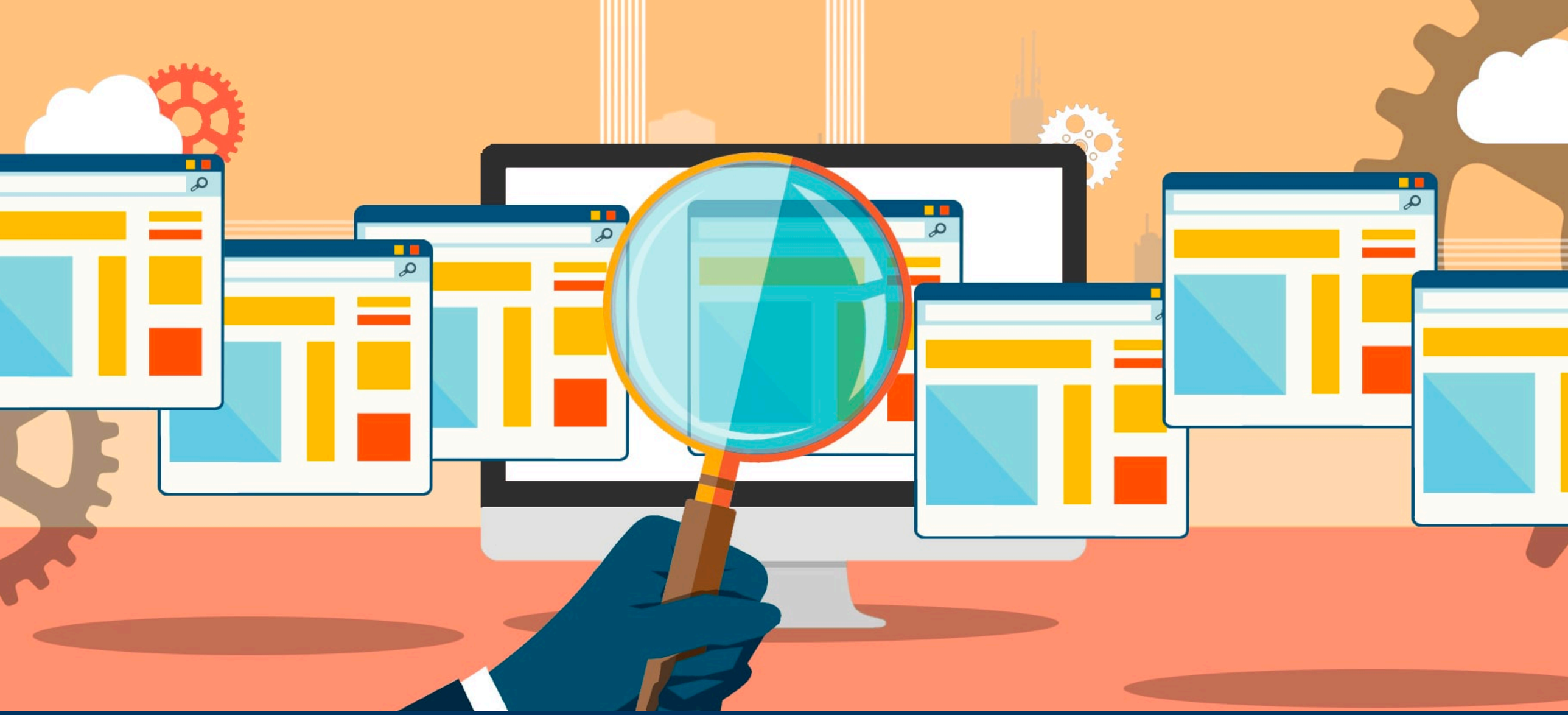


Extracting and Analyzing Graphs

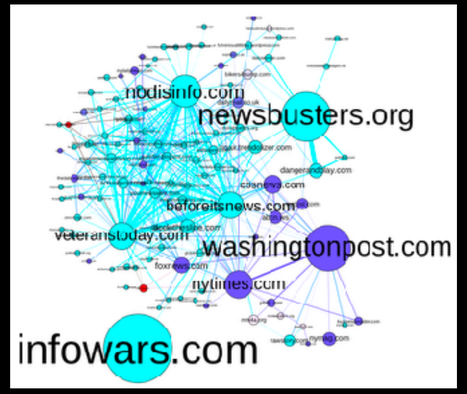
INST414 - Data Science Techniques

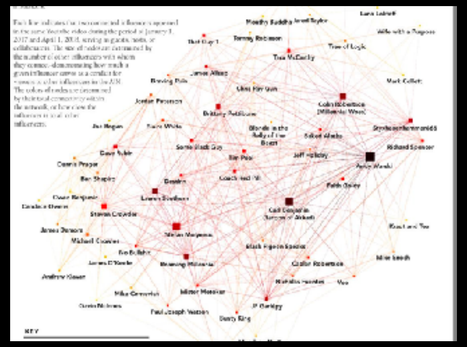
Six Core Learning Objectives

1. Collect and clean large-scale datasets
2. Articulate the math behind supervised and unsupervised techniques
3. Execute supervised and unsupervised machine learning techniques
4. Select and evaluate various types of machine learning techniques
5. Explain the results coming out of the models
6. Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

Where are we?

1. Collect and clean large-scale datasets

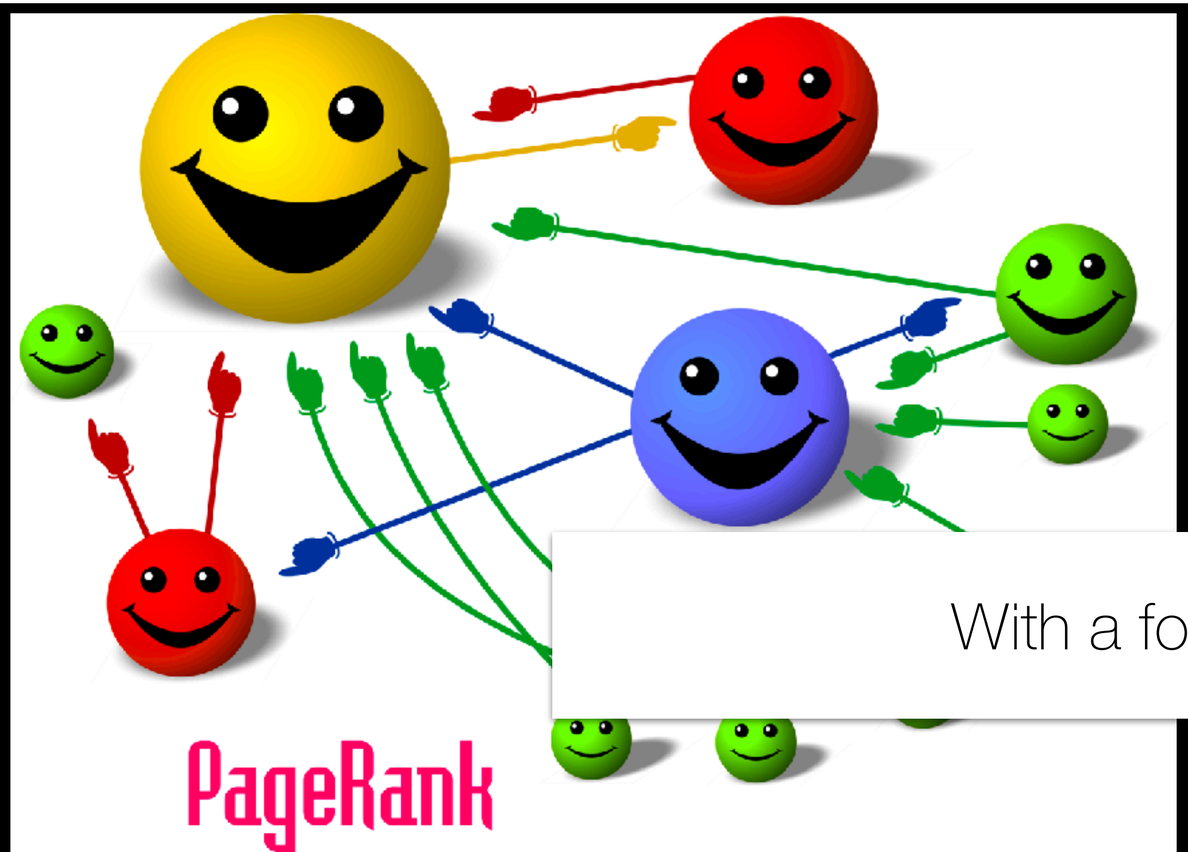




With a focus on network structures

Where are we?

2. Articulate the math behind supervised and unsupervised techniques



With a focus on graph analysis

This Lecture's Learning Objectives

- Construct graphs from web structures
- Describe at least two ways how one may try to exploit search engines
- Explain how PageRank captures "importance" in a graph structure
- Construct a transition matrix to model a collection of web pages
- Differentiate hubs and authorities

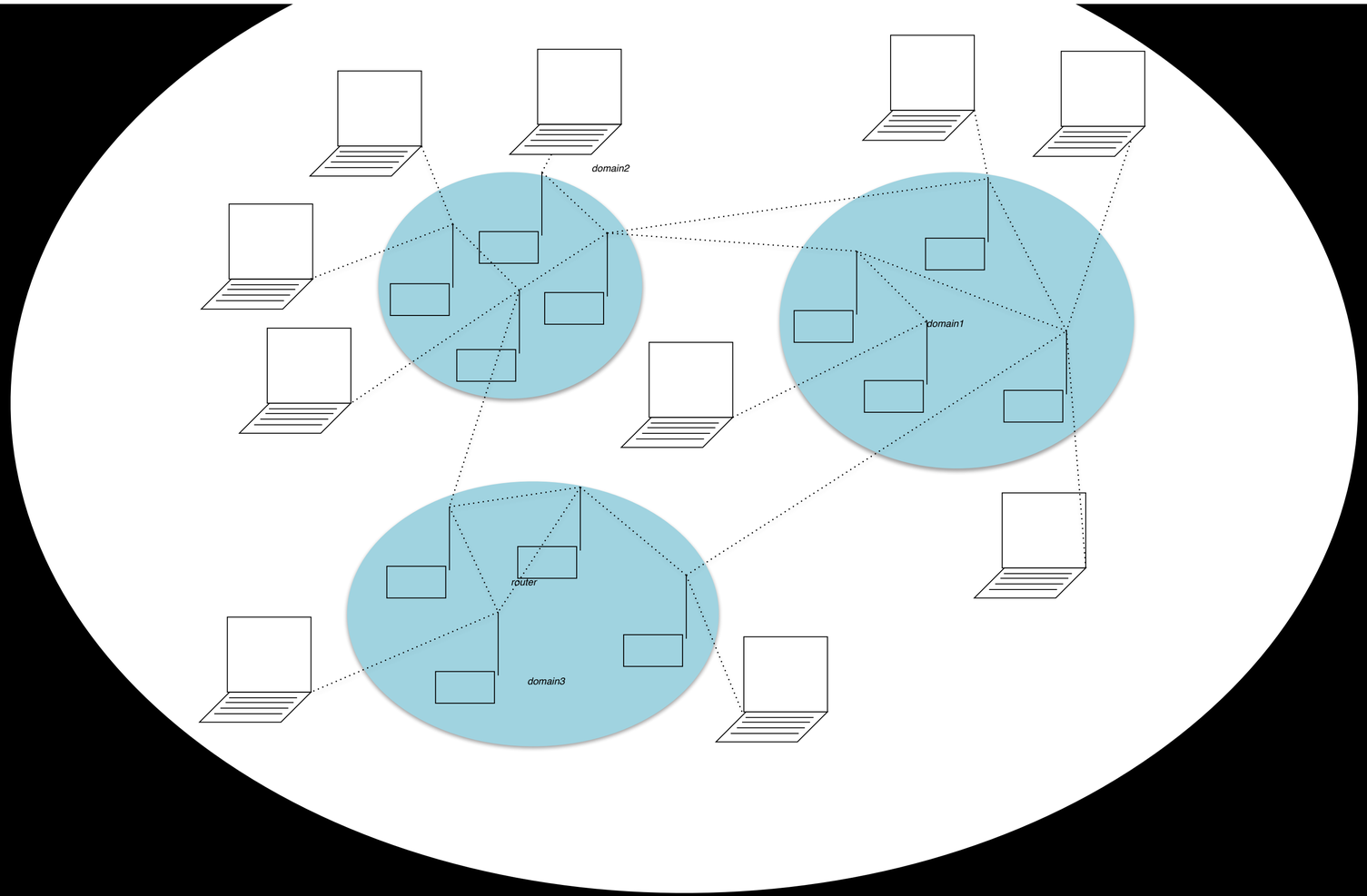
This Lecture's Learning Objectives

- Construct graphs from web structures
- Describe at least two ways how one may try to exploit search engines
- Explain how PageRank captures "importance" in a graph structure
- Construct a transition matrix to model a collection of web pages
- Differentiate hubs and authorities

