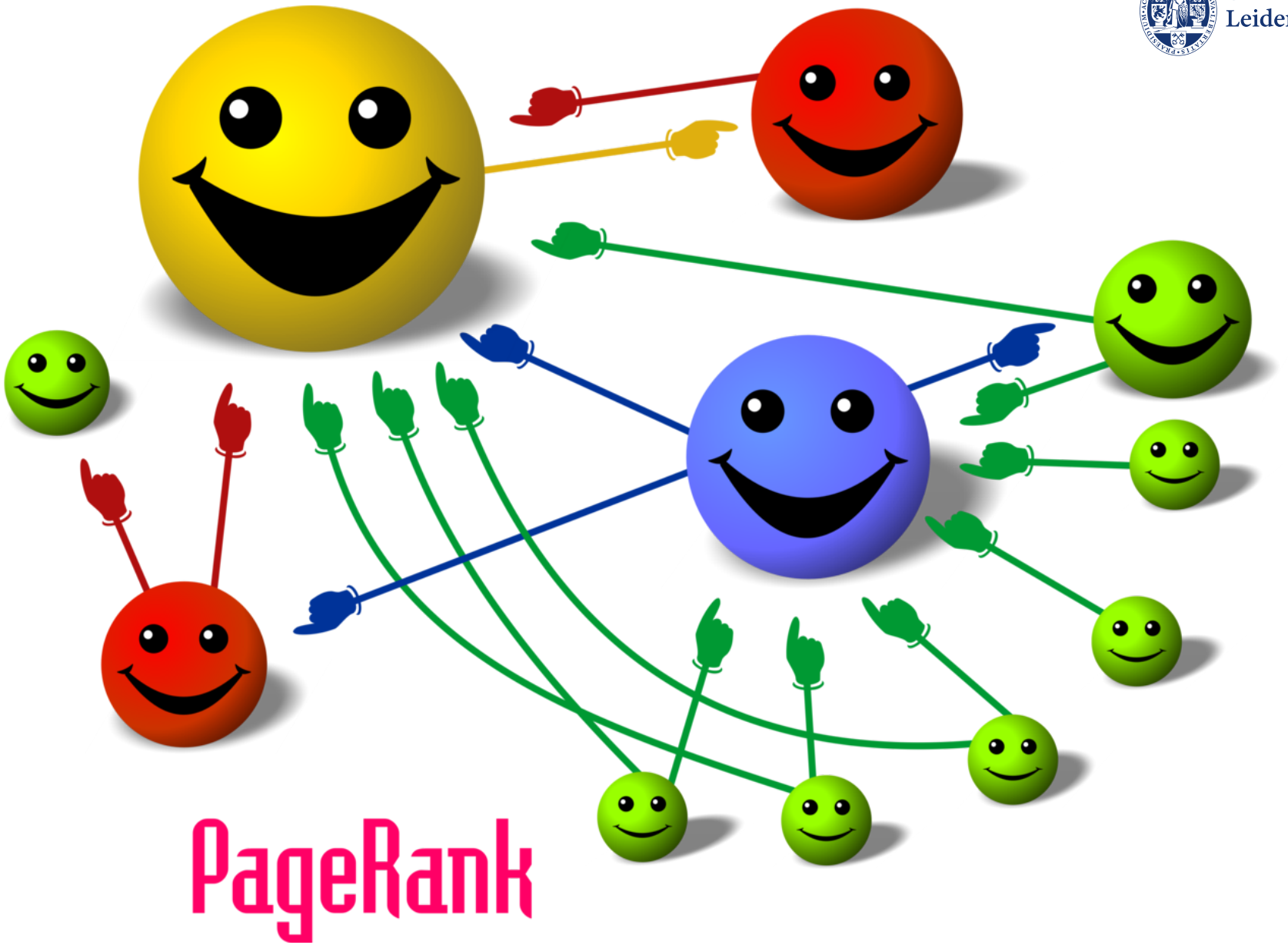
- For all nodes $v \in V$, initialize $PR^0(v) = (1/n)$
- $t = 0$
- Repeat:
  1. $t = t + 1$
  2. $PR^t(v) = \frac{1-p}{n} + p \cdot \sum_{w \in N'(v)} \frac{PR^{t-1}(w)}{outdeg(w)}$
  3. Normalize so that $\sum_{v \in V} PR(v) = 1$
     (just divide each value by the sum of all values)
- Until scores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

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

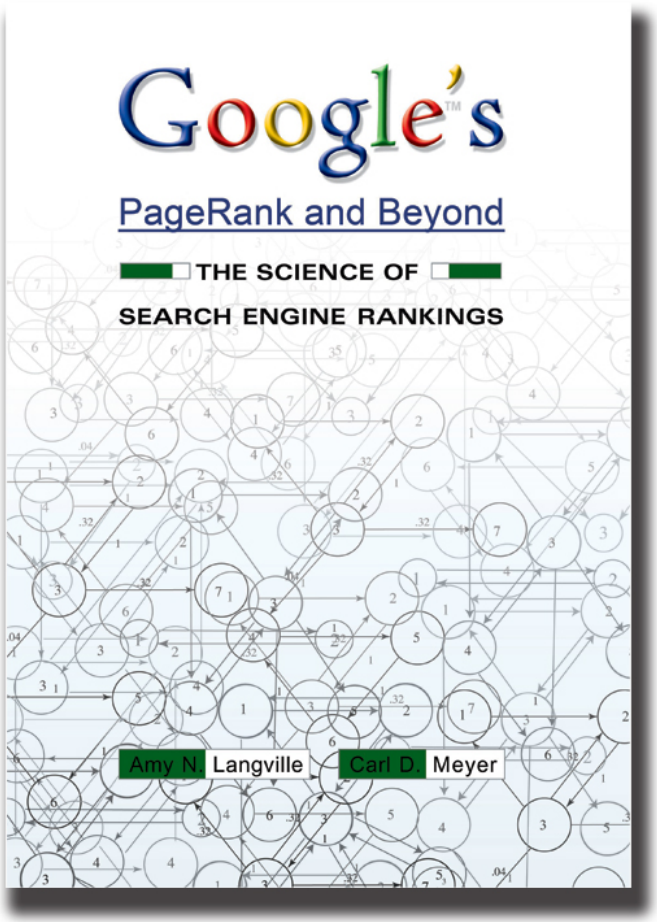
  for each of the nodes $v \in V$ and then normalizing the values so that they sum to 1, where $PR(v)$ is initialized to $1/n$ and $N'(v)$ is the set of nodes that links to node v and $p = 0.85$
- 100 iterations is usually enough
- Time $100 \cdot m$, similar to HITS.


