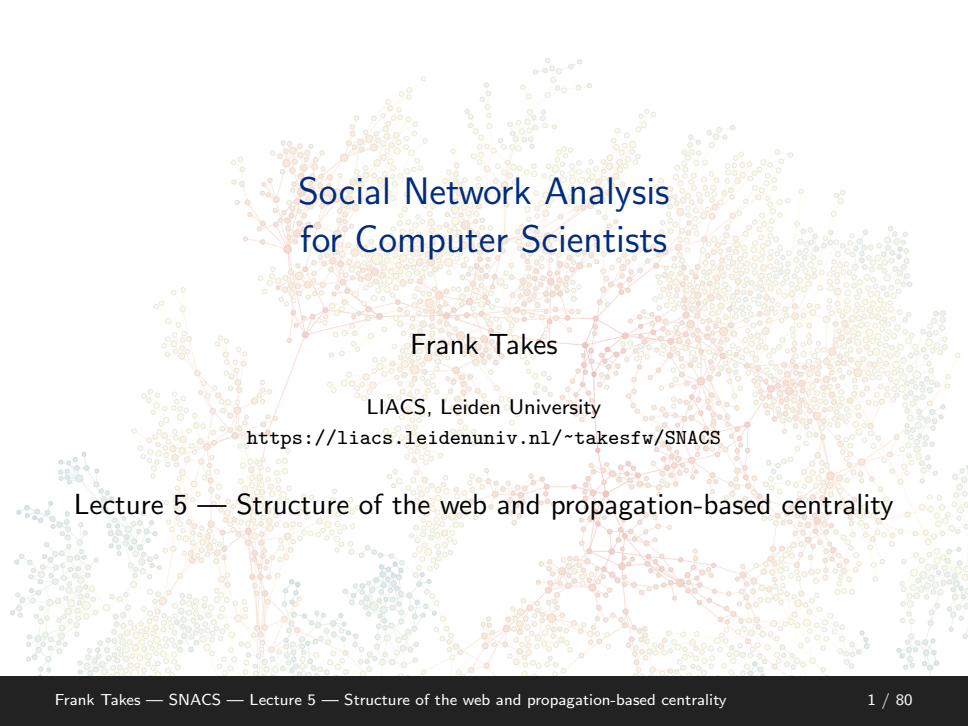




Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

Lecture 5 — Structure of the web and propagation-based centrality

Today

- Recap
- (Random) Walks
- The web
- Propagation-based centrality
- Network structure of the web
- Human network traversal (Wiki Game)

Recap

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
- Many examples: social networks, communication networks, citation networks, collaboration networks, protein interaction networks, information networks, webgraphs, financial networks, ...

Advanced concepts

- Reciprocity
- Complete graph, complement of a graph
- Planar graph, lattice, regular graph
- Tree and dag
- Subgraphs
- Ego network
- Spanning trees
- Diameter
- Bridges
- Power law exponent

Centrality measures

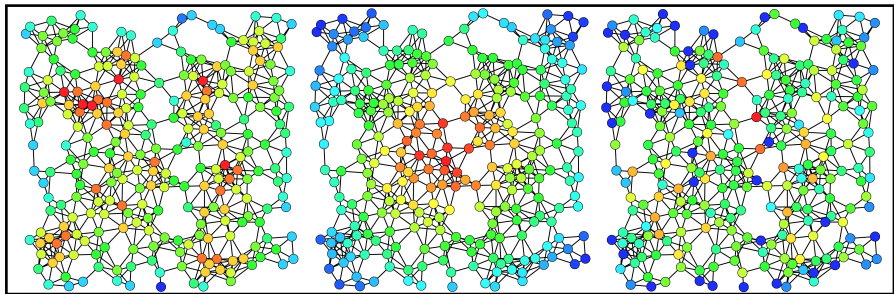


Figure: Degree, closeness and betweenness centrality.