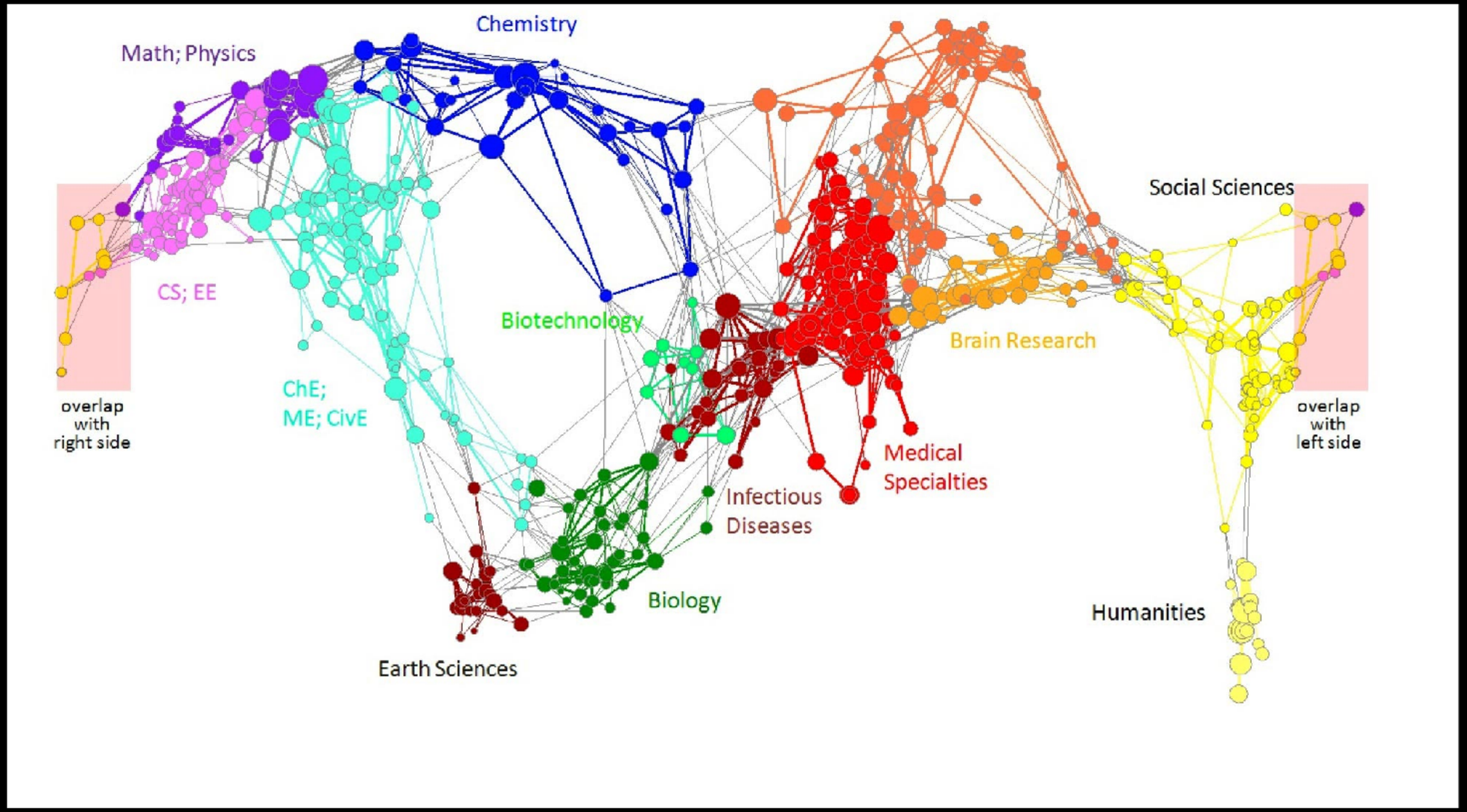
Web data is full of graphs/networks

What are some examples?

Graphs in the Internet

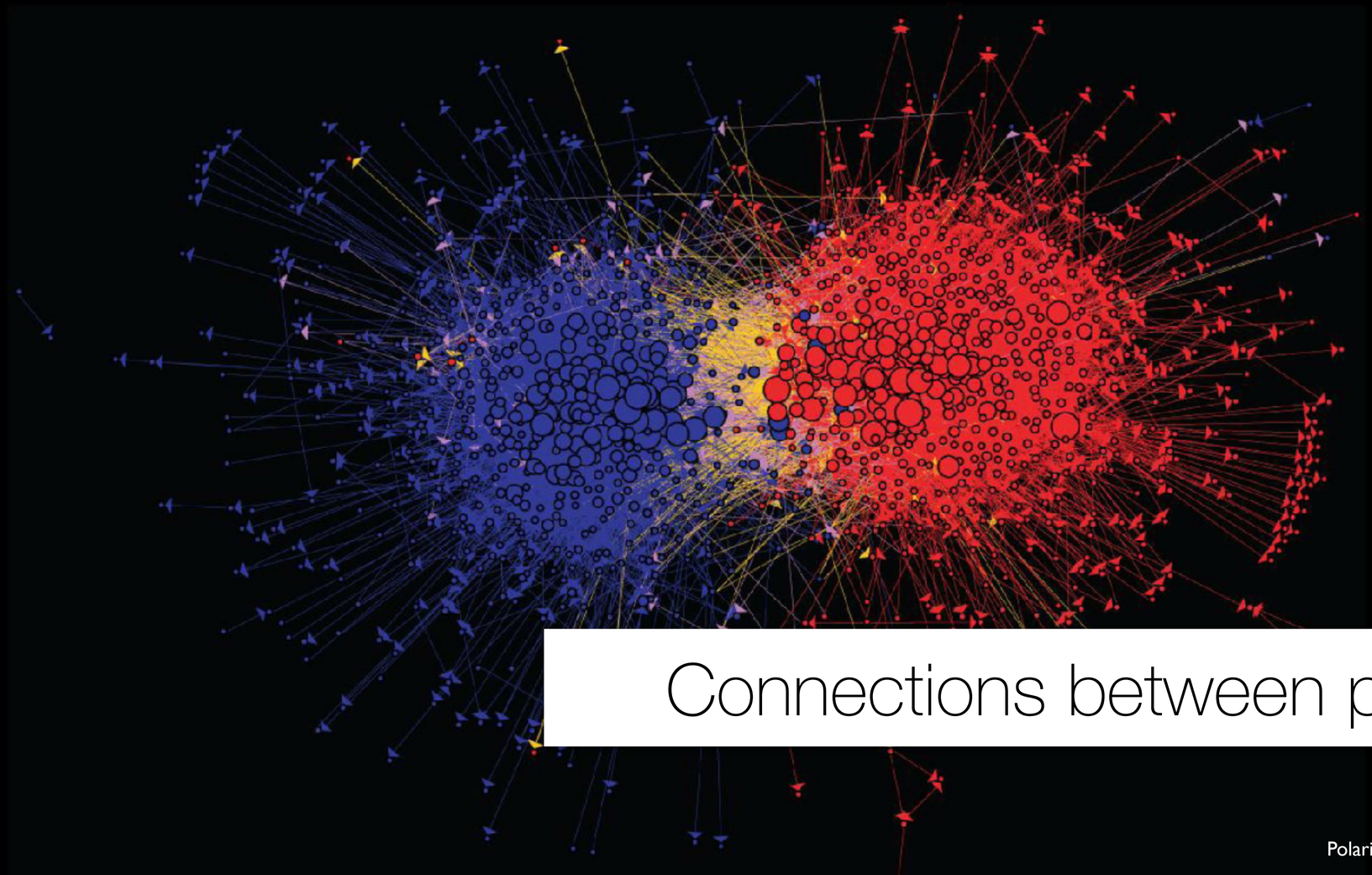


Graphs in Citations Networks



[Börner et al., 2012]

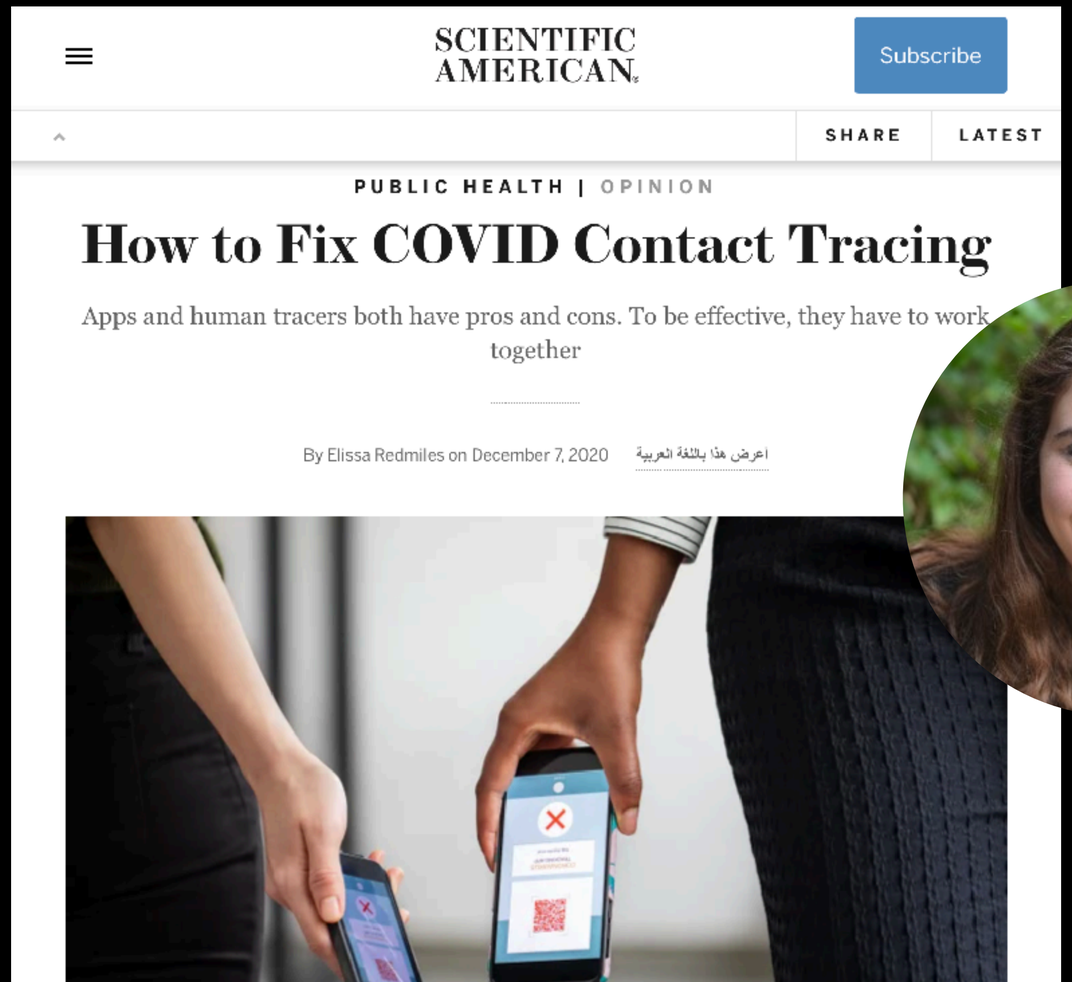
Graphs in Media Networks



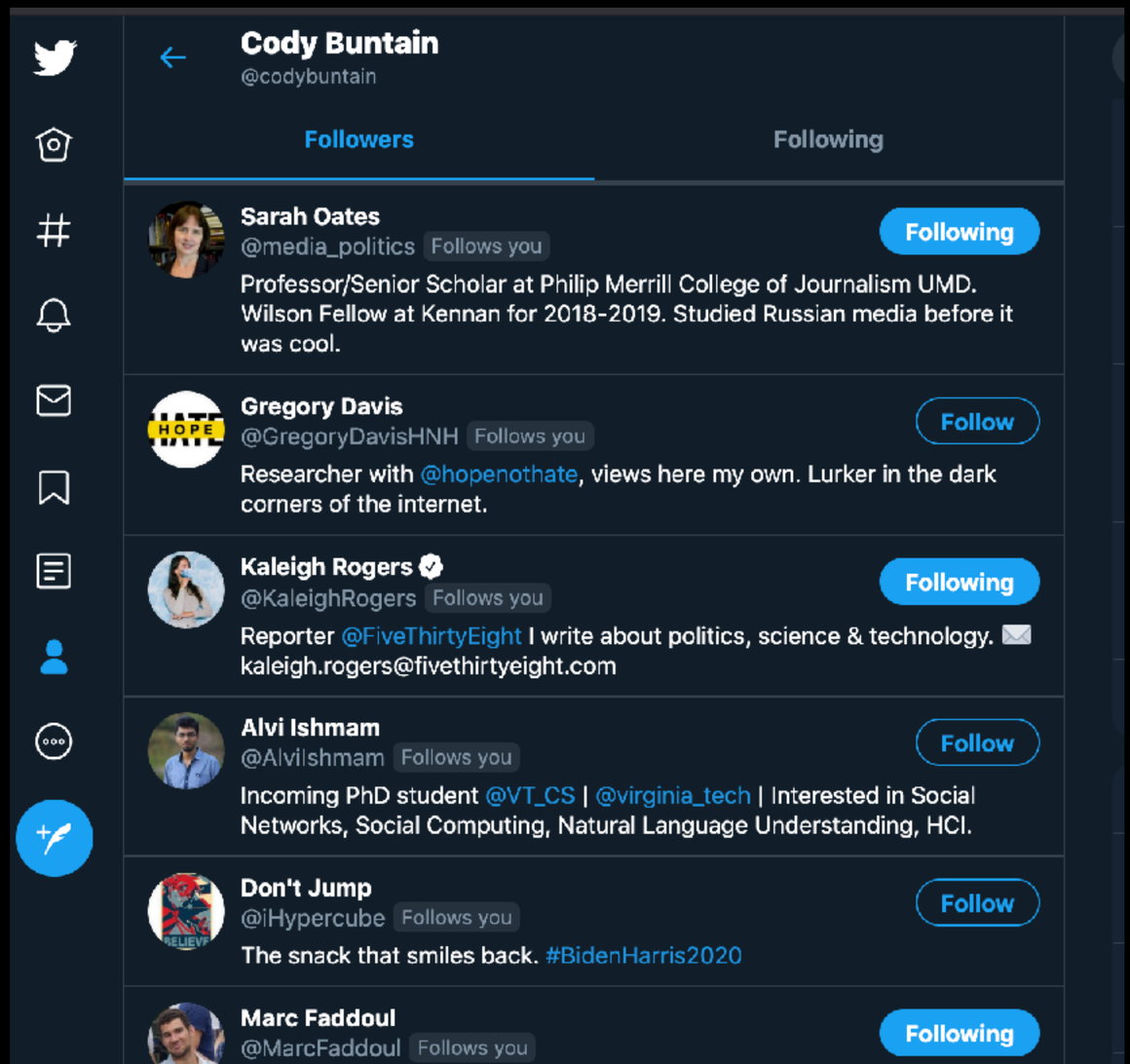
Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