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

Universiteit Leiden

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

Universiteit Leiden

- 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 $O(m)$

Universiteit Leiden

Structure of the web

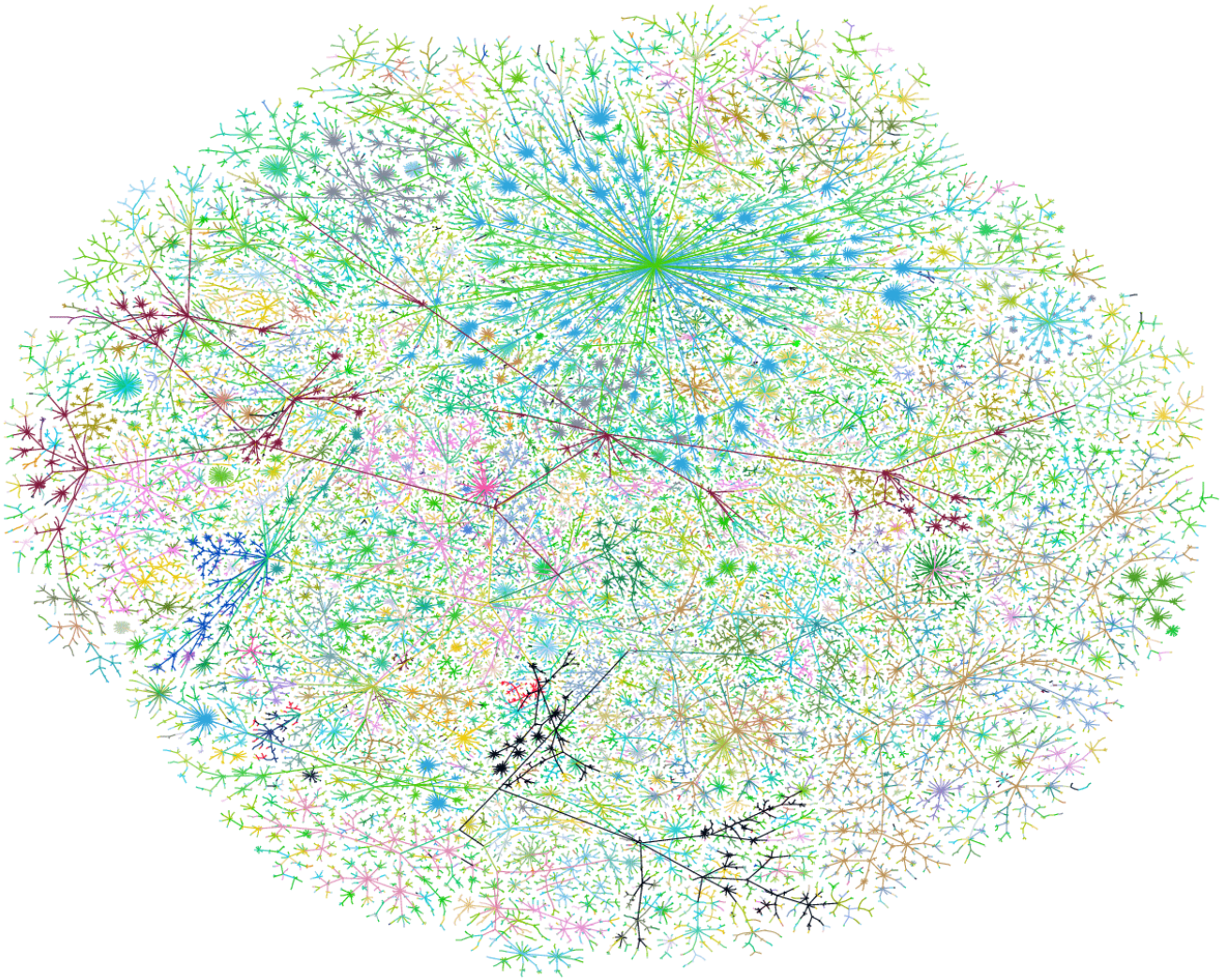
Webgraph

- **Webgraph**: directed unweighted network $G = (V, E)$
- Nodes V are webpages
- Links E are "hyperlinks" to other pages
- Many dense subgraphs ...

Webgraph

- **Webgraph**: directed unweighted network $G = (V, E)$
- Nodes V are webpages
- Links E are "hyperlinks" to other pages
- Many dense subgraphs . . . because pages (nodes) may belong to the same domain
- Alternative: draw webgraph with only (sub)domains as nodes, referred to as **host graph**
- Idea: search engine ranks webpages using the structure of the webgraph
- Centrality measures

The web



http://www.cheswick.com/ches/map/gallery/index.html

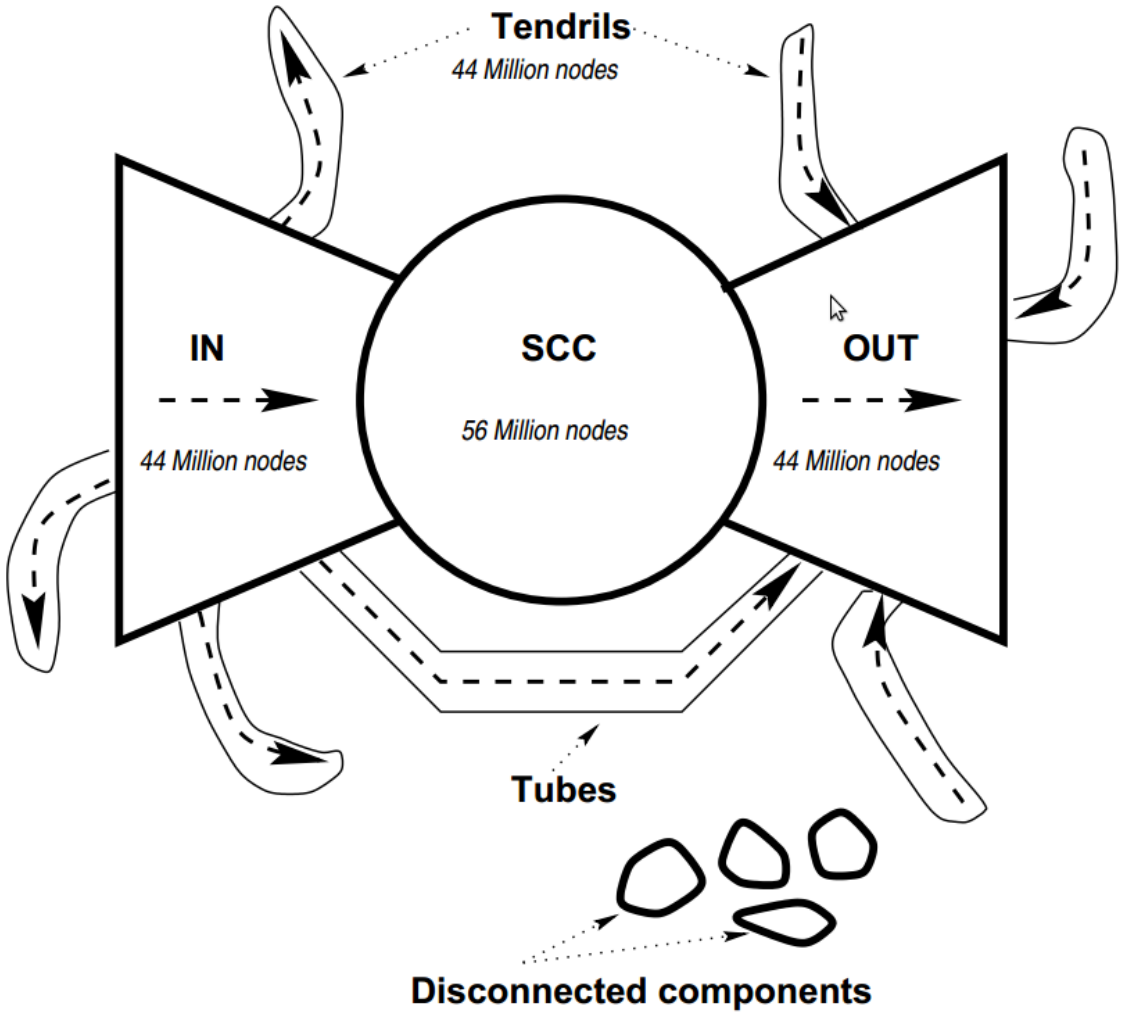
Why study the webgraph?

