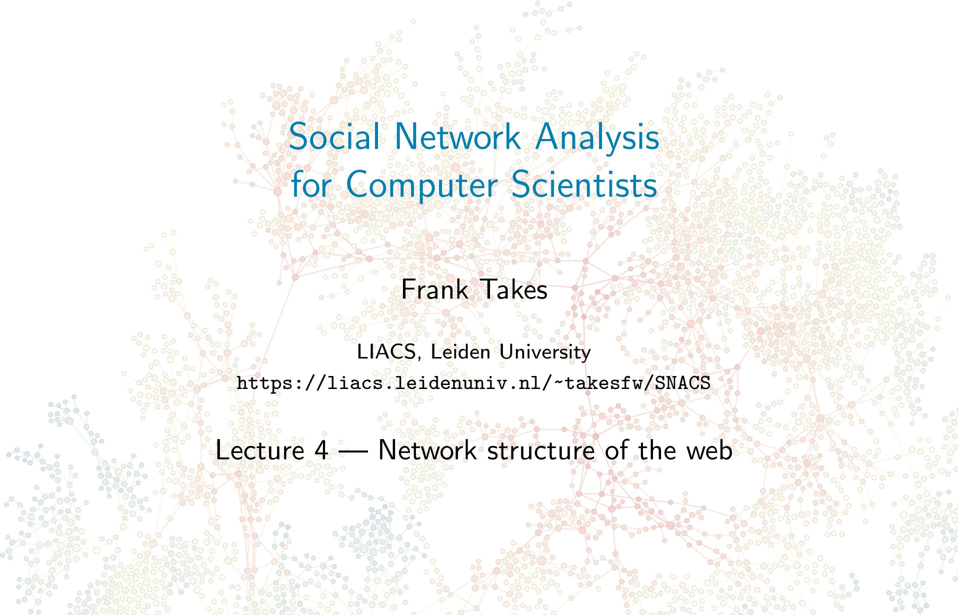


Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

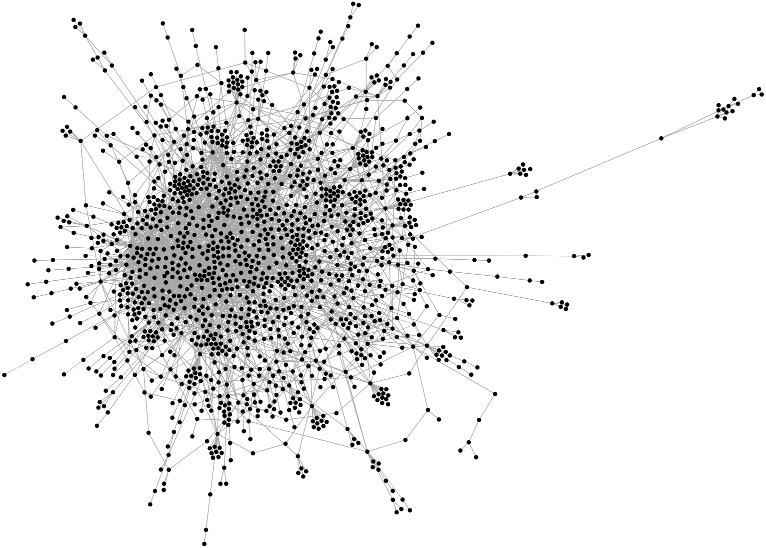
Lecture 4 — Network structure of the web

Today

- Recap
- Propagation-based centrality
- Structure of the web
- Data Science Lab
- Example presentation
- Planning of all presentations