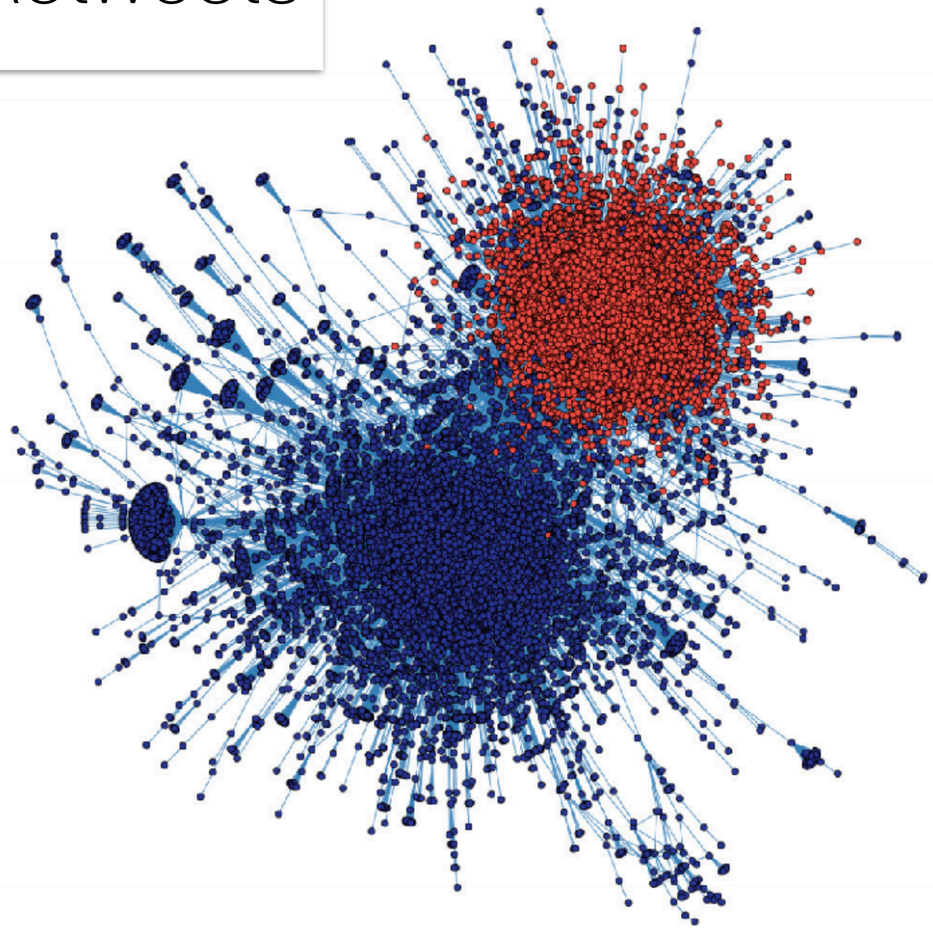
Graphs in Author Networks



Graphs in Social Media



Retweets



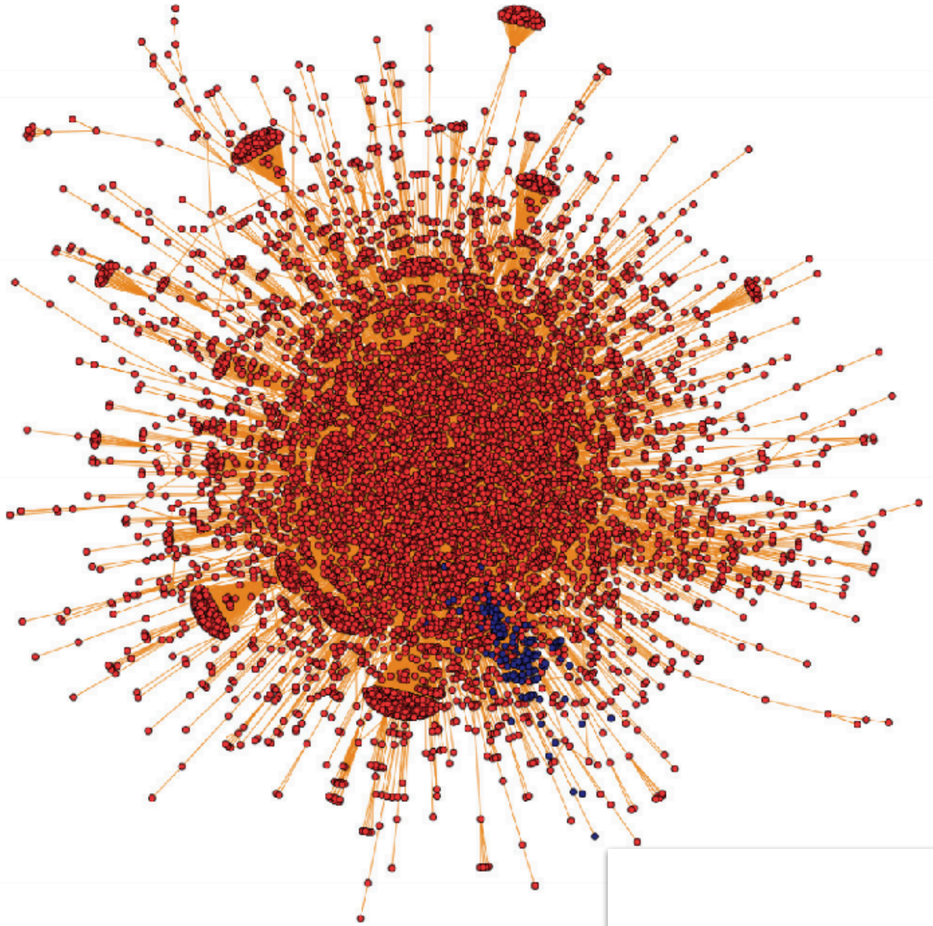
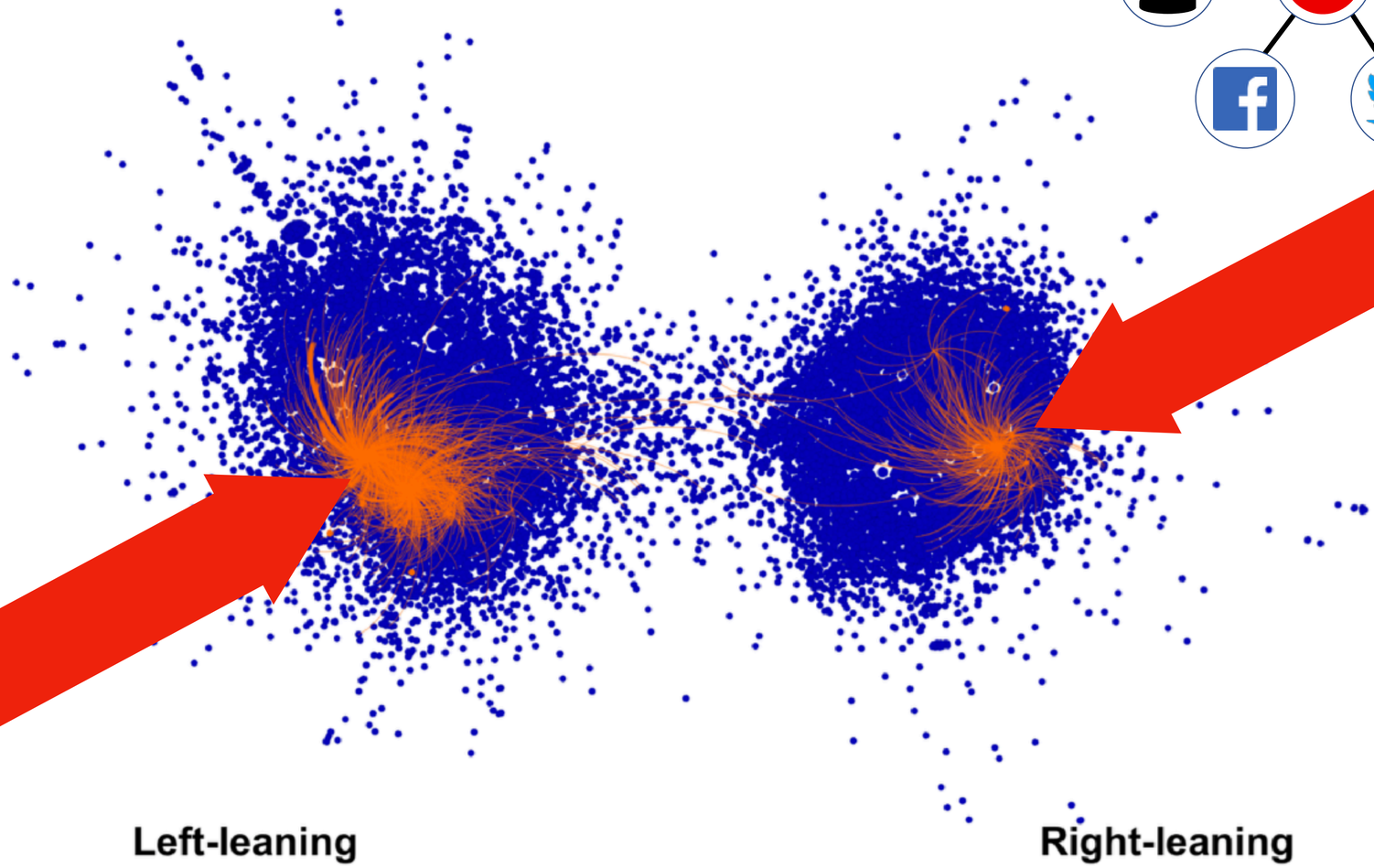


Figure 1: The political retweet (left) and mention (right) networks, laid out using a force-directed algori
cluster assignments (see § 3.1). Community structure is evident in the retweet network, but less so in th
show in § 3.3 that in the retweet network, the red cluster A is made of 93% right-leaning users, while the blue cluster B is made
of 80% left-leaning users.

Mentions

Conover, Michael D., Jacob Ratkiewicz, Matthew R. Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. "Political polarization on twitter." Icwsm 133, no. 26 (2011): 89-96.

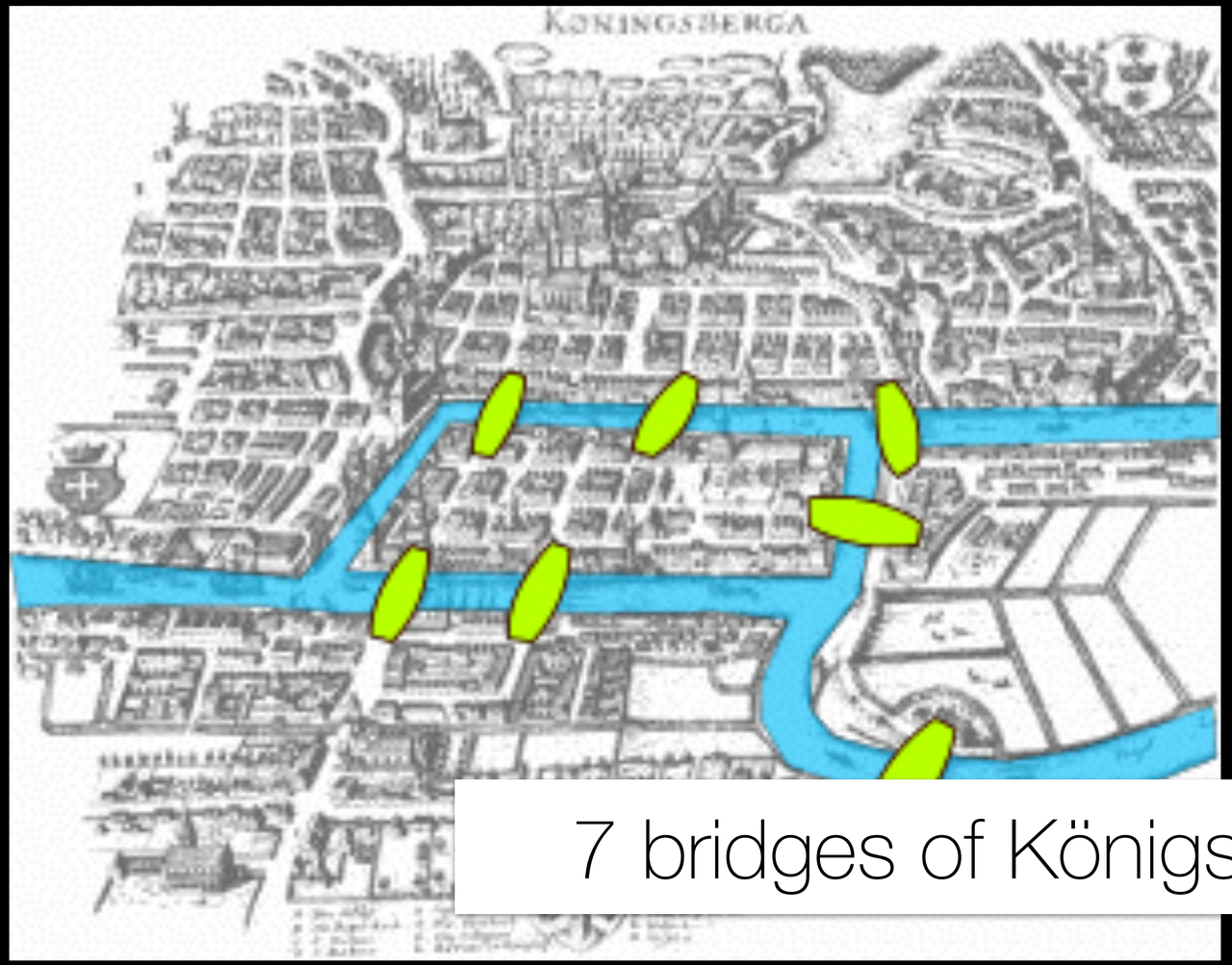
with the IRA accounts in orange and other accounts in blue.