Universiteit Leiden

- Understanding social mechanisms that govern growth
- Designing ranking methods
- Devising better crawling algorithms
- Creating accurate models of the web's structure

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Webgraph in 1999



Broder et al., Graph structure in the web, Computer Networks 33(1): 309–320, 2000.

Webgraph in 1999

- Altavista: 200 million nodes
- 186 million nodes in the weakly connected component (90% of the links)
- 56 million nodes in the strongly connected component
- Power law degree distribution
- Average distance of 16 (if there is a path, 25% of the cases)
- Average (undirected) distance of 6.83 (small world!)
- Diameter is 28

Broder et al., Graph structure in the web, Computer Networks 33(1): 309–320, 2000.

Crawling the webgraph in 2012

- Crawled by Common Crawl Foundation
- First half of 2012
- Breadth-first visiting strategy
- Heuristics to detect spam pages
- Seeded with the list of domains from a previous crawl and a set of URLs from Wikipedia

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Webgraph in 2012

- Page graph
- Host graph
- PLD graph (Pay-Level Domain (PLD): a subdomain of a public top-level domain, for which users have to pay. PLDs identify a single user or organization)

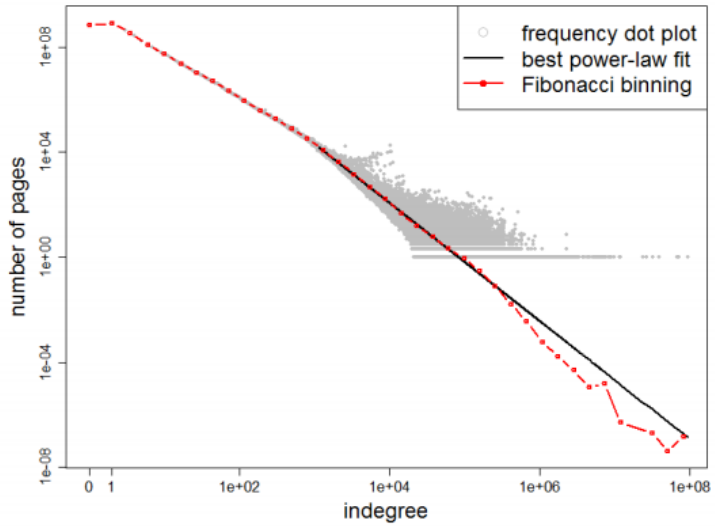
Granularity	# Nodes in millions	# Arcs in millions
Page Graph	3 563	128 736
Host Graph	101	2 043
PLD Graph	43	623

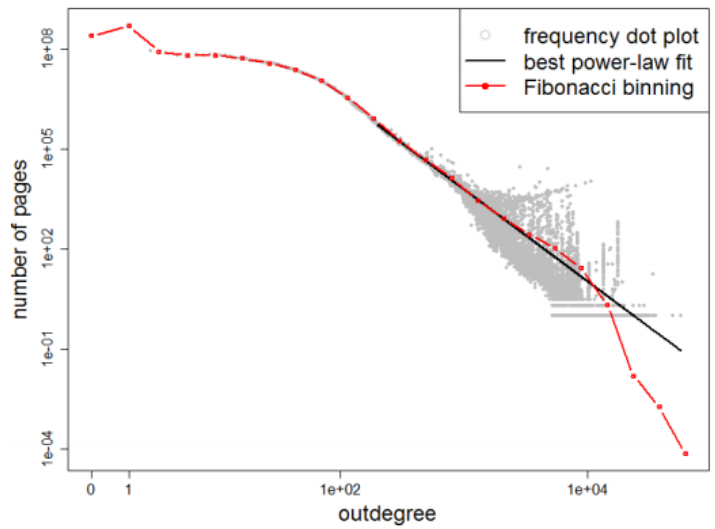
Table 1: Sizes of the graphs

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Degree

- Compared to 1999, average degree increased from 7.5 to 36.8
- Perhaps due to use of content management systems (they tend to create dense websites)





Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

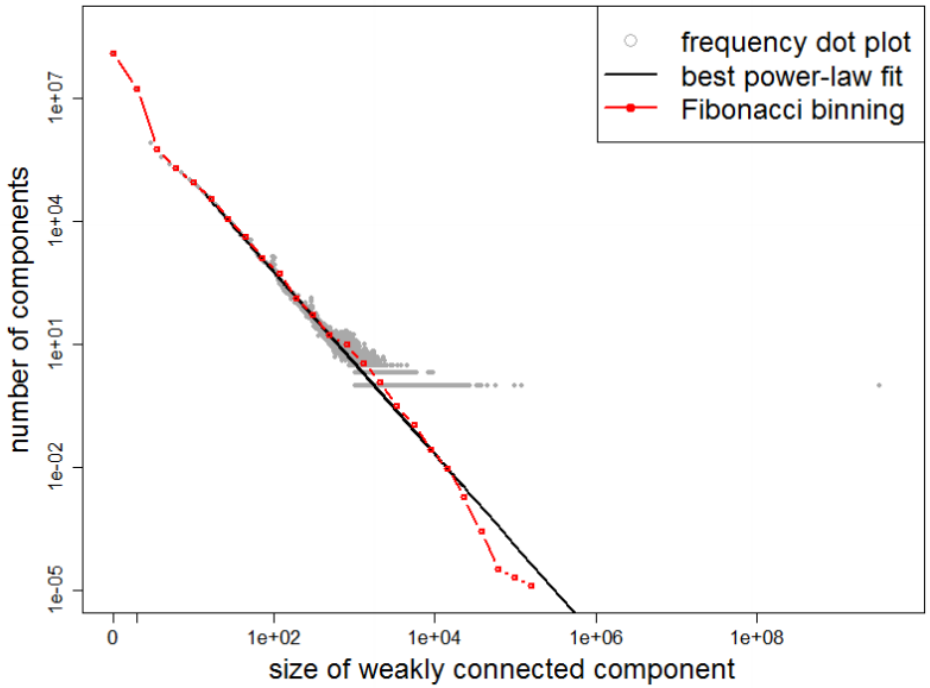
Centrality in the webgraph

PageRank	Indegree	Harmonic Centrality
gmpg.org	wordpress.org	youtube.com
wordpress.org	youtube.com	en.wikipedia.org
youtube.com	gmpg.org	twitter.com
livejournal.com	en.wikipedia.org	google.com
tumblr.com	tumblr.com	wordpress.org
en.wikipedia.org	twitter.com	flickr.com
twitter.com	google.com	facebook.com
networkadvertising.org	flickr.com	**apple.com**
promodj.com	**rtalabel.org**	vimeo.com
skriptmail.de	wordpress.com	creativecommons.org
parallels.com	**mp3shake.com**	**amazon.com**
tistory.com	w3schools.com	**adobe.com**
google.com	domains.lycos.com	**myspace.com**
miibeian.gov.cn	**staff.tumblr.com**	**w3.org**
phpbb.com	**club.tripod.com**	**bbc.co.uk**
blog.fc2.com	creativecommons.org	**nytimes.com**
tw.yahoo.com	vimeo.com	**yahoo.com**
w3schools.com	miibeian.gov.cn	**microsoft.com**
wordpress.com	facebook.com	**guardian.co.uk**
domains.lycos.com	phpbb.com	**imdb.com**

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

WCC of the webgraph

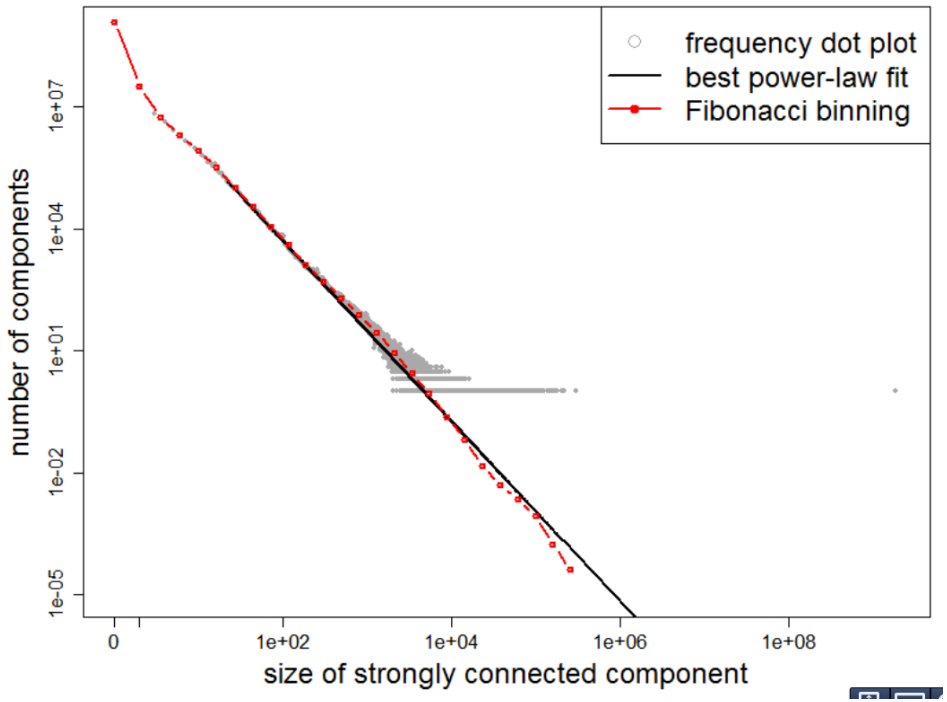
- Weakly Connected Component (WCC)
- 91.8% in 1999, 94% in 2012
- Component size distribution



Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

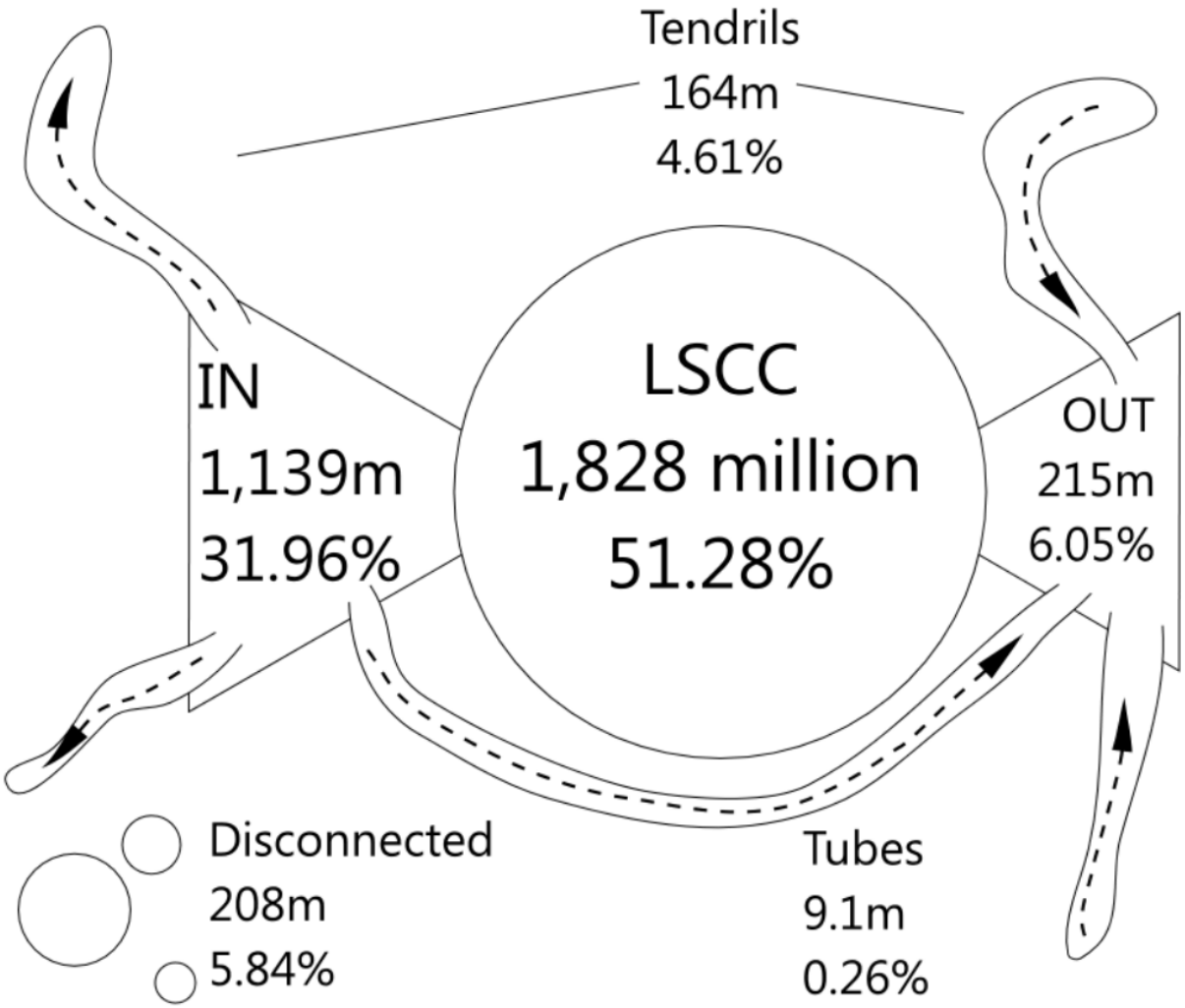
SCC of the webgraph

- Strongly Connected Component (SCC): 51.3% of the nodes
- Computation required 1TB of RAM
- Graph compression framework WebGraph was used



Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Bow-tie structure in 2012



Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Bow-tie structure in 2012

Component	Common Crawl 2012 # nodes (in thousands)	Common Crawl 2012 % nodes (in %)	Broder *et al.* # nodes (in thousands)	Broder *et al.* % nodes (in %)
LSCC	1 827 543	51.28	56 464	27.74
IN	1 138 869	31.96	43 343	21.29
OUT	215 409	6.05	43 166	21.21
TENDRILS	164 465	4.61	43 798	21.52
TUBES	9 099	0.26	-	-
DISC.	208 217	5.84	16 778	8.24

Table 3: Comparison of sizes of bow-tie components

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Distances and diameter

- Diameter lower bound: $5,282$

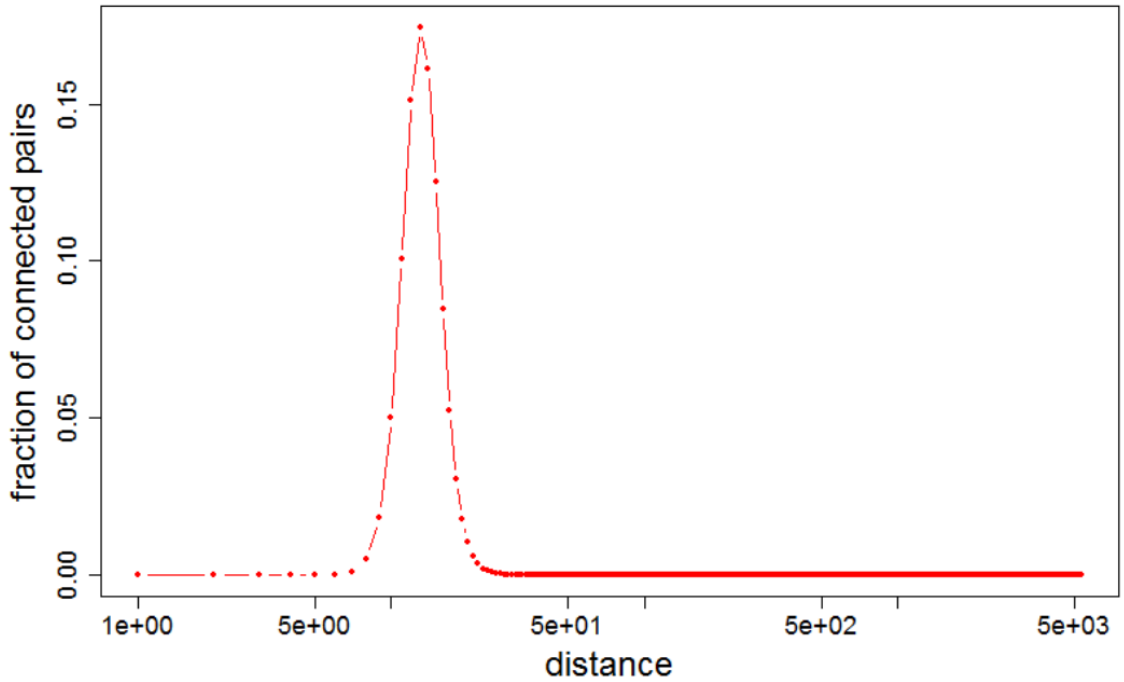


Figure: Distance distribution

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Webgraph conclusions

- Measurements on the largest webgraph available to the public
- Average degree has significantly increased, almost by a factor of 5
- Connectivity has increased, average distance has decreased
- Structure of the web appears dependent on the specific web crawl
- The distribution of indegrees and outdegrees is extremely different

Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Making a "better" webgraph

Universiteit Leiden

- Not just an unweighted unlabeled directed network
- **Resource Description Framework (RDF)**: link is a triple `[subject] [predicate] [object]`
- Link weighting: define a weight for outgoing links (to give hints to PageRank algorithm)
- Link annotation: make more use of the `rel=""` attribute to describe the kind of link: `alternate`, `search`, `next`, etc.
- Requires new algorithms for ranking ...

Universiteit Leiden

Walks in information networks

F.W. Takes and W.A. Kosters, Mining User-Generated Path Traversal Patterns in an Information Network, in Proceedings of the 12th IEEE/ACM International Conference on Web Intelligence (WI 2013), pp. 284-289, IEEE, 2013.
F.W. Takes and W.A. Kosters, The Difficulty of Path Traversal in Information Networks, in Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2012), pp. 138-144, 2012.