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

Attribute assortativity

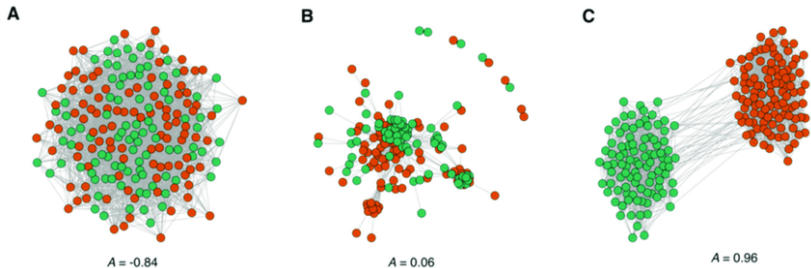


Figure: Attribute assortativity.

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

Projection

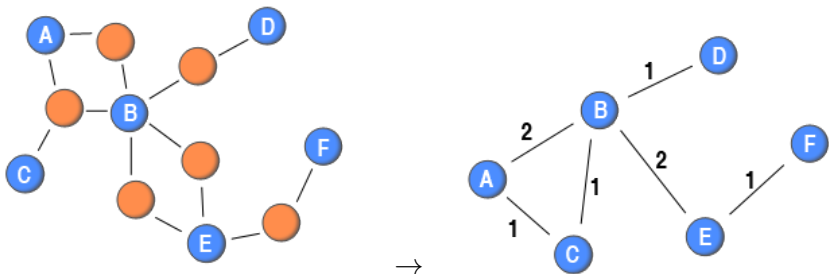


Image: <http://toreopsahl.com>

Community detection

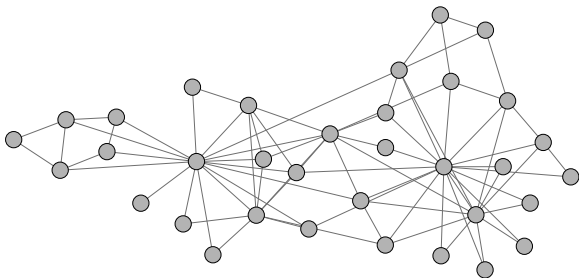


Figure: Communities: node subsets connected more strongly with each other.

- Partitioning a given graph into densely connected groups of nodes

Community detection

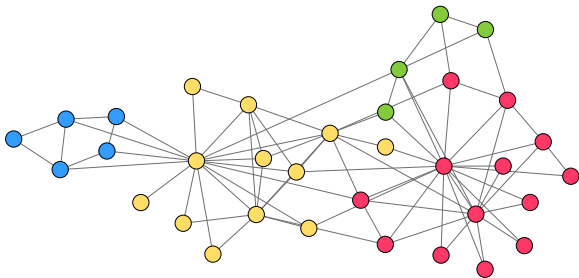


Figure: Communities: node subsets connected more strongly with each other.

- Divisive algorithms (Girvan-Newman), local search algorithms (Louvain) optimizing a quality function (e.g., modularity), label propagation and random walk based approaches.

Core/periphery structure

- Many large networks have a densely connected **core** and a **periphery** of nodes attaching to only a few neighbors
- **Core/periphery structure**: there is **one** dense area
- **Community structure**: there are **several**, with little periphery

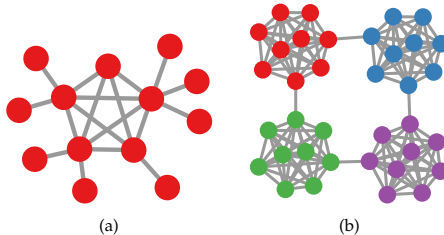


Figure: (a) Core/periphery structure and (b) Community structure.

(Random) Walks

Random walks and the stochastic matrix



Figure: Graph, adjacency matrix and stochastic adjacency matrix.