Retweet Network Graph plus IRA Troll Accounts. Courtesy of Kate Starbird/University of Washington

https://medium.com/s/story/the-trolls-within-how-russian-information-operations-infiltrated-online-communities-691fb969b9e4

Graphs aren't new



7 bridges of Königsberg [Euler 1735]

Other examples of graphs in the Web?

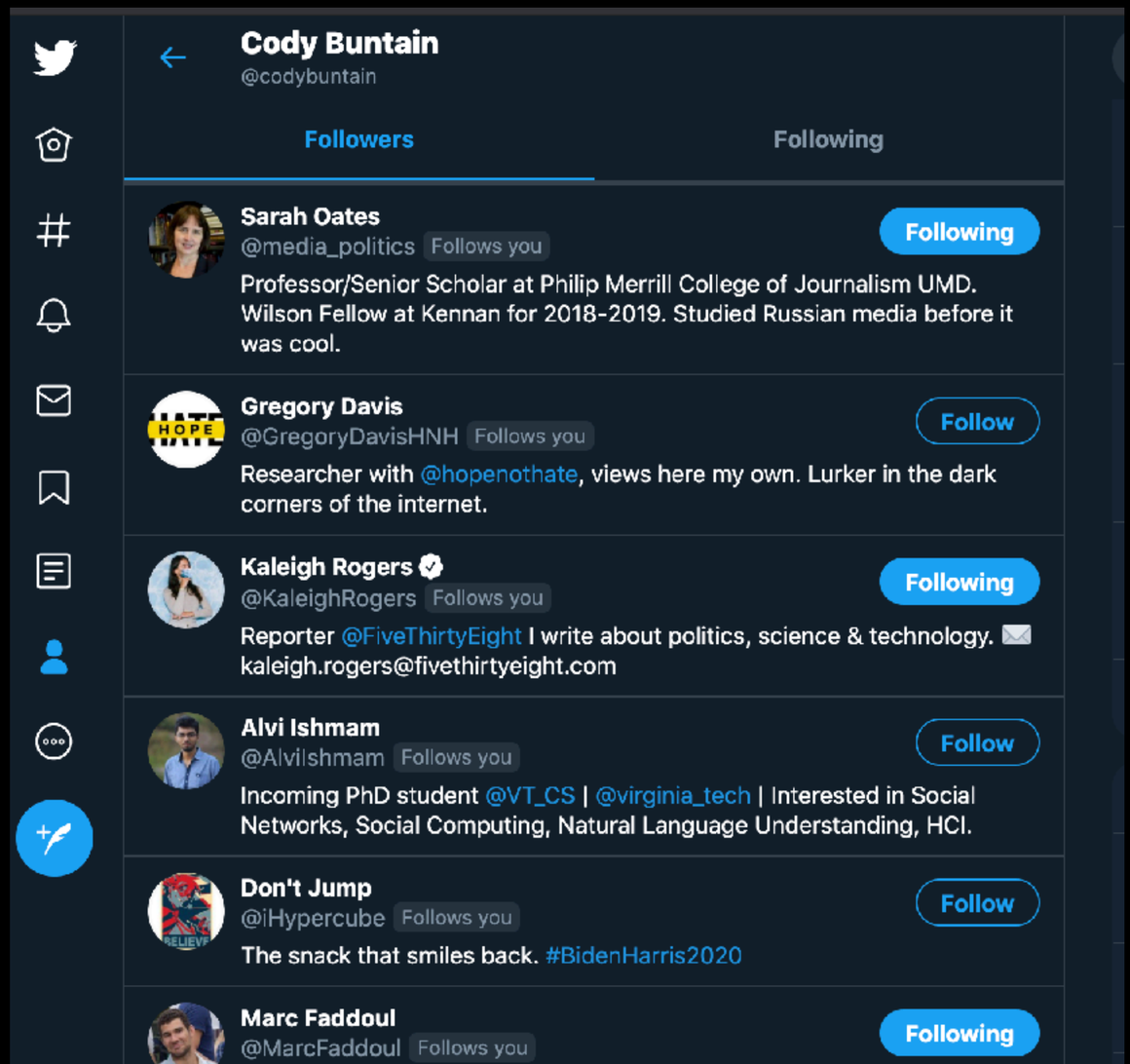
Several kinds of graphs

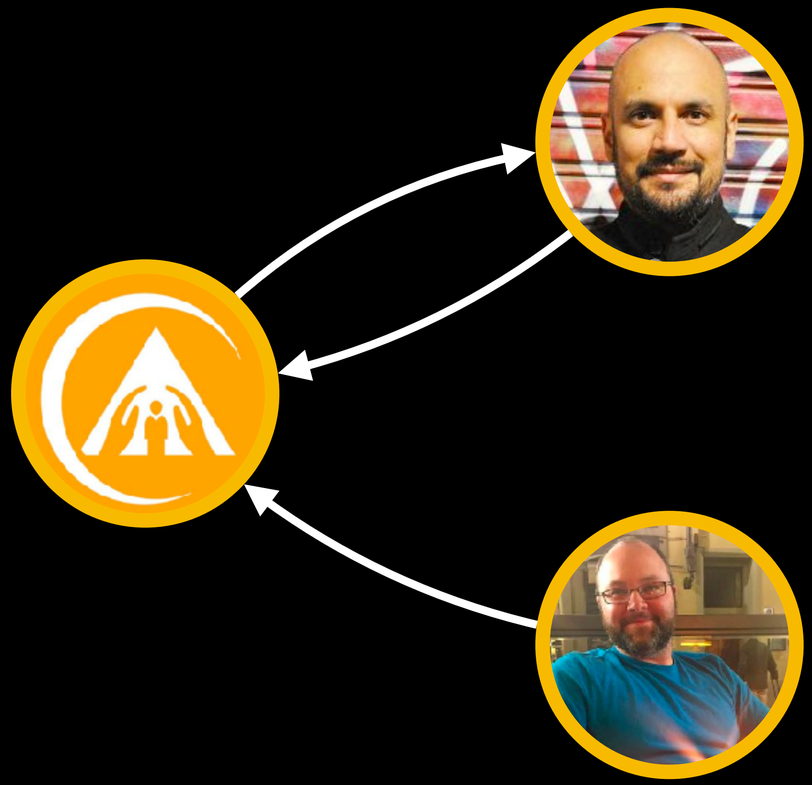
Directed graphs

Undirected graphs

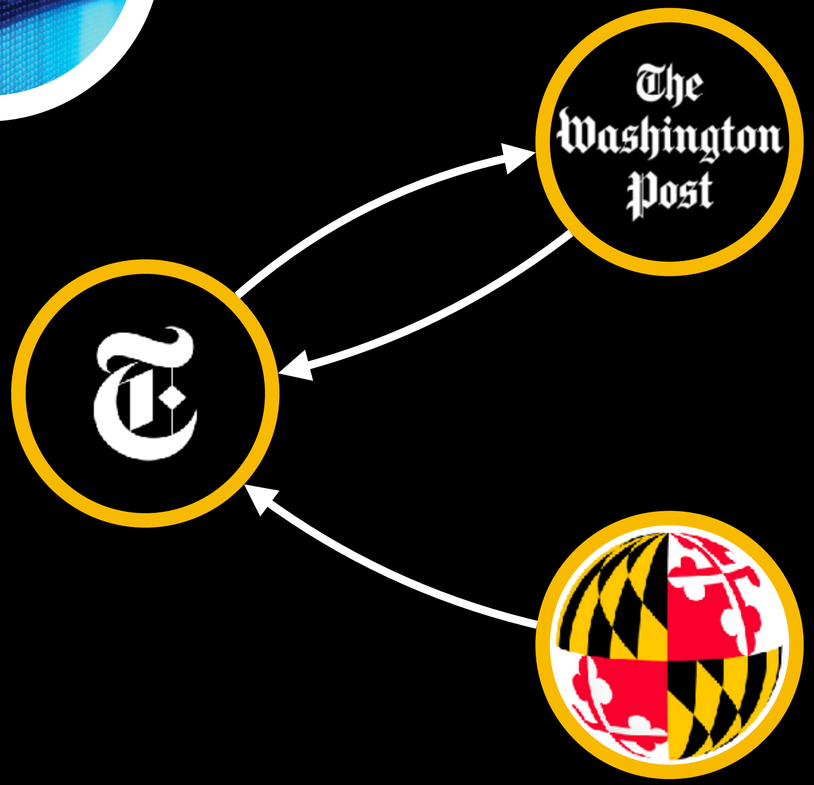
And others (bipartite, multigraphs, hypergraphs, etc.)

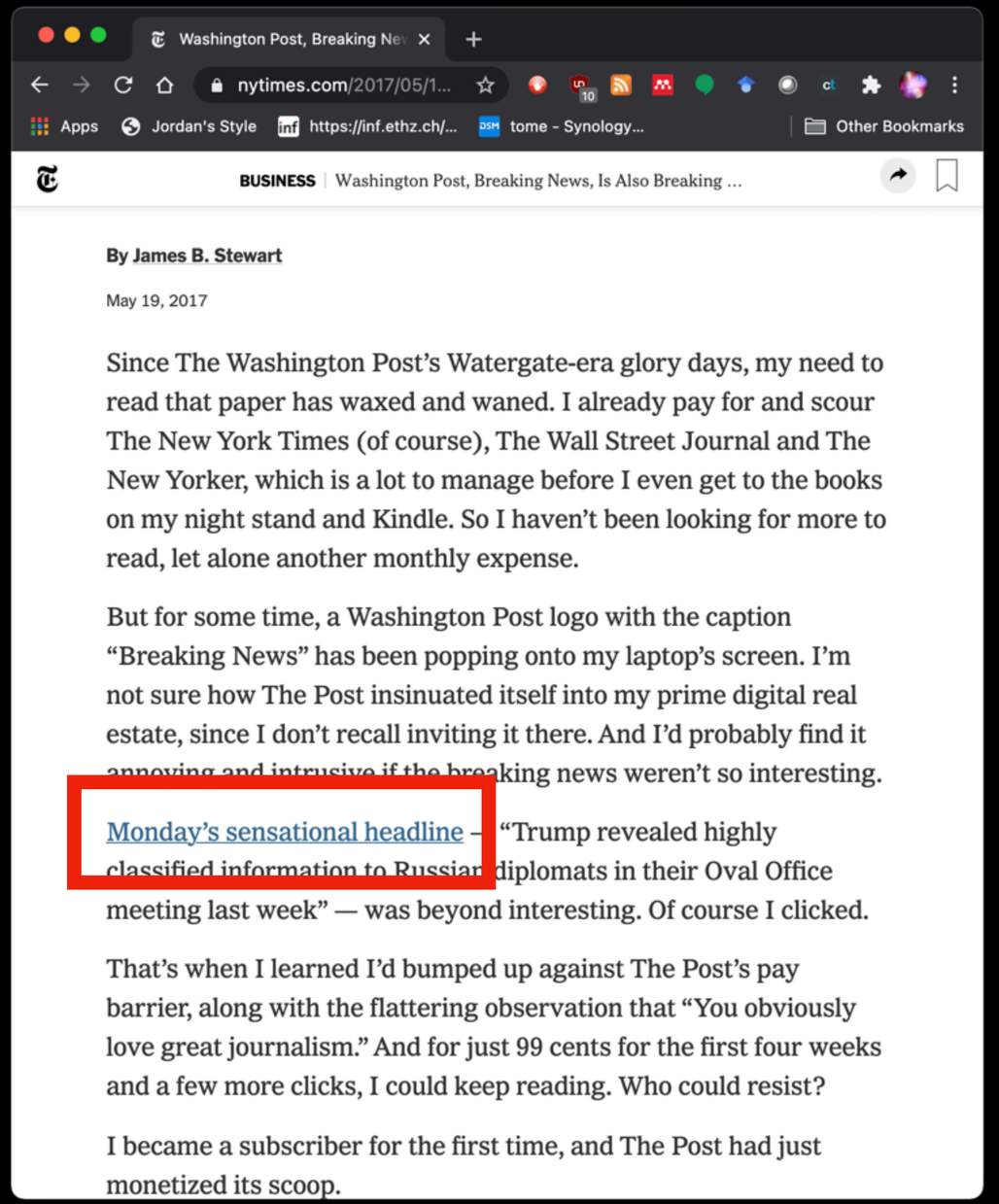
Graphs in Social Media





Directed Graph





Washington Post, Breaking Ne

nytimes.com/2017/05/1...

Apps | Jordan's Style | https://inf.ethz.ch/... | tome - Synology... | Other Bookmarks

BUSINESS | Washington Post, Breaking News, Is Also Breaking ...

By James B. Stewart

May 19, 2017

Since The Washington Post's Watergate-era glory days, my need to read that paper has waxed and waned. I already pay for and scour The New York Times (of course), The Wall Street Journal and The New Yorker, which is a lot to manage before I even get to the books on my night stand and Kindle. So I haven't been looking for more to read, let alone another monthly expense.

But for some time, a Washington Post logo with the caption "Breaking News" has been popping onto my laptop's screen. I'm not sure how The Post insinuated itself into my prime digital real estate, since I don't recall inviting it there. And I'd probably find it annoying and intrusive if the breaking news weren't so interesting.

Monday's sensational headline – "Trump revealed highly classified information to Russian diplomats in their Oval Office meeting last week" — was beyond interesting. Of course I clicked.

That's when I learned I'd bumped up against The Post's pay barrier, along with the flattering observation that "You obviously love great journalism." And for just 99 cents for the first four weeks and a few more clicks, I could keep reading. Who could resist?