Walks

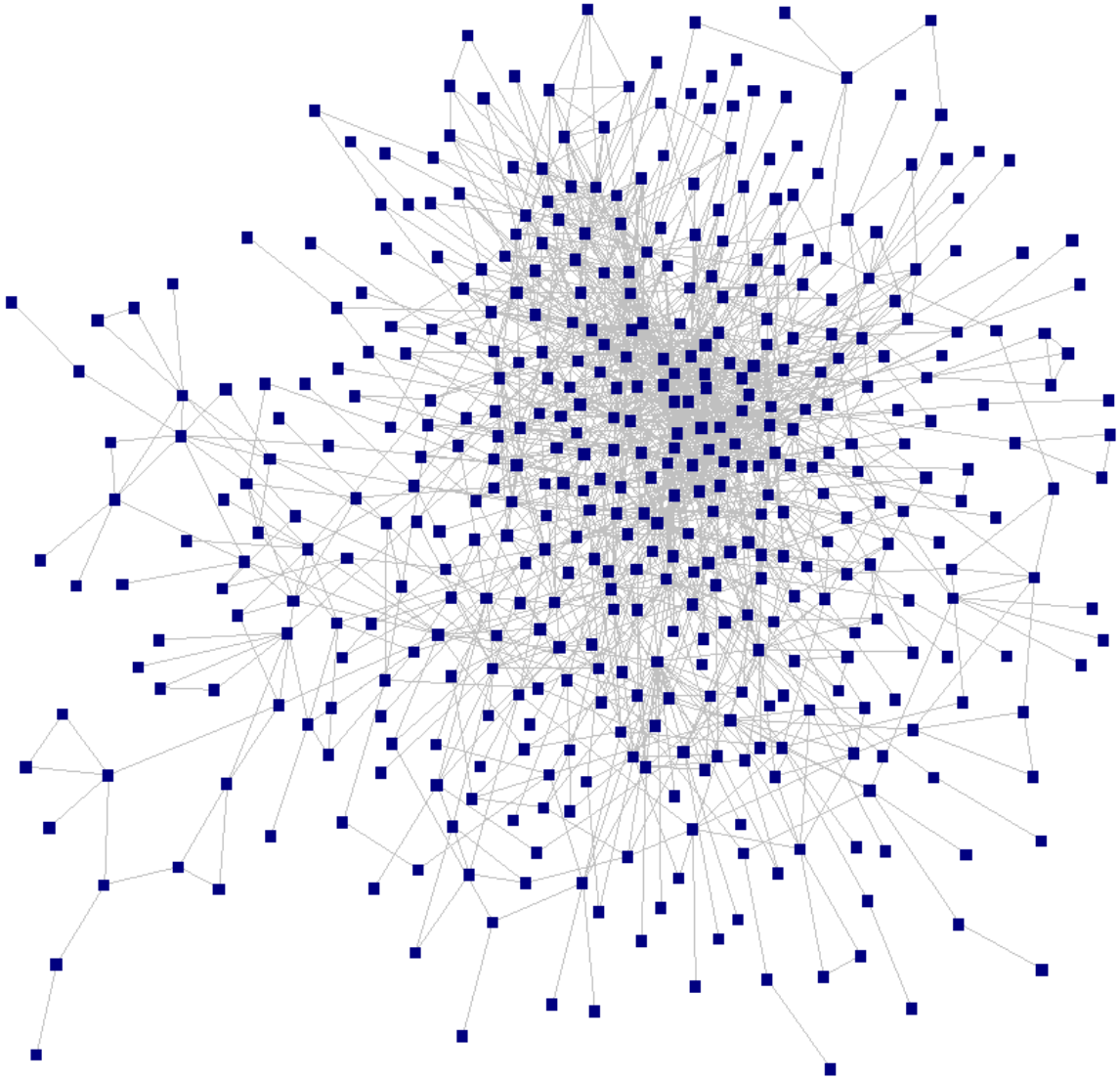
- Random walk
- Biased random walk
- **Targeted walk**

Walks

- Random walk
- Biased random walk
- **Targeted walk**
  - By humans
  - On Wikipedia

The Wiki Game

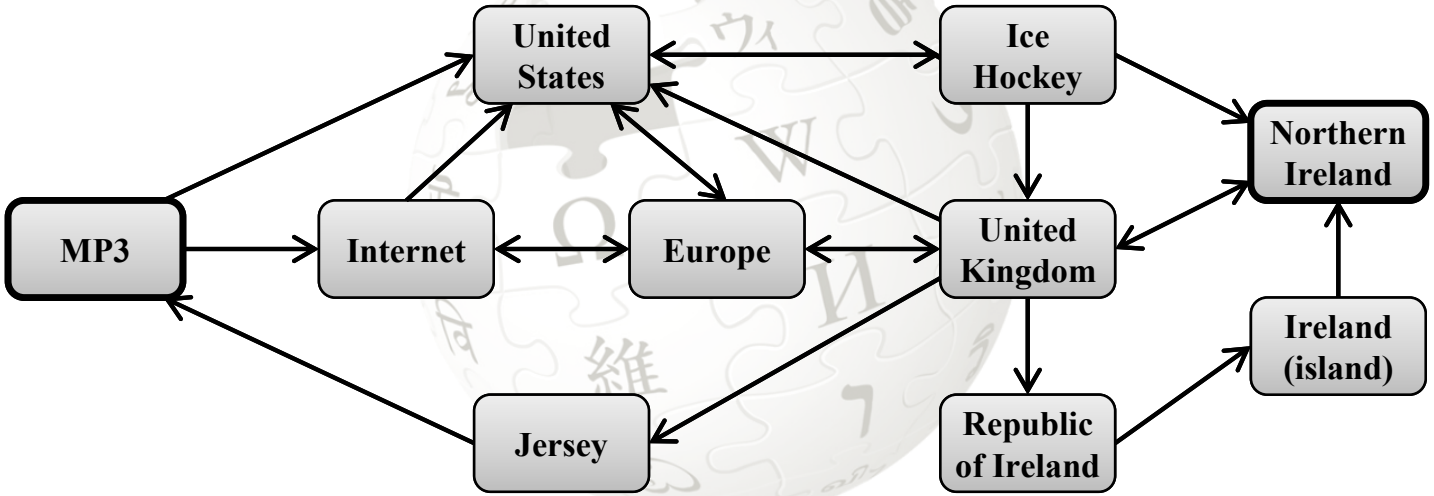
Information Networks



The topic explained

The Difficulty of Path Traversal in Information Networks

How hard is it for a human to find a path between two Wikipedia pages?



Motivation & Goals

- Search strategies
- Human search behavior
- Navigating the "Deep Web"
- Path traversal patterns
- Pattern mining

Concepts

- **Information network**: directed graph $G = (V, E)$ with n nodes and m links (Wikipedia)
- Path traversal task: navigate from u to v (with $u, v \in V$)
- Path: sequence of nodes connected by edges (u, a, b, c, a, d, v)
- Path length: number of traversed edges (6)
- Shortest path: minimum number of traversed edges (u, a, d, v)
- Distance: shortest path length $d(u, v) = 3$
- Failed path: $(u, a, b, c, u, a, e, a, d)$

Wikipedia dataset

- DBpedia 3.7 (`http://dbpedia.org`) from August 2011
- Small world network

Articles (n)	$3,464,902$
Directed links (m)	$82,019,786$
Largest WCC	99.9%
Average indegree	26
Average outdegree	22
Average distance ($\overline{d}$)	4.81
Effective diameter	7
Diameter	11