- **Stochastic adjacency matrix** A with row u valued at $\frac{1}{\deg(u)}$ for every neighbor of u , and 0 otherwise

Random walks and the stochastic matrix

- A **random walk** repeatedly steps from the current node to one of its neighbors, chosen uniformly at random
- Stochastic matrix A^k gives the probability of being at a node after k steps, starting from each node

0	0	0	0	0.33	0	0	0.33	0.33	0.3	0.2	0.08	0	0.07	0.1	0	0.07	0.2	0.08	0.09	0.04	0.05	0.2	0.1	0.07	0.2	0.2
0	0	0	0	0.25	0.25	0.25	0	0.25	0.1	0.3	0.04	0.04	0.05	0.2	0.04	0.09	0.1	0.06	0.09	0.07	0.04	0.1	0.2	0.1	0.1	0.2
0	0	0	0.33	0	0.33	0	0.33	0	0.08	0.06	0.3	0.06	0	0.2	0.06	0.06	0.1	0.04	0.1	0.08	0.1	0.07	0.2	0.06	0.2	0.1
0	0	0.5	0	0	0.5	0	0	0	0	0.08	0.08	0.2	0	0.2	0.08	0.2	0.08	0.08	0.09	0.2	0.06	0.04	0.3	0.05	0.07	0.1
0.33	0.33	0	0	0	0	0	0	0.33	0.07	0.07	0	0	0.3	0.1	0.08	0.2	0.2	0.2	0.2	0.07	0.02	0.08	0.1	0.04	0.09	0.2
0	0.17	0.17	0.17	0	0	0.17	0.17	0.17	0.07	0.1	0.1	0.06	0.07	0.3	0.04	0.09	0.08	0.06	0.1	0.1	0.1	0.07	0.2	0.09	0.1	0.2
0	0.5	0	0	0	0.5	0	0	0	0	0.08	0.08	0.08	0.1	0.1	0.2	0.08	0.2	0.1	0.2	0.08	0.05	0.06	0.3	0.04	0.09	0.1
0.25	0	0.25	0	0	0.25	0	0	0.25	0.05	0.09	0.04	0.1	0.1	0.1	0.04	0.3	0.1	0.1	0.1	0.1	0.04	0.06	0.2	0.05	0.08	0.2
0.2	0.2	0	0	0.2	0.2	0	0.2	0	0.1	0.1	0.08	0.03	0.1	0.1	0.08	0.1	0.3	0.1	0.2	0.06	0.04	0.1	0.2	0.04	0.1	0.1

-0.5em

Figure: Stochastic matrix A (left), A^2 (middle) and A^3 (right).

The stationary distribution

- After a long enough walk it no longer matters where you started; all rows become identical
- **Stationary distribution** π , the probability of finding the walker at each node, with $\pi A = \pi$ (π is the eigenvector of A for eigenvalue 1)
- A high $\pi(v)$ means the walker visits v often (related to how PageRank measures centrality)

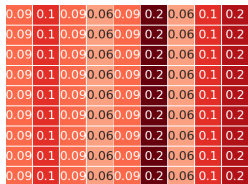


Figure: A^{30} : every row is the same vector π .

But the walker can get stuck

- A walker can never leave the component it starts in, so an unconnected graph has one stationary distribution per component, with zeros everywhere else

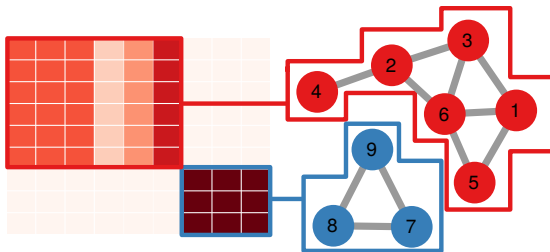


Figure: Two components and the stationary distributions.

Non-backtracking walks

- A walk is **non-backtracking** if it may not re-use the same edge twice in a row
- Useful when going straight back where you came from is not a meaningful step
- The transitions are now between **edges** rather than nodes, so this needs a different matrix

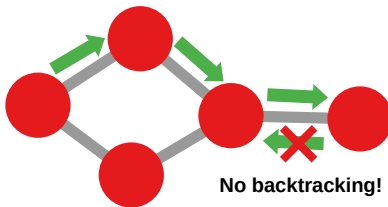


Figure: The green arrows are allowed state transitions, the red one is not.