I became a subscriber for the first time, and The Post had just monetized its scoop.

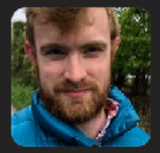
Friends

Search

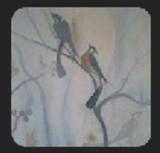
Friend Requests

Find Friends

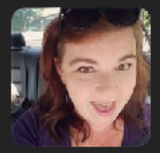
All Friends | Birthdays | Work | College | High School | Hometown | Following



Kristopher Micinski
45 mutual friends
Friends



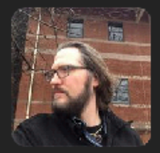
Chris Harrington
15 mutual friends
Friends



Leigh Cook Buntain
272 mutual friends
Friends

Chris Kilroy
73 mutual friends
Friends

Leslie Harrington
36 mutual friends
Friends



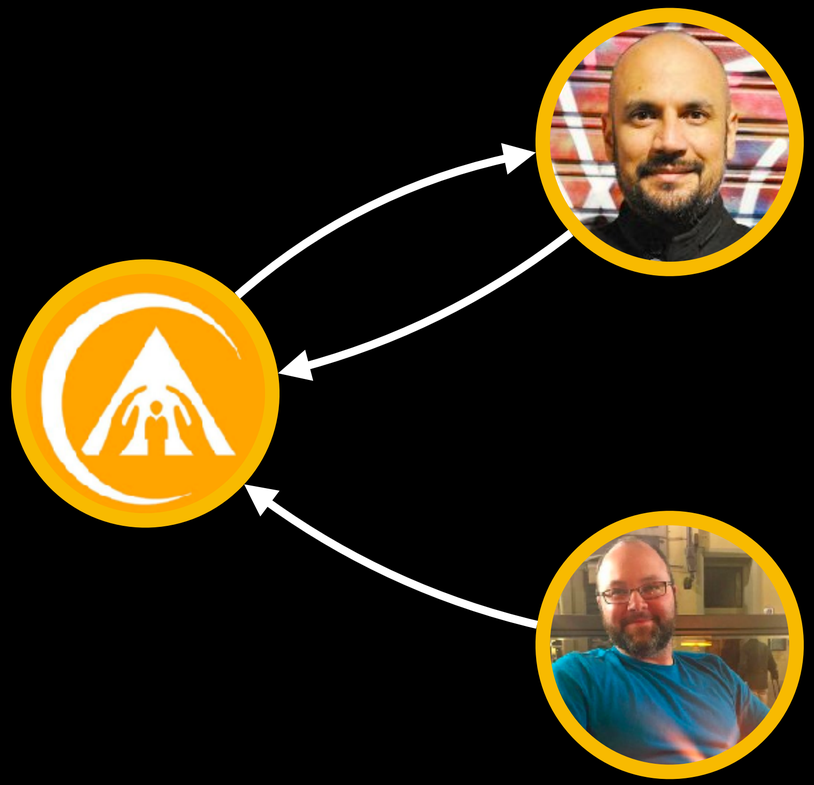
Greg Bennett
155 mutual friends
Friends

Haley Elizabeth Allen
42 mutual friends
Friends

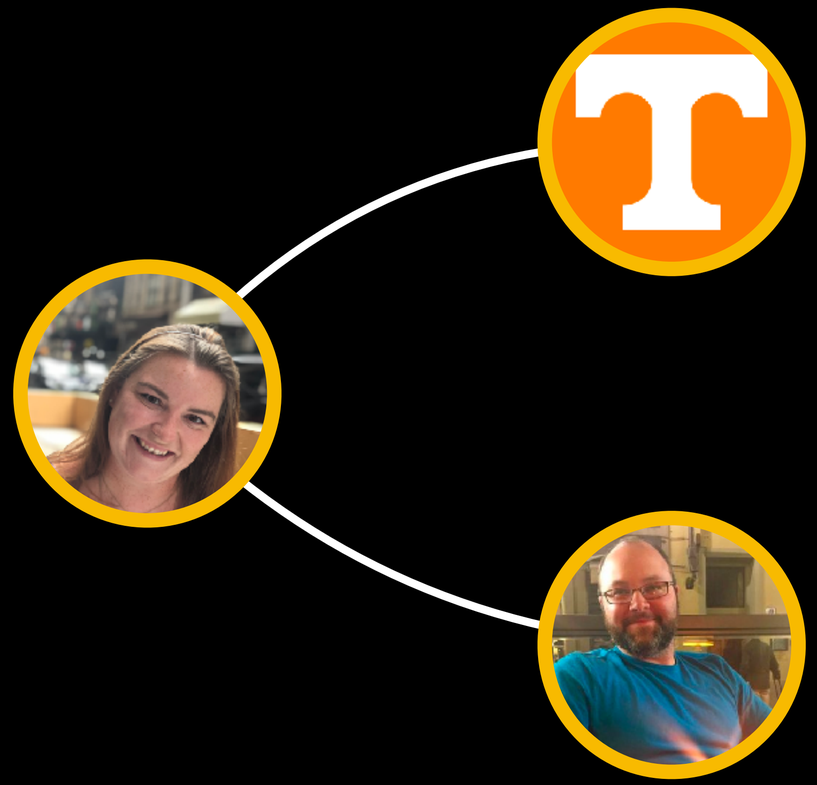
Debra Moriarity
3 mutual friends
Friends

Joshua Tucker
12 mutual friends
Friends

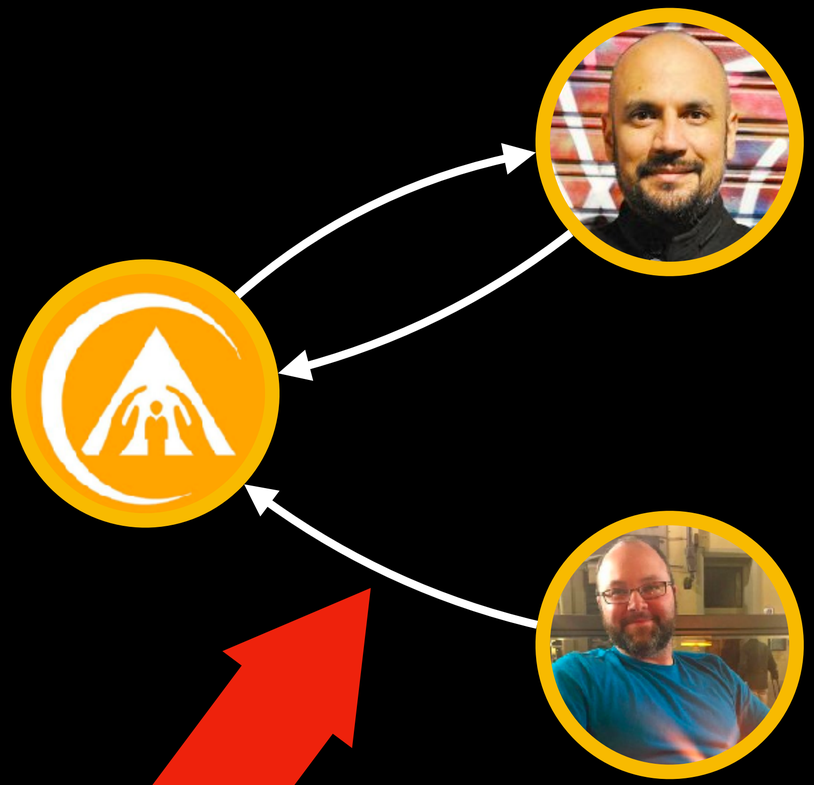
Christine Moulder
78 mutual friends
Friends



Directed Graph



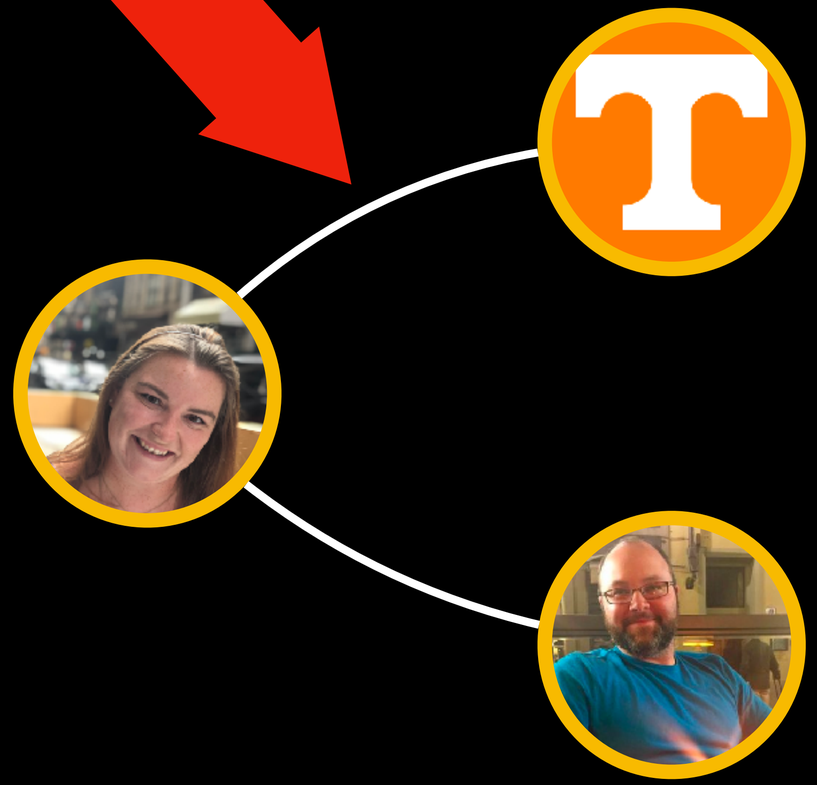
Undirected Graph



Directed Graph

Not Reciprocal

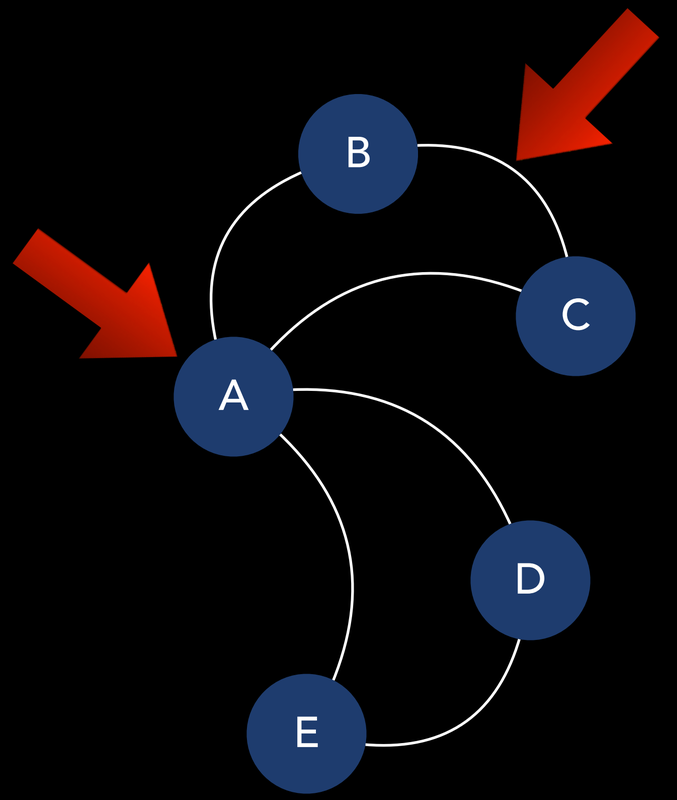
Reciprocal



Undirected Graph

WHAT IS A NETWORK?

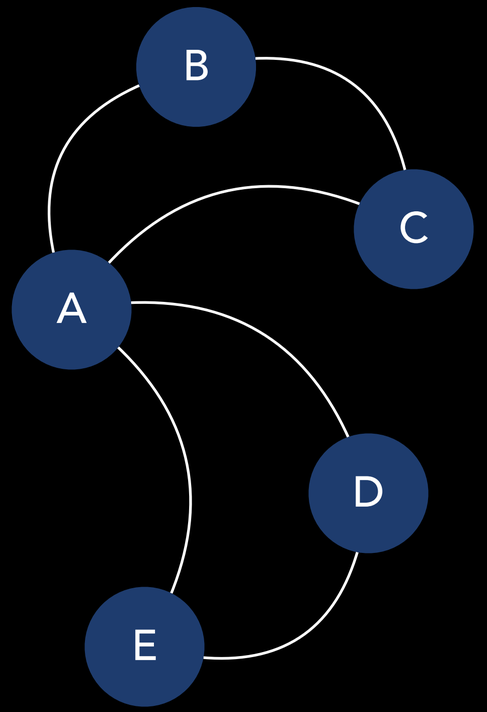
- A “network” or “graph” is:
  - A collection of nodes and edges
- Nodes - objects or entities
- Edges - connections between nodes



I’ll use “network” and “graph” interchangeably

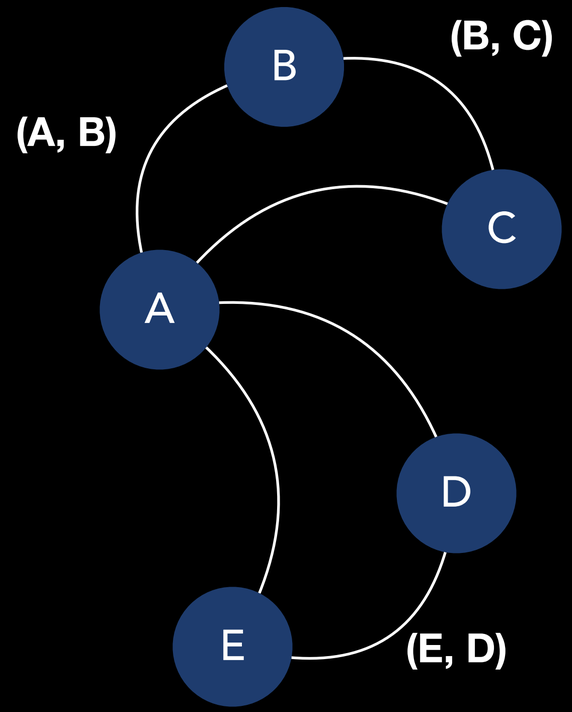
WHAT IS A NETWORK?