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

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

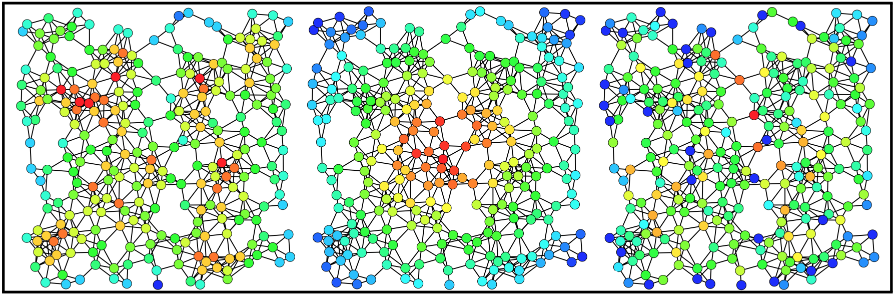


Figure: Degree, closeness and betweenness centrality

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

Community detection

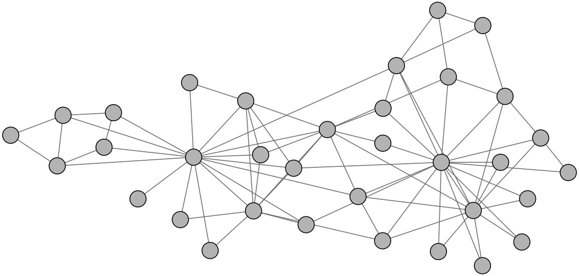


Figure: Communities: node subsets connected more strongly with each other

Community detection

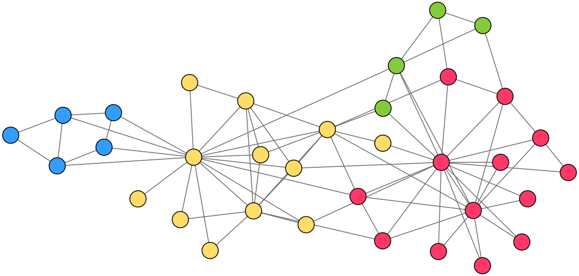


Figure: Communities: node subsets connected more strongly with each other

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.

The web

World wide web

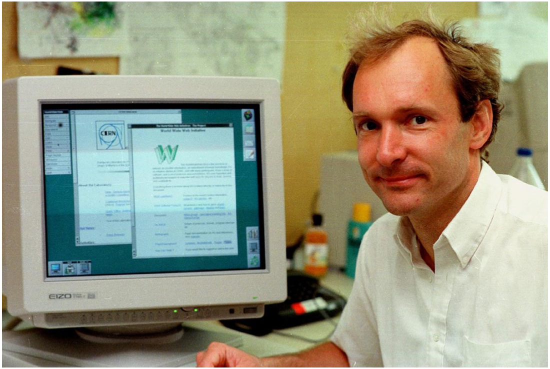


Figure: Tim Burners-Lee

Turing award

SIR TIM BERNERS-LEE
Inventor of the World Wide Web, the first web browser, and the protocols and algorithms allowing the Web to scale
acm
A.M. TURING AWARD 2016

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?

HITS

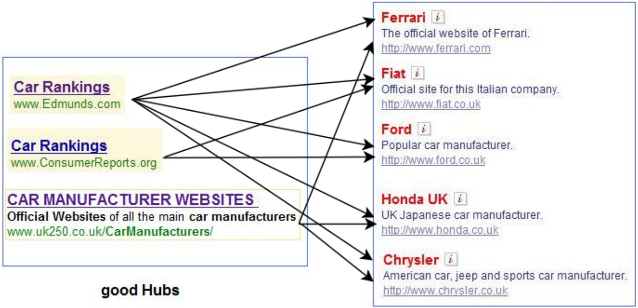
Centrality measures

- 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



Car Rankings
www.Edmunds.com

Car Rankings
www.ConsumerReports.org

CAR MANUFACTURER WEBSITES
Official Websites of all the main **car manufacturers**
www.uk250.co.uk/**CarManufacturers**/

Ferrari
The official website of Ferrari.
http://www.ferrari.com

Fiat
Official site for this Italian company.
http://www.fiat.co.uk

Ford
Popular car manufacturer.
http://www.ford.co.uk

Honda UK
UK Japanese car manufacturer.
http://www.honda.co.uk

Chrysler
American car, jeep and sports car manufacturer.
http://www.chrysler.co.uk

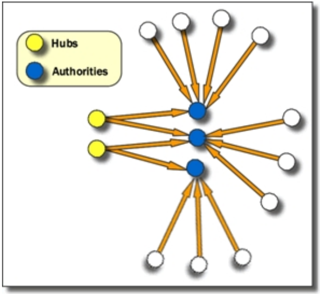
good Hubs

good Authorities

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.