Table: Wikipedia dataset

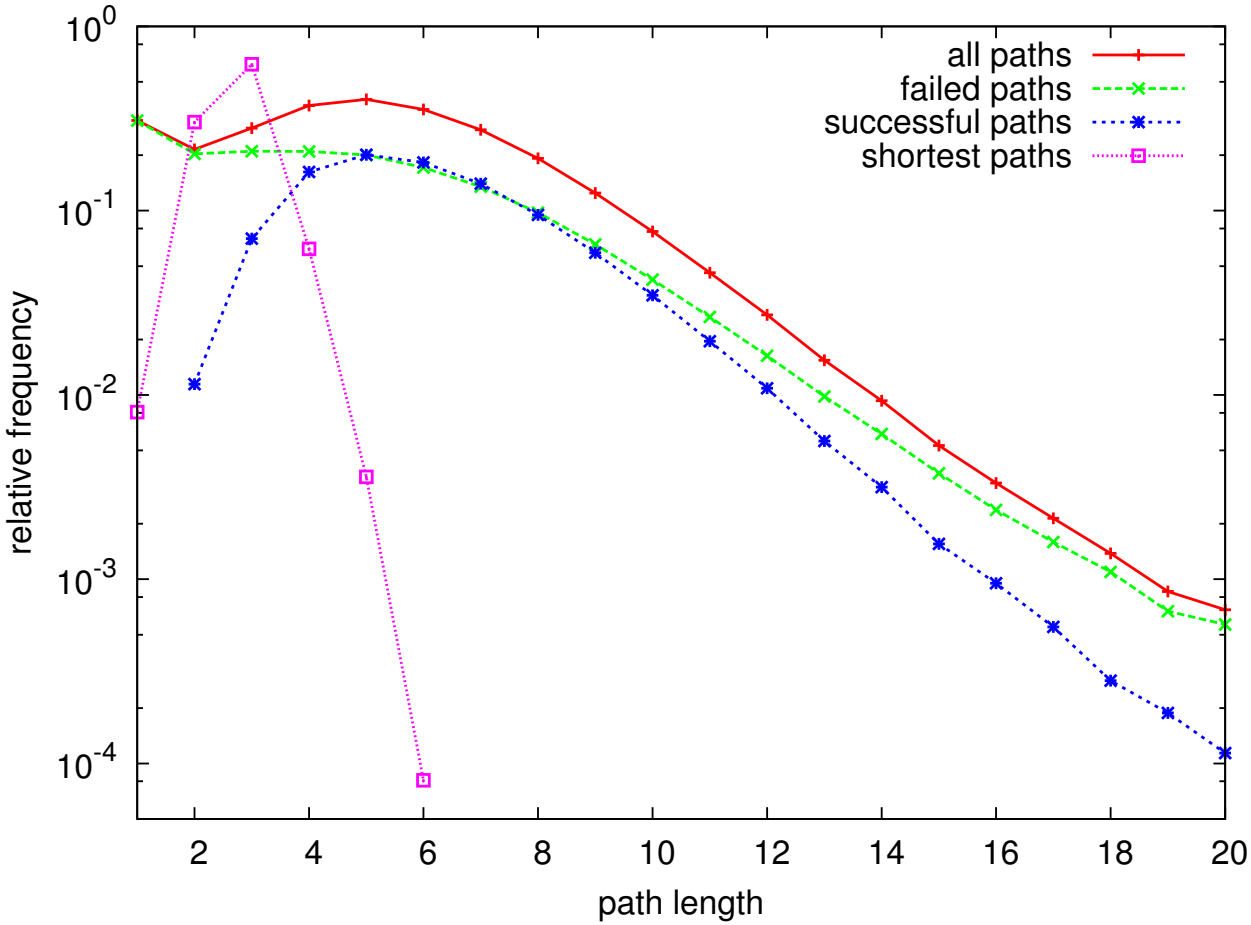
The Wiki Game dataset

- The Wiki Game (`http://thewikigame.com`)
- User-generated paths from 2009–2012
- Original set: $3,219,641$ paths consisting of $17,151,824$ clicks
- Only consider succesful paths (33.8%) of length 3–20.

User-generated paths	$1,137,337$
Clicks	$7,135,060$

Table: Wiki Game dataset

Successful and failed paths



Difficulty measures

- Start outdegree:

$$f_{outdeg}(u, v) = outdeg(u)$$

- Goal indegree:

$$f_{indeg}(u, v) = indeg(v)$$

- start 2-neighborhood size:

$$f_{|N_2|}(u, v) = |N_2(u)|$$

- goal reversed 2-neighborhood size:

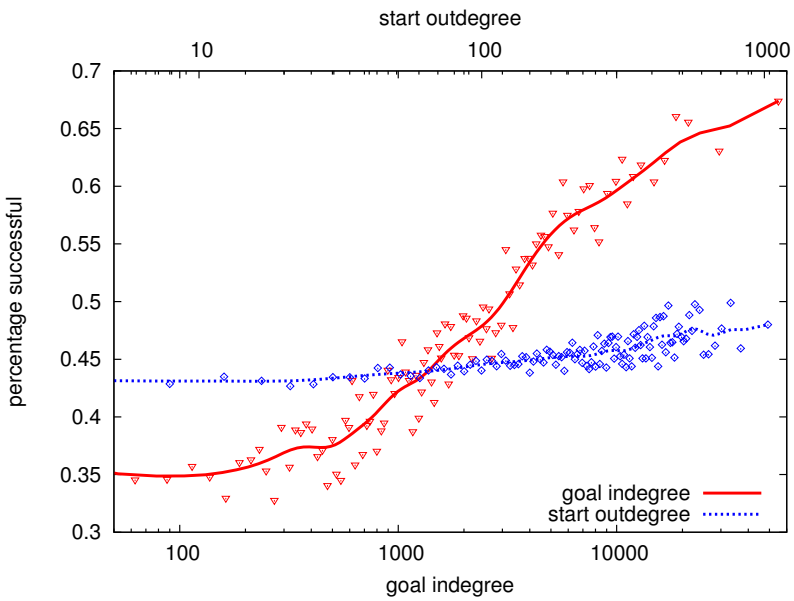
$$f_{|N_2'|}(u, v) = |N_2'(v)|$$

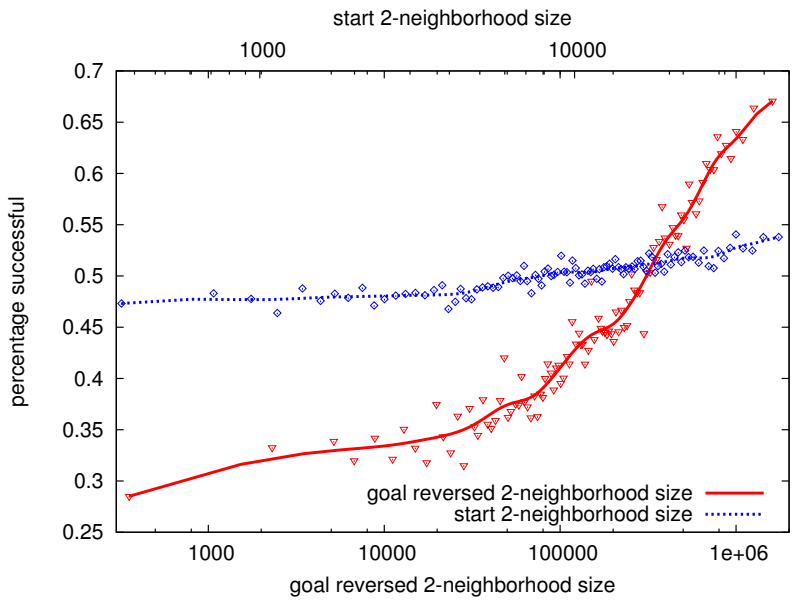
- Path length (distance):