The web

World wide web

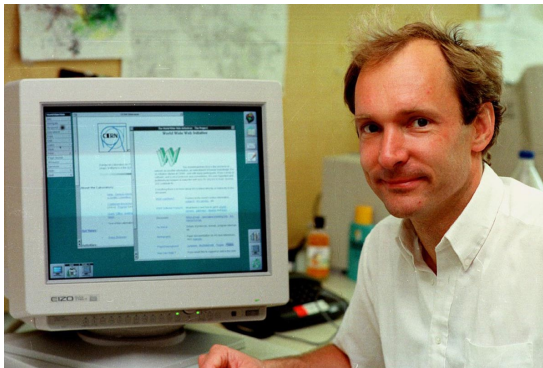


Figure: Tim Berners-Lee.

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?

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:

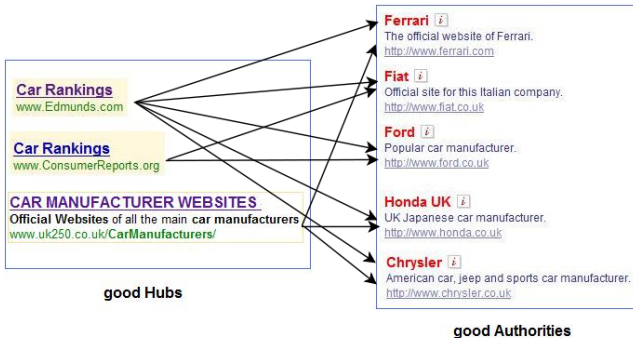
- Hyperlink Induced Topic Search (HITS)

- 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.

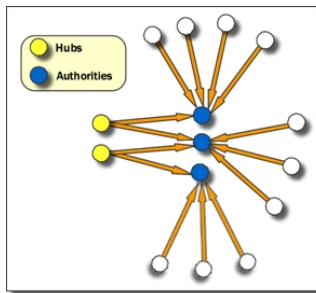
Hyperlink Induced Topic Search



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 $v \in V$:
$$a^t(v) = \sum_{w \in N'(v)} h^{t-1}(w)$$
 - 3 Update the hub scores, so for all nodes $v \in V$:
$$h^t(v) = \sum_{w \in N(v)} a^{t-1}(w)$$
 - 4 Normalize both scores so that $\sum_{v \in V} (a^t(v))^2 = \sum_{v \in V} (h^t(v))^2 = 1$
- Until scores converge:
$$\sum_{v \in V} (a^t(v) - a^{t-1}(v))^2 < \epsilon \text{ and}$$
$$\sum_{v \in V} (h^t(v) - h^{t-1}(v))^2 < \epsilon$$

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 the sum of their squares is 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

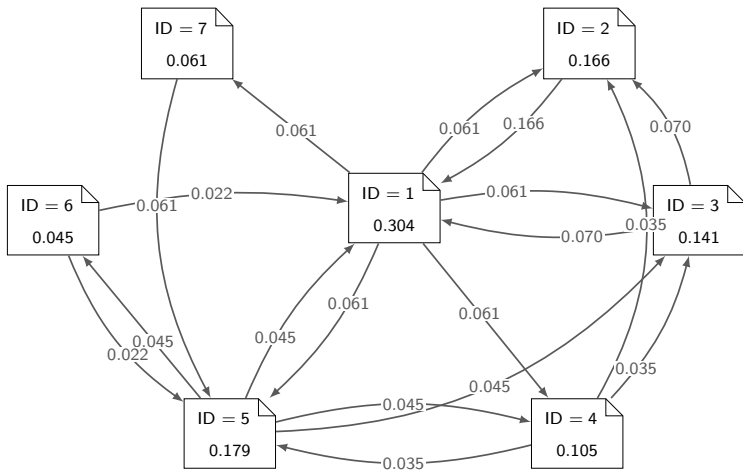
PageRank

- 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 $\text{outdeg}(v)$ is the outdegree of node $v \in V$
- Each page has its own importance $PR(v)$
- Each page v casts equal votes of size $\frac{PR(v)}{\text{outdeg}(v)}$ 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: $PR(v) = \sum_{w \in N'(v)} \frac{PR(w)}{\text{outdeg}(w)}$
- Again recursive
- Does it converge?