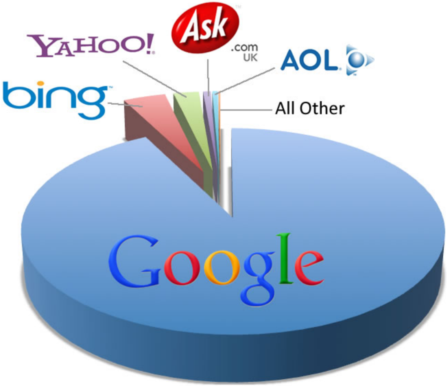
HITS algorithm (easy mode)

- 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

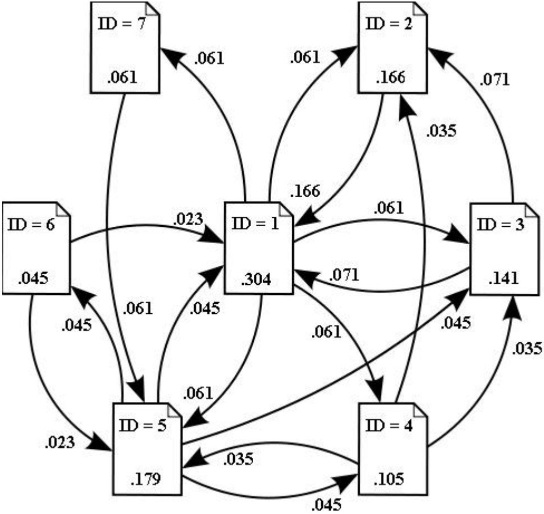
PageRank

- 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?

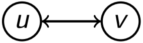
Towards PageRank example



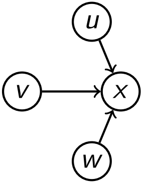
Challenges

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

- **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)

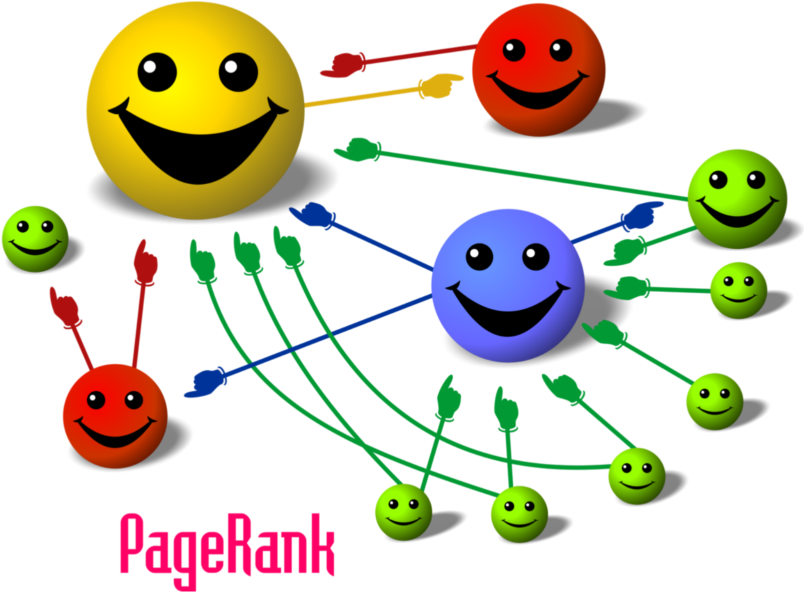
PageRank centrality

- 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 $p = 0.85$

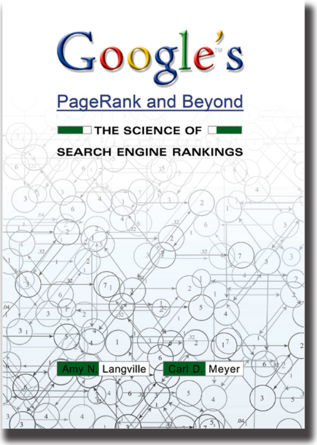
- 100 iterations is usually enough
- Time $100 \cdot m$, similar to HITS.


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

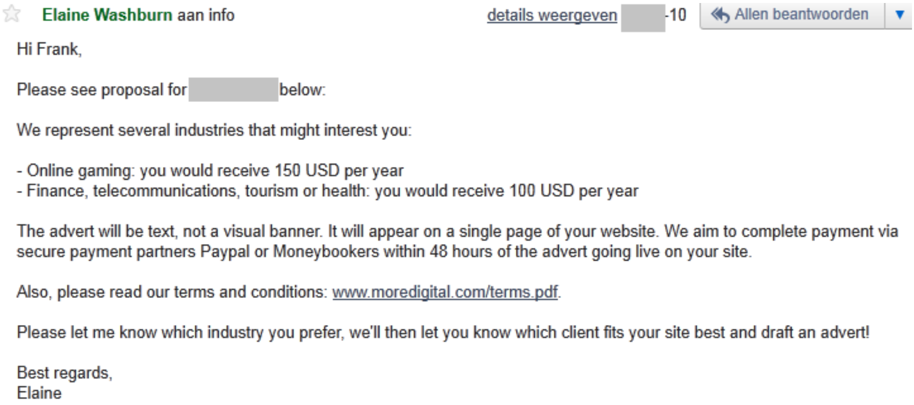
PageRank and beyond


Google's
PageRank and Beyond
THE SCIENCE OF
SEARCH ENGINE RANKINGS
Amy N. Langville
Carl D. Meyer