- Networks can represent many things
  - Friend/follower
  - Links between websites
  - Communication links between servers
  - etc.



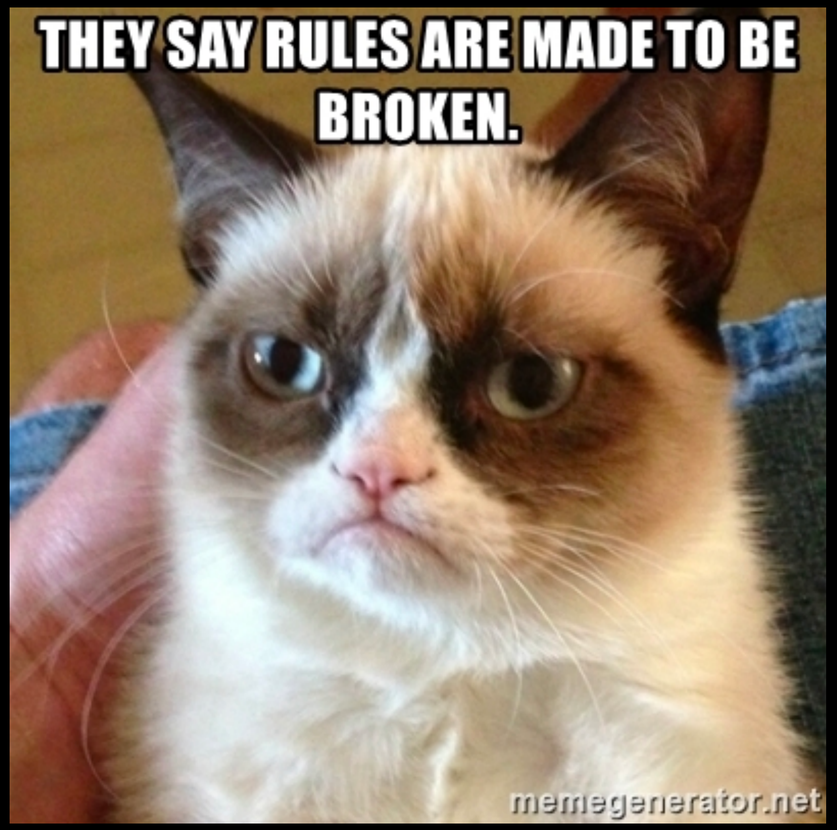
WHAT IS A NETWORK?

- Mathematical Definition:
  - *G(V, E)*
- *V* – Vertices (i.e., nodes)
- *E* – Edges
  - Tuples of vertices (v1, v2)
  - $E \subset V \times V$



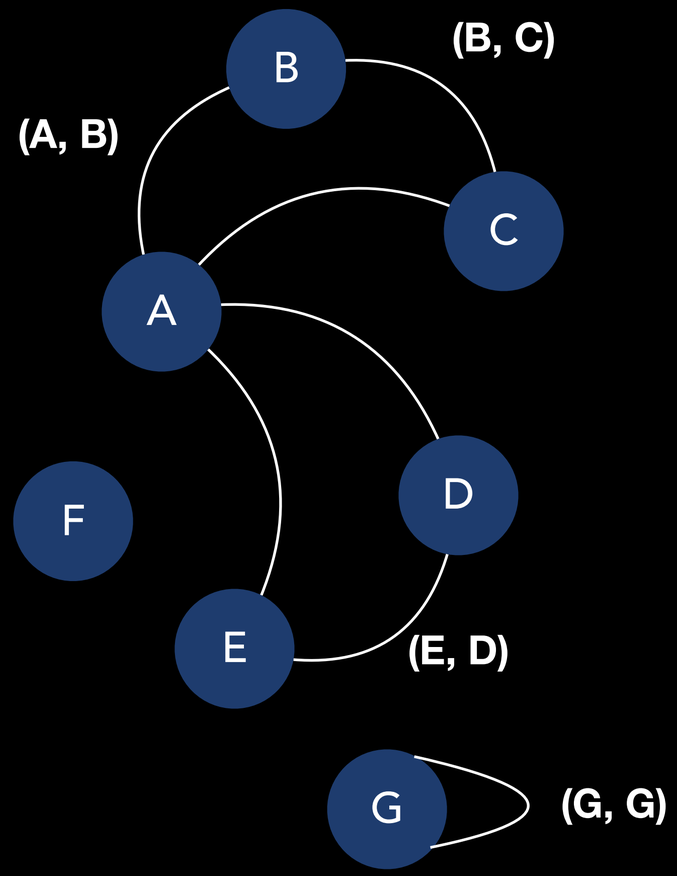
RULE OF THUMB ABOUT CONSISTENCY

- Nodes should generally of the same type of entity
- Edges should represent the same kind of relationship



WHAT IS A NETWORK? - VOCABULARY

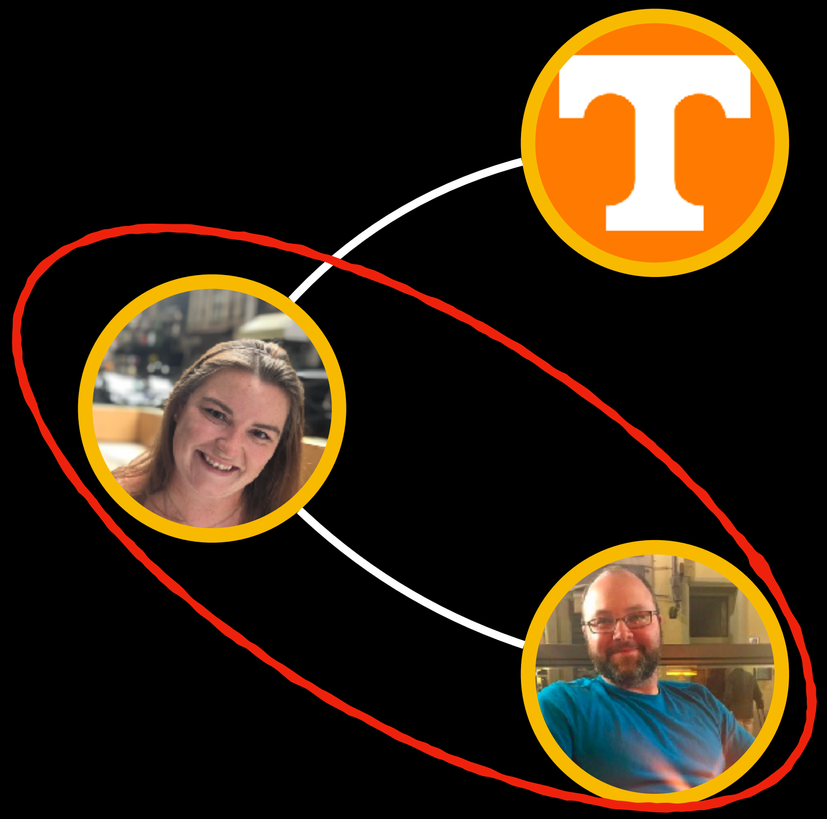
- A and B are neighbors
- F has no incident edges
  - No neighbors
  - I.e., a “singleton”
- Nodes can be connected to themselves
  - I.e., a “loop”



Can also integrate “weights” into graphs

UNWEIGHTED AND WEIGHTED GRAPHS

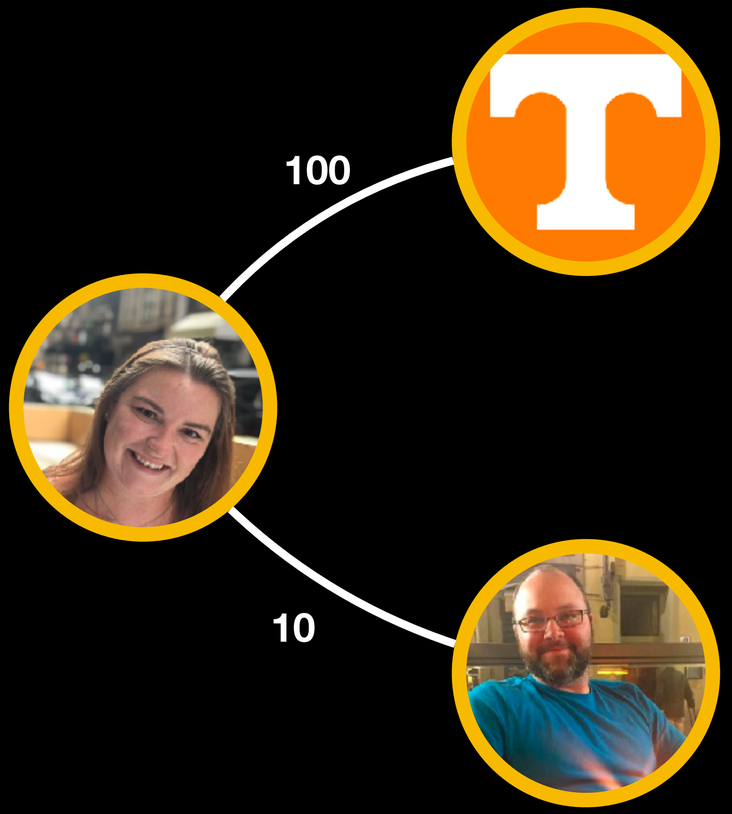
- Edges:
  - Tuple (pair) of vertices
  - All edges mean the same thing



But do all edges mean the same thing?

UNWEIGHTED AND WEIGHTED GRAPHS

- *Weighted* Edges:
  - Add numeric “weight” to tuple of vertices
- Example:
  - E = {
    (Leigh, Cody, 10),
    (Leigh, UT, 100)
    }



What other factors might edge “weights” represent?

UNWEIGHTED AND WEIGHTED GRAPHS

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases.

I will argue, in this paper, that the analysis of processes in interpersonal networks provides the most fruitful micro-macro bridge. In one way or another, it is through these networks that small-scale interaction becomes translated into large-scale patterns, and that these, in turn, feed back into small groups.

Sociometry, the precursor of network analysis, has always been curiously

- Link count
- Duration
- Proximity

What does the weight/tie strength mean?

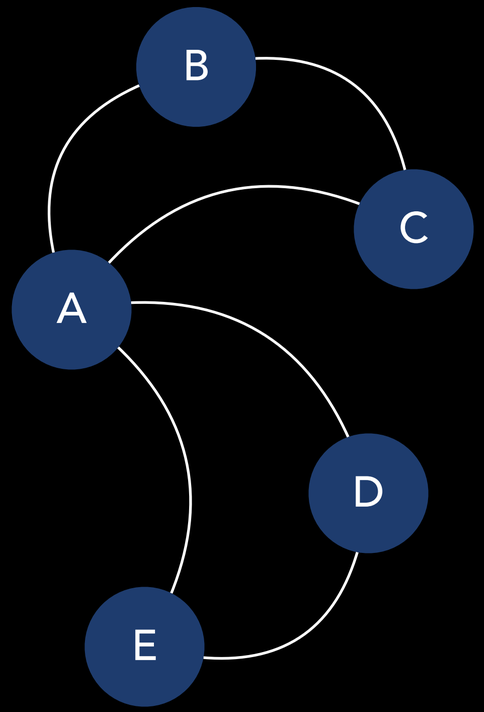
Importance of domain expertise

Representing Graphs

Representing and storing graphs

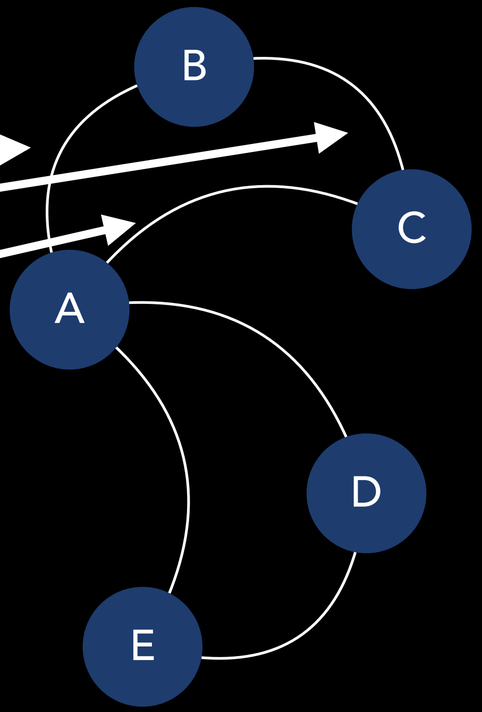
REPRESENTING GRAPHS - EDGE LISTS

- Mathematical Definition: *G(V, E)*
- *E* – Edges
  - Tuples of vertices (v1, v2)
- Edge Lists:
  - A list of all edge tuples in the graph