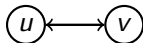
Towards PageRank example



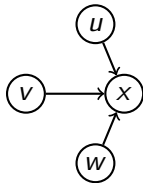
Challenges

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

- **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 $PR^t(v) = \frac{1-p}{n} + p \cdot \sum_{w \in N'(v)} \frac{PR^{t-1}(w)}{\text{outdeg}(w)}$
 - 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:

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

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 $p = 0.85$

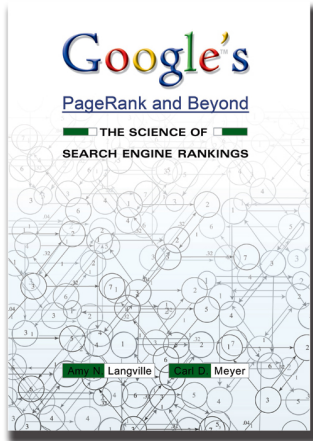
- 100 iterations is usually enough
- Time $100 \cdot m$, similar to HITS.



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



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

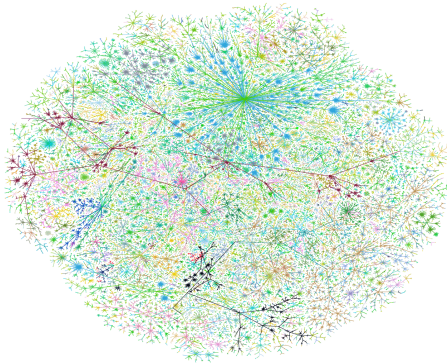
- **Webgraph**: directed unweighted network $G = (V, E)$, in which
 - Nodes V are webpages
 - Links E are “hyperlinks” to other pages
- Many dense subgraphs ...

Webgraph

- **Webgraph**: directed unweighted network $G = (V, E)$, in which
 - 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

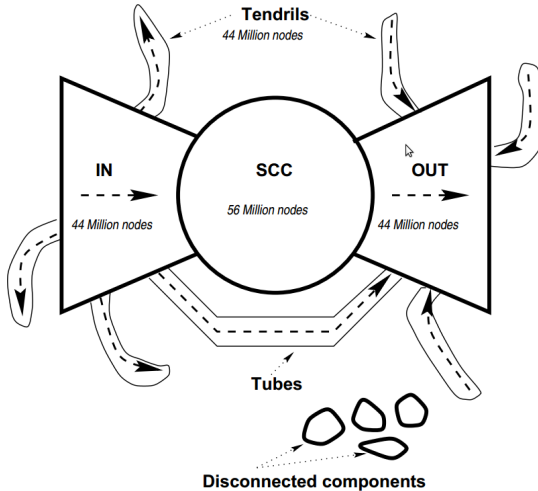
Why study the webgraph?

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



[/www.cheswick.com/ches/map/gallery](http://www.cheswick.com/ches/map/gallery) and 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 nodes)
- 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)

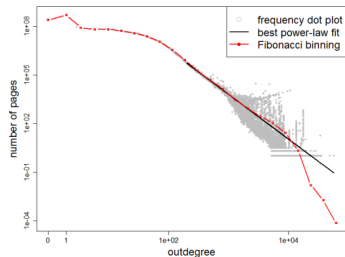
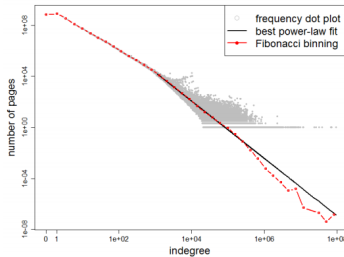
Granularity	# Nodes in millions	# Arcs in millions
Page Graph	3 563	128 736
Host Graph	101	2 043
PLD Graph	43	623

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.