Actual Google (Page)Rank

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

PageRank hunters



Elaine Washburn aan info

[redacted]-10

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 $O(m)$

Structure of the web

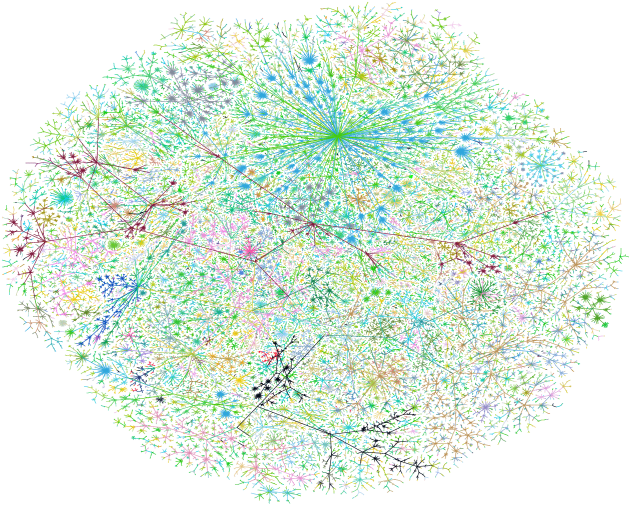
Webgraph

- **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



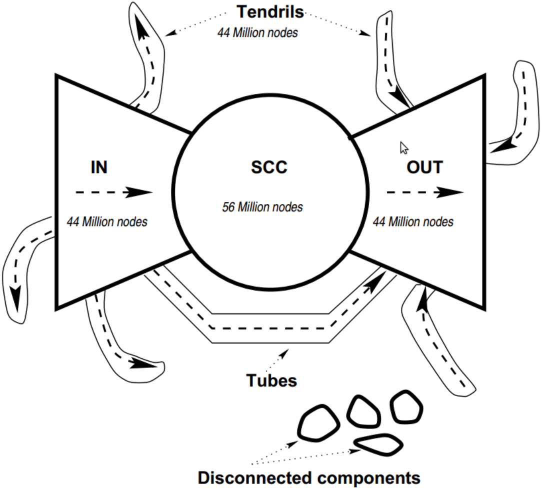
http://www.cheswick.com/ches/map/gallery/index.html

Why study the webgraph?

- 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.

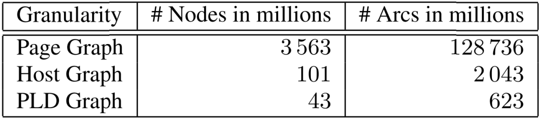
Crawling the webgraph in 2012

- 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)



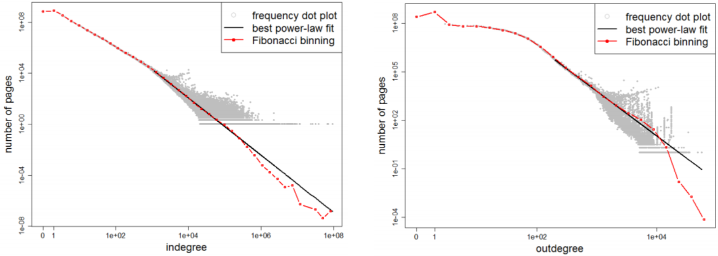
Granularity	# Nodes in millions	# Arcs in millions
Page Graph	3 563	128 736
Host Graph	101	2 043
PLD Graph	43	623