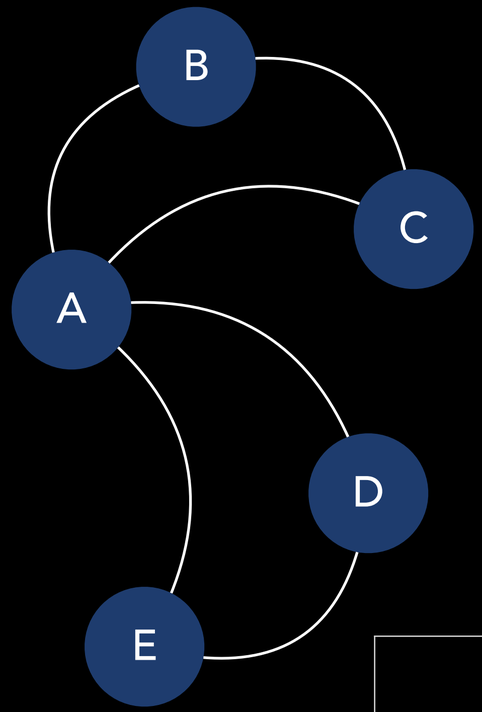
REPRESENTING GRAPHS - EDGE LISTS

- A,B
- B,C
- A,C
- A,E
- A,D
- E,D



REPRESENTING GRAPHS - ADJACENCY MATRIX

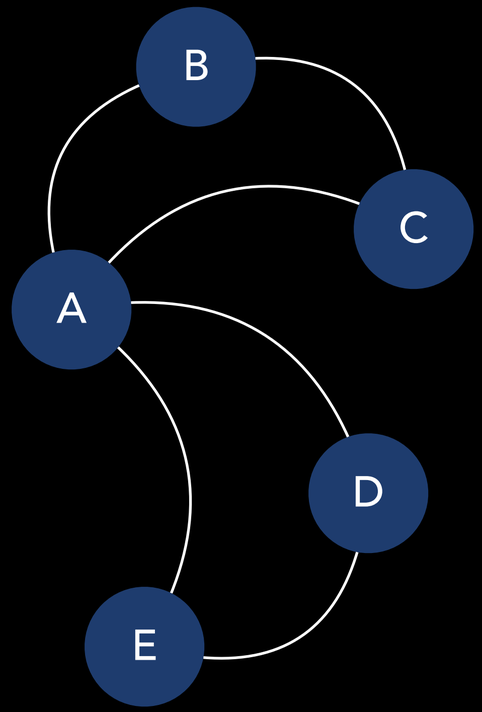
- Alternate Mathematical Definition: *G(V, E)*
- N = |V|, number of nodes
- Matrix A $\in R^{NxN}$
  - "Adjacency matrix"
  - $a_{i,j}$ = value at i^{th} row and j^{th} column in A
- $a_{i,j}$ = 0 if node i and j are unconnected



	A	B	C	D	E
A	0	1	1	1	1
B	1	0	1	0	0
C	1	1	0	0	0
D	1	0	0	0	1
E	1	0	0	1	0

REPRESENTING GRAPHS - ADJACENCY MATRIX

	A	B	C	D	E
A	0	1	1	1	1
B	1	0	1	0	0
C	1	1	0	0	0
D	1	0	0	0	1
E	1	0	0	1	0



Adjacency Matrix has many uses

REPRESENTING GRAPHS - OTHER OPTIONS

- Adjacency lists
  - Node1 neighbor1 neighbor2…
  - Node2 neighbor1 neighbor2…
- GraphML - Graph markup language
- GML
- GEXF
- And many others

Contact

Mailing list
Issue tracker
Source

Releases

Stable (notes)

2.5 — August 2020
download | doc | pdf

Latest (notes)

2.6 development
github | doc | pdf

Archive

NetworkX

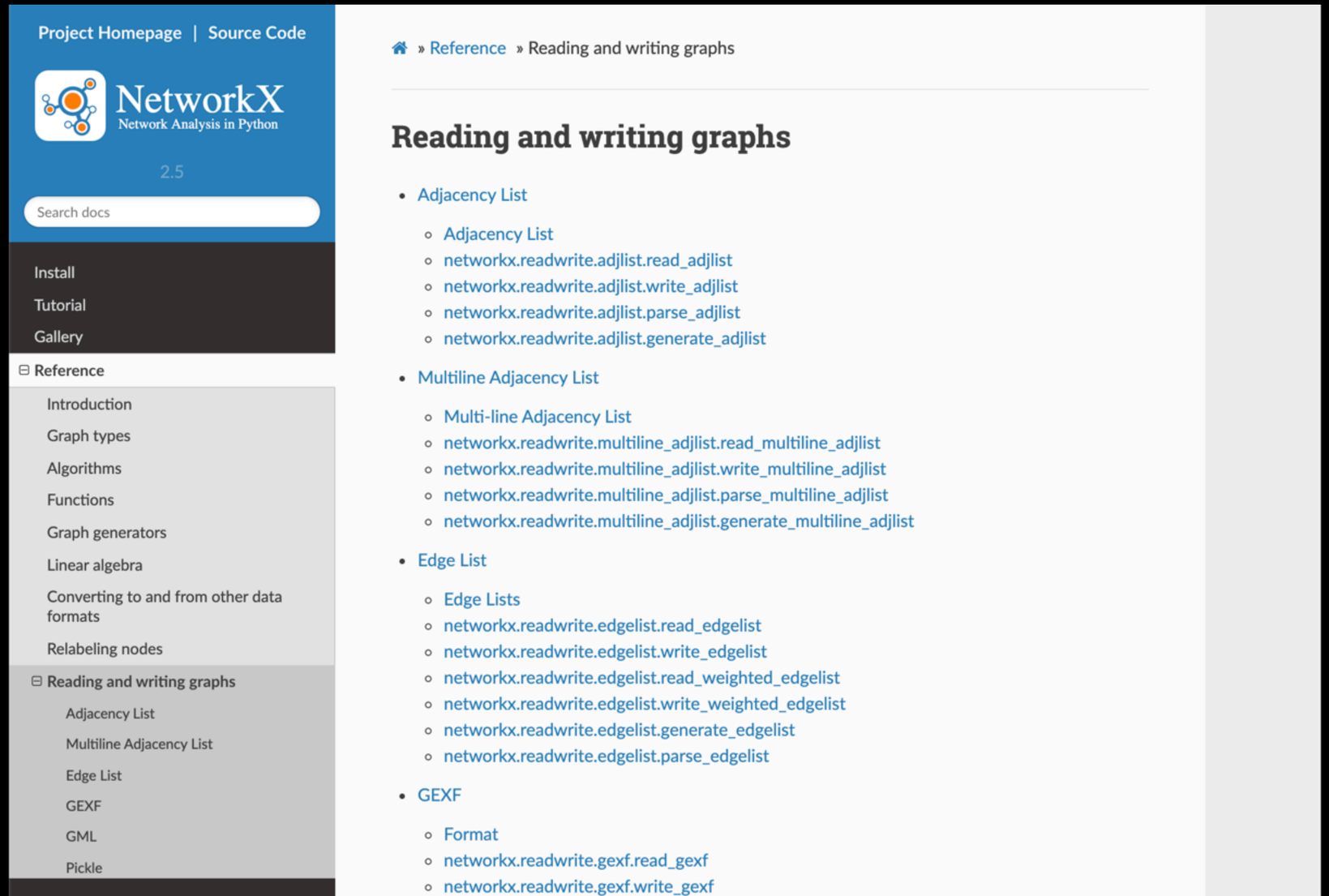
Network Analysis in Python

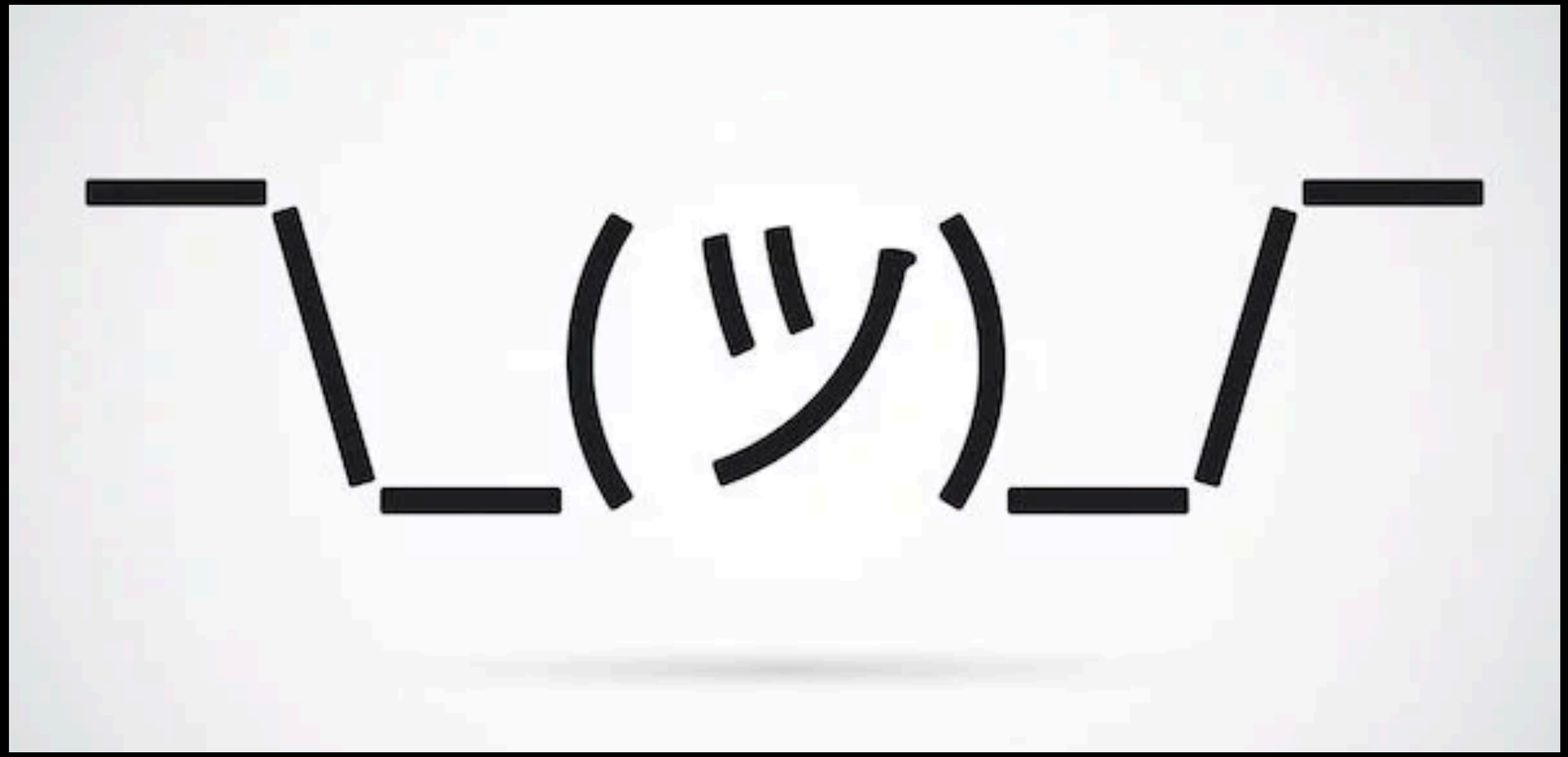
NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Software for complex networks

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

REPRESENTING GRAPHS - OTHER OPTIONS





What does this have to do with the Data Science?

This Lecture's Learning Objectives

Construct graphs from web structures

Describe at least two ways how one may try to exploit search engines

Explain how PageRank captures "importance" in a graph structure

Construct a transition matrix to model a collection of web pages

Differentiate hubs and authorities

You likely have an intuition about networks and their structure

Motivating Question: How do find insights from Web data?

botw.org

Apps | Jordan's Style | https://inf.ethz.ch/... | tome - Synology... | Other Bookmarks