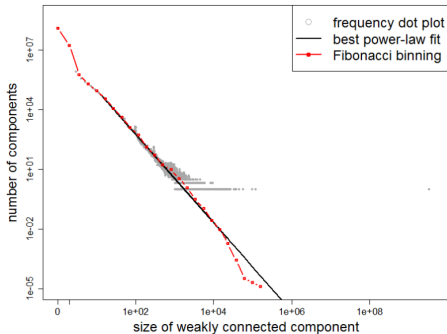
Centrality in the webgraph

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

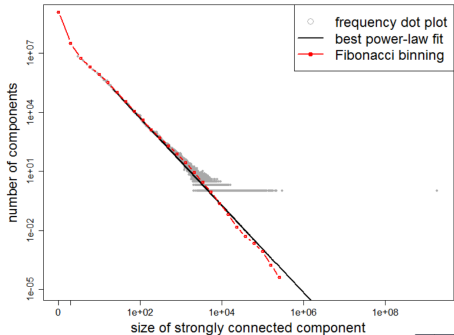
- Weakly Connected Component (WCC)
- 91.8% in 1999, 94% in 2012
- Component size distribution



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

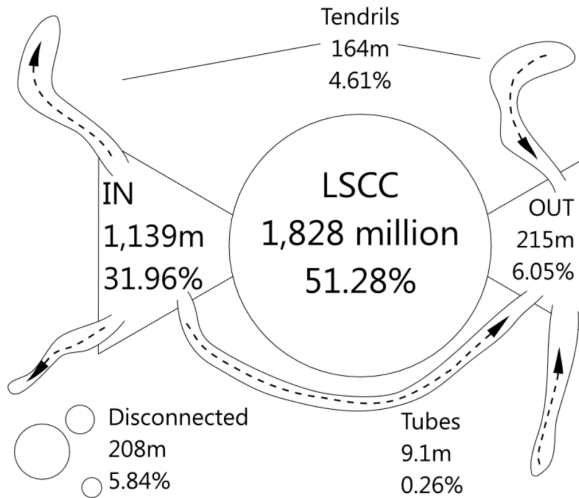
SCC of the webgraph

- 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		Broder <i>et al.</i>	
	# 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.

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.

Walks in information networks

F.W. Takes and W.A. Kusters, 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. Kusters, 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

Like Send 11,355 people like this. Be the first of your friends. [Login](#) or [Signup now for Username & Stats!](#)

EXPLORE & RACE THROUGH WIKIPEDIA ARTICLES

THE WIKI GAME

CURRENT GAME
START
BLUE DRAGON
GOAL
NELLY

 **PLAY NOW!**
With other cool brainiacs!

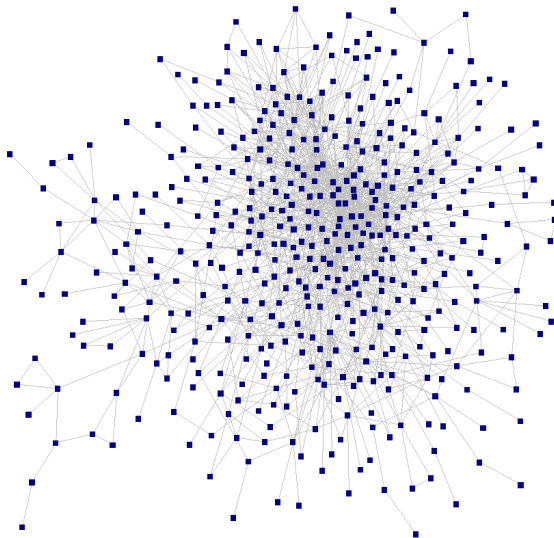
AT LAST!
PLAY ANYWHERE WITH THE
IPHONE APP!
(It's transcendently exquisite)

Available on the App Store

WIKI GAME

SPEED RACE
Question to Answer
13/20 32/60

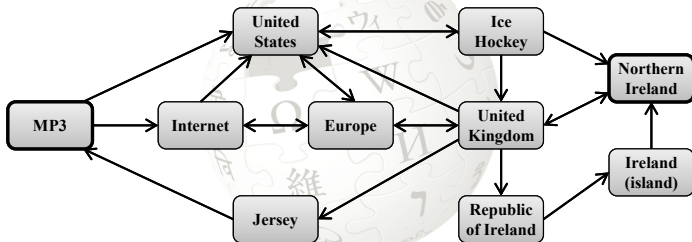
Information Networks



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 ($\bar{d}$)	4.81
Effective diameter	7
Diameter	11

Table: Wikipedia dataset.