$$f_d(u, v) = d(u, v)$$

Local difficulty measures

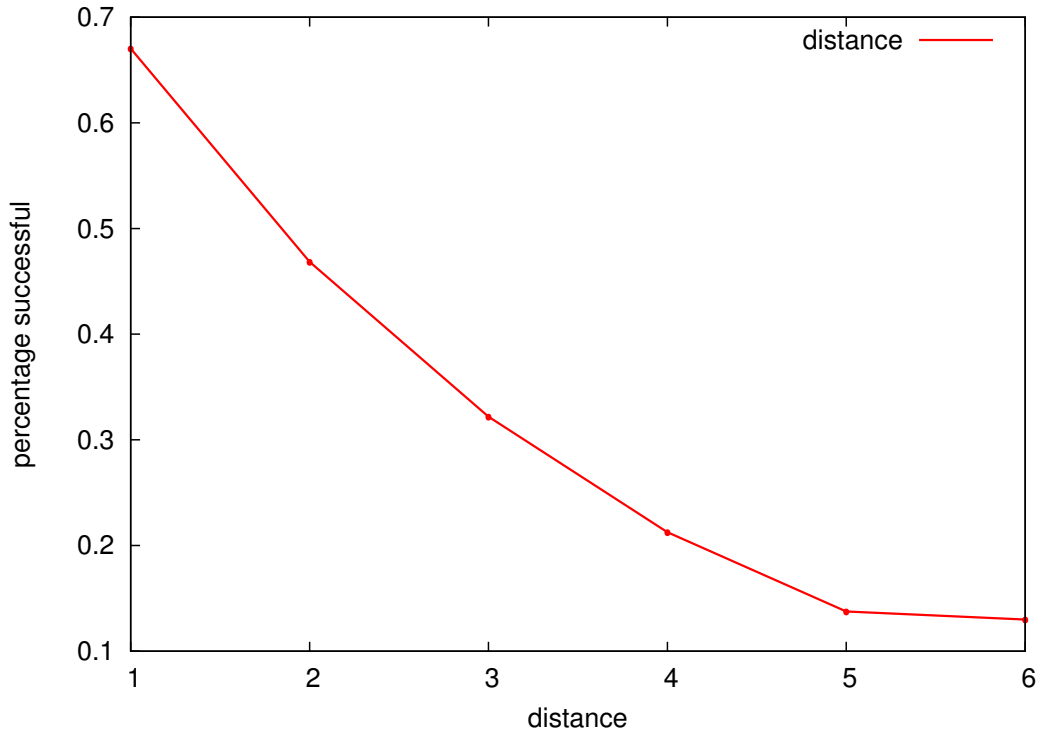
- Start page is of little influence
- Neighborhood of goal page is of high influence





Global measure: path length

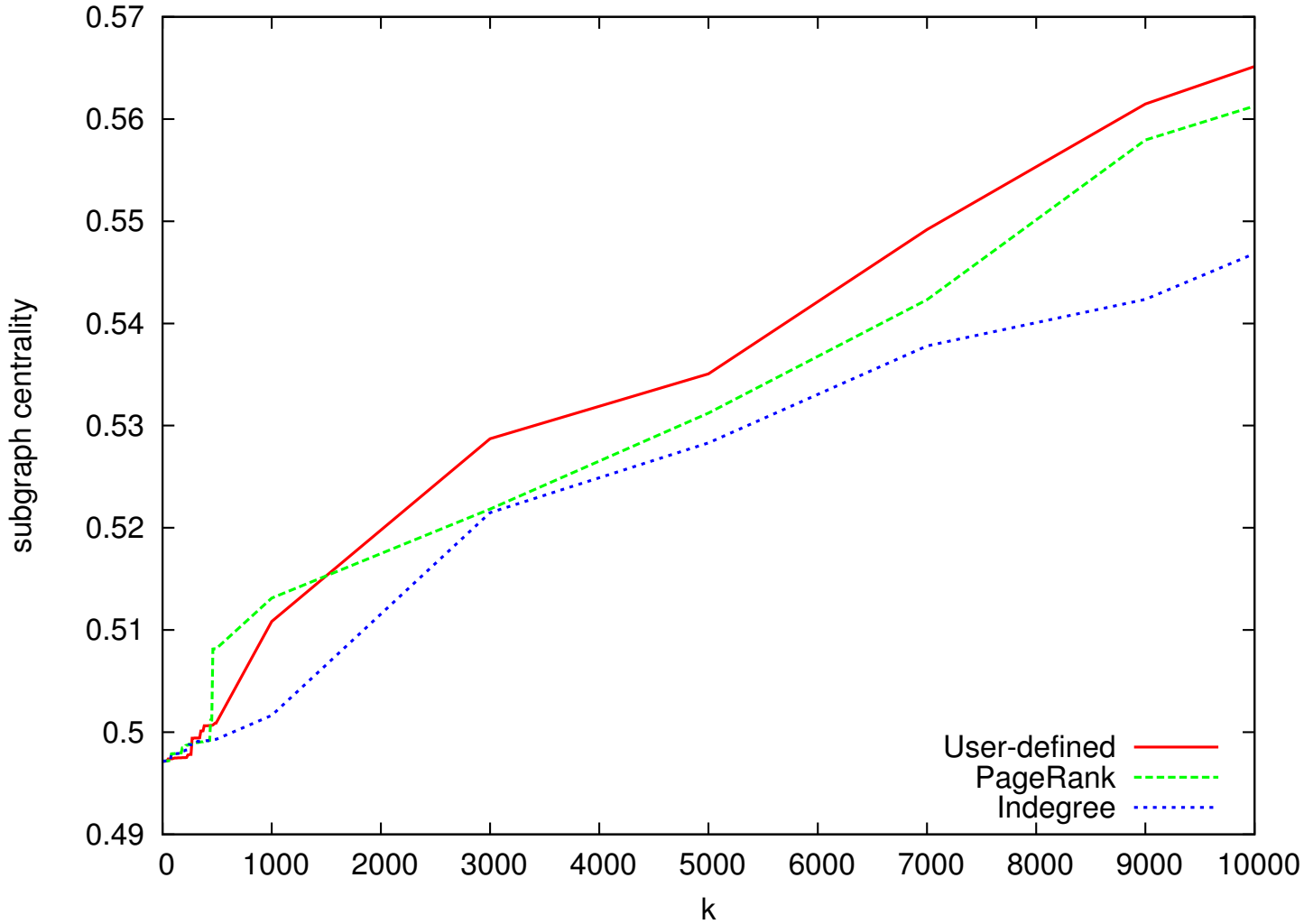
- Path length (distance) is directly related to path formation difficulty



Subgraph centrality

- Users are able to select a central set of nodes for reaching their navigation goals.
  But is this a good and efficient set?
- Node centrality: the importance of a single node
- **Subgraph centrality**: the importance of a set of nodes
- Subgraph consisting of the top-k user-defined central nodes
- Determine subgraph's closeness centrality, for top-k nodes of different centrality measures

Results: subgraph centrality



Findings

- Human paths are roughly twice as long as shortest paths
- Progress is easy close to start and goal
- High-level reasoning causes users to miss opportunities
- Hubs are crucial in the beginning, but
- Local properties of goal page explain difficulty
- Users are apparently able to select an efficient portion of the graph that is useful for traversing it

Lab session & "Homework"

- Lab session today: ask your final questions about Assignment 1
- Finalize Assignment 1
- After that: read your course project paper
- E-mail me somewhere in the coming week when you want to present; first come, first serve.
- Have a first look at Assignment 2