Table 1: Sizes of the graphs

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

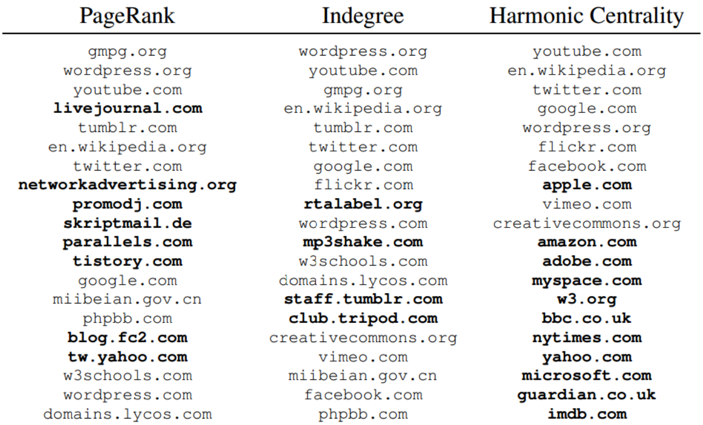
Degree

- 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

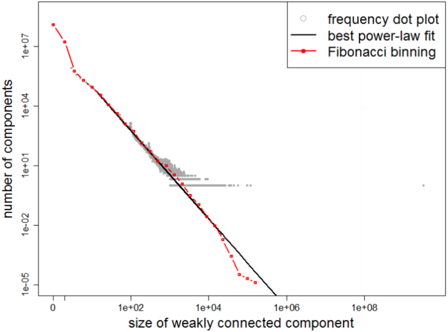


PageRank	Indegree	Harmonic Centrality
gmpg.org	wordpress.org	youtube.com
wordpress.org	youtube.com	en.wikipedia.org
youtube.com	gmpg.org	twitter.com
livejournal.com	en.wikipedia.org	google.com
tumblr.com	tumblr.com	wordpress.org
en.wikipedia.org	twitter.com	flickr.com
twitter.com	google.com	facebook.com
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

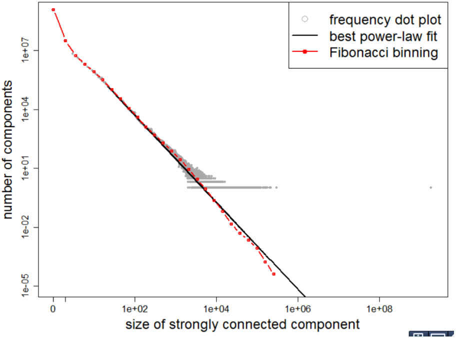
- Weakly Connected Component (WCC)
- 91.8% in 1999, 94% in 2012
- Component size distribution



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

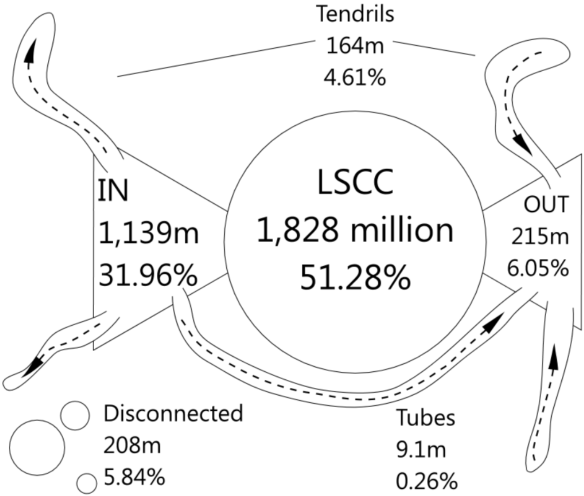
SCC of the webgraph

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



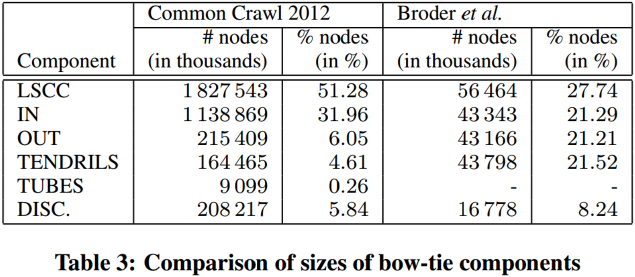
Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Bow-tie structure in 2012



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

Bow-tie structure in 2012



	Common Crawl 2012		Broder *et al.*	
Component	# nodes (in thousands)	% nodes (in %)	# nodes (in thousands)	% 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: $5,282$

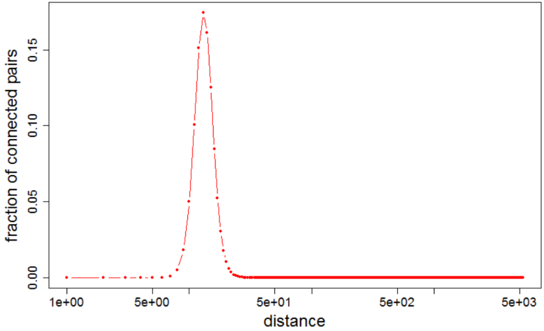


Figure: Distance distribution

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

- 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 . . .

Data Science Lab

`http://rel.liacs.nl/labs/dslab`

“Homework”

- Read your course project paper(s); think about which paper you will present
- Make progress with Assignment 2