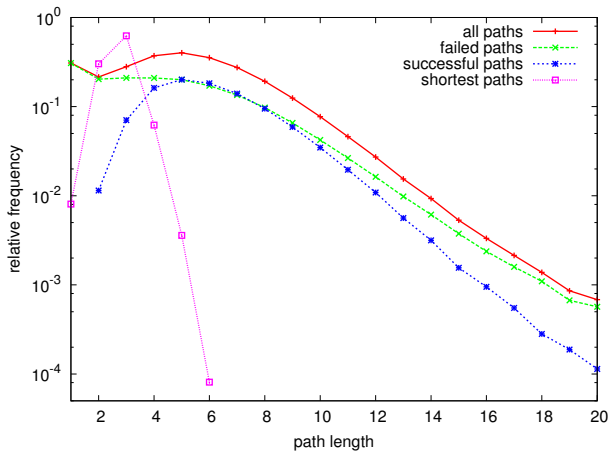
The Wiki Game dataset

- 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 successful 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:

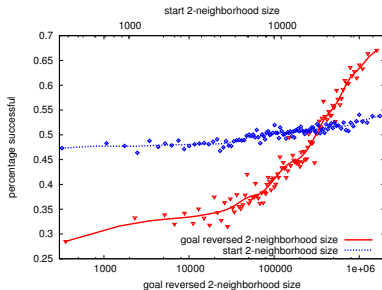
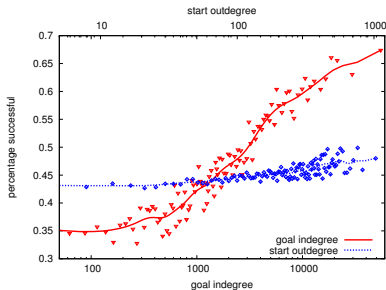
$$f_{|N'_2|}(u, v) = |N'_2(v)|$$

- Path length (distance):

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

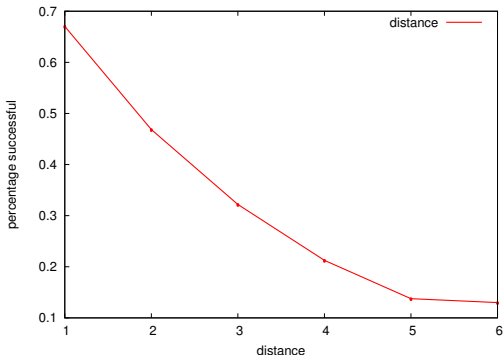
Local difficulty measures

- 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

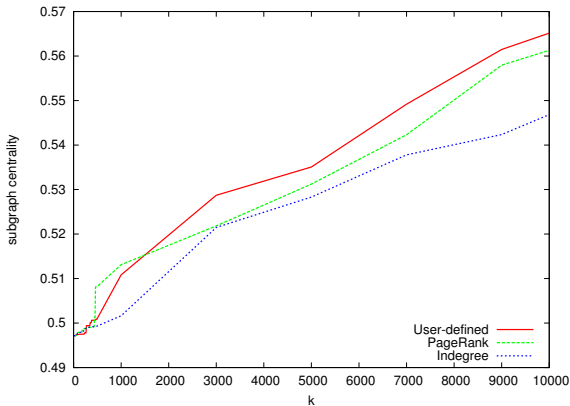


Figure: Comparison of subgraph centrality (vertical axis) of various centrality measures for different values of k (horizontal axis).

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
- 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

Upcoming lab session and “Homework”

- Finalize Assignment 1

Credits

- Several images are gratefully reused from Michele Coscia, The Atlas for the Aspiring Network Scientist, v3 of <https://arxiv.org/abs/2101.00863>