Best of the Web — Since 1994

Best of the Web TRUSTED

Directory | Submit Site | Blog | Sign In

Search

Arts
Movies Television Music

Business
Jobs B2B Investing

Computers
Internet Software Hardware Hosting

Developers and Designers
Web Development Tutorials Designers

Finance
Insurance Banking Loans Mortgages
Financial Planning

Family Real Estate Gardening

Home Services
HVAC Movers Appliance Repair Pest Co
Plumbers

rs and Firms
Injury Divorce Attorneys Bankruptcy
lanning Criminal Defense

Marketing
Software Industry Specific SEO Compan
Social Media Local Marketing

nce
s Maps Education Libraries

Science
Biology Physics Technology Social Scier

Baseball Soccer Basketball

Regional
United States Canada Australia United K

Trust Badge - Get Your Site Verified

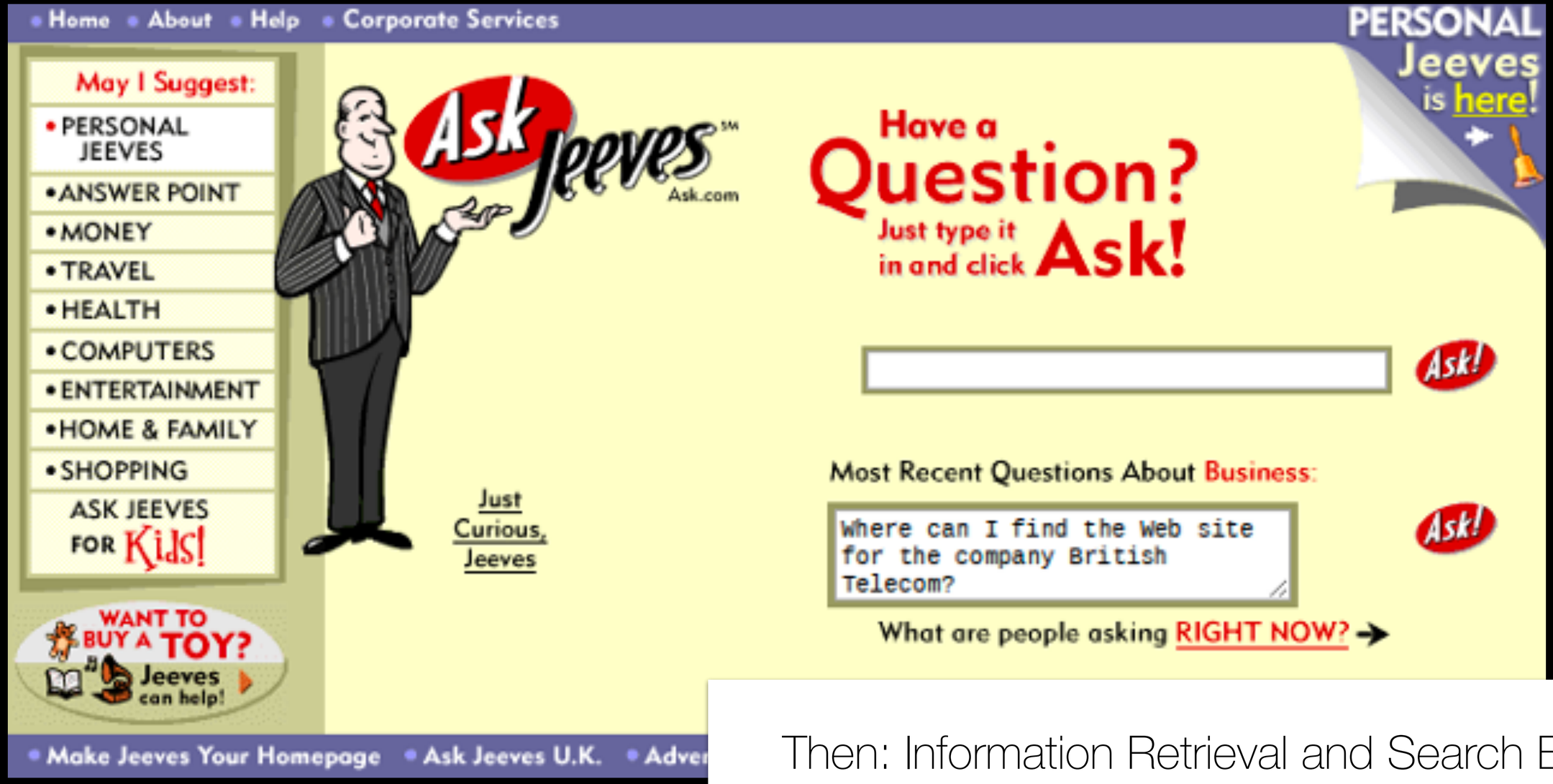
Best of the Web

The Internet's Oldest Business Directory

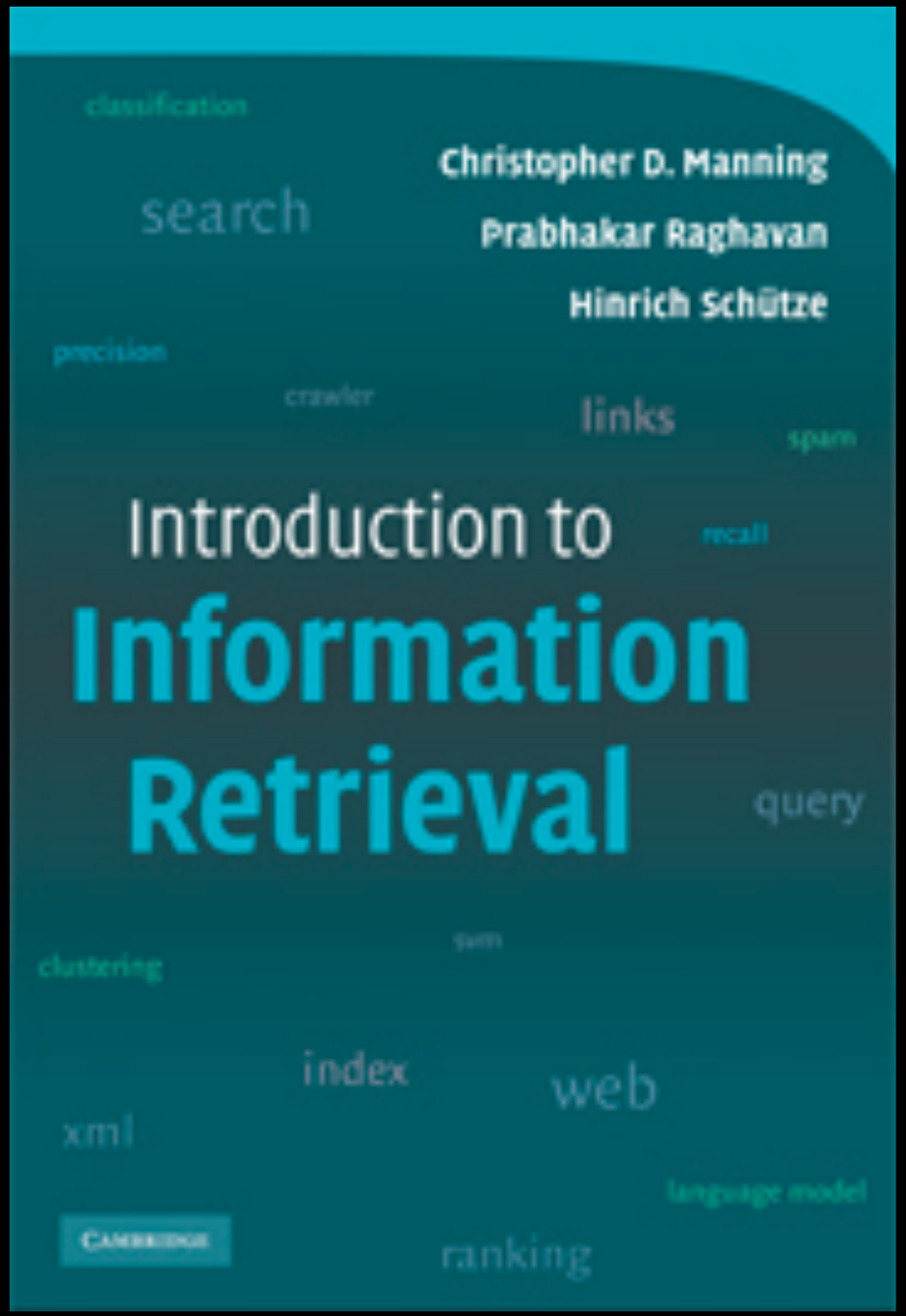
Welcome to Best of the Web - the world's most authoritative online directory. If you want to increase your business website's visibility and see a real boost to your traffic, a BOTW listing is an easy win. Best of the Web is one of the most visited online directories so we consistently drive targeted traffic to our listed sites - and with over 400,000 unique categories to choose from, it's easy to find the perfect relevant place for your business' directory listing. Every website we list is hand-reviewed by our expert editors, ensuring a spam-free

Every business owner or operator understands the power of online reviews, right? According to Podium, online reviews are becoming ever more important to how we as consumers do business, with 93% of consumers saying that online reviews do

Early Methods: Curated Web directories



Then: Information Retrieval and Search Engines



Query: Amanda Gorman

Amanda Gorman

From Wikipedia, the free encyclopedia

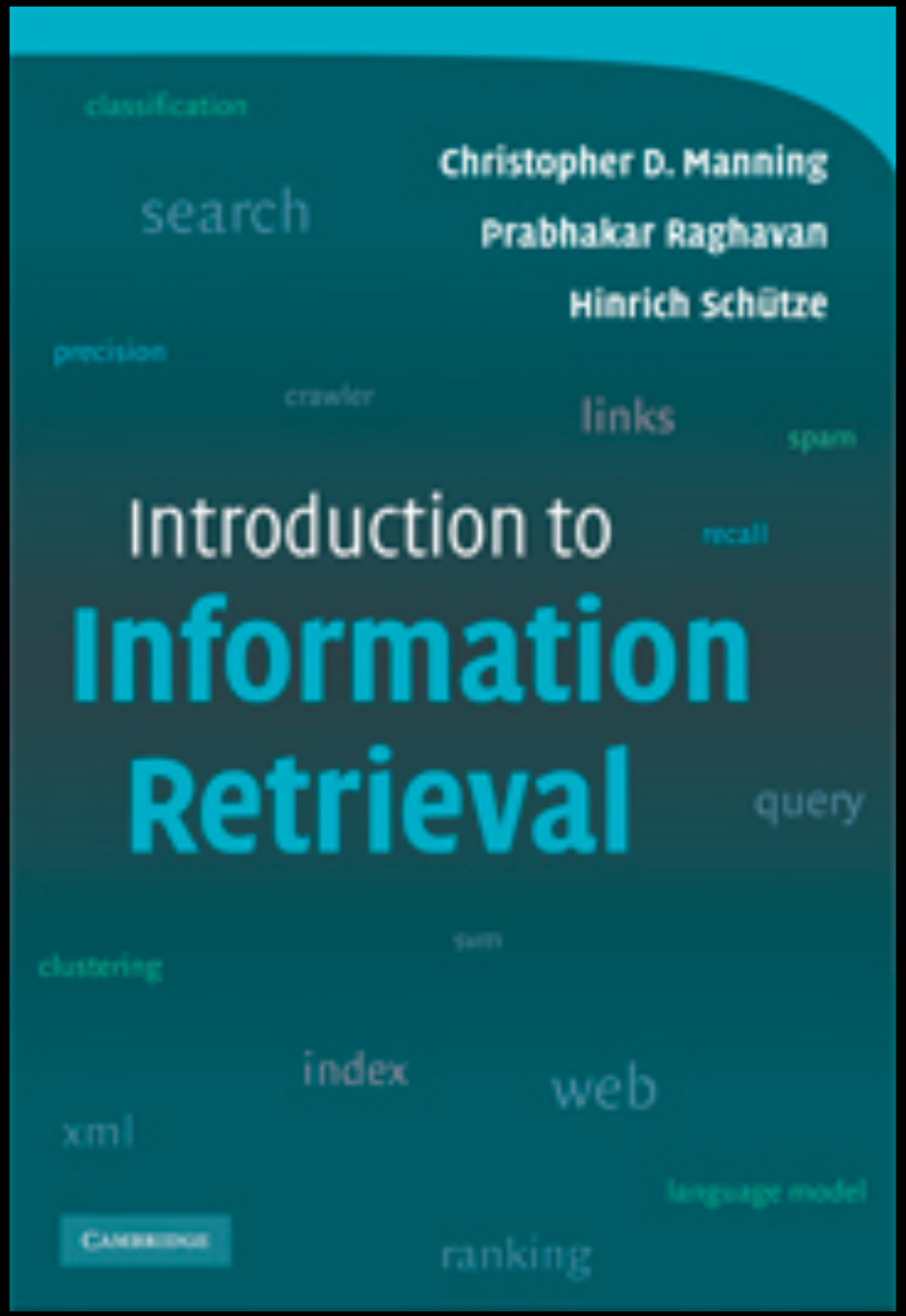
Amanda S. C. Gorman[1] (born 1998) is an American poet and activist. Her work focuses on issues of oppression, feminism, race, and marginalization, as well as the African diaspora. Gorman was the first person to be named National Youth Poet Laureate. She published the poetry book *The One for Whom Food Is Not Enough* in 2015. In 2021, she delivered her poem "The Hill We Climb" at the inauguration of U.S. President Joe Biden. Her inauguration poem generated international acclaim, stimulated her two books to reach best-seller

Wordsmith. Change-maker.

Amanda Gorman is the youngest inaugural poet in U.S. history, as well as an award-winning writer and *cum laude* graduate of Harvard University, where she studied Sociology. She has written for the *New York Times* and has three books forthcoming with Penguin Random House.

Born and raised in Los Angeles, she began writing at only a few years of age. Now her words have won her invitations to the Obama White

Such approaches are subject to manipulation though



Query: Amanda Gorman

Amanda Gorman

From Wikipedia, the free encyclopedia

Amanda S. C. Gorman[1] (born 1998) is an American poet and activist. Her work focuses on issues of oppression, feminism, race, and marginalization, as well as the African diaspora. Gorman was the first person to be named National Youth Poet Laureate. She published the poetry book *The One for Whom Food Is Not Enough* in 2015. In 2021, she delivered her poem "The Hill We Climb" at the inauguration of U.S. President Joe Biden. Her inauguration poem generated international acclaim, stimulated her two books to reach best-seller

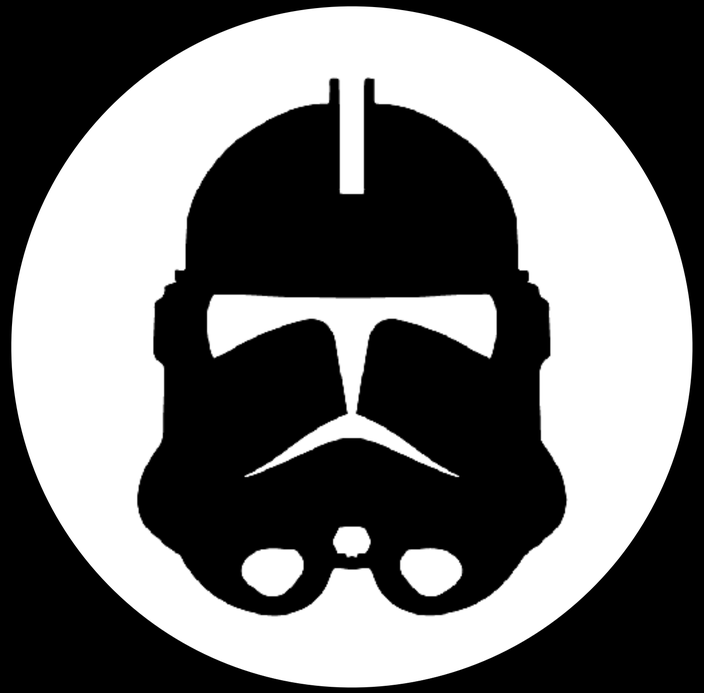
Wordsmith. Change-maker.

Amanda Gorman is the youngest inaugural poet in U.S. history, as well as an award-winning writer and *cum laude* graduate of Harvard University, where she studied Sociology. She has written for the *New York Times* and has three books forthcoming with Penguin Random House.

Born and raised in Los Angeles, she began writing at only a few years of age. Now her words have won her invitations to the Obama White

Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,

How might this weakness be exploited?



Relying on websites to self-identify as “relevant” leaves us open to exploitation

Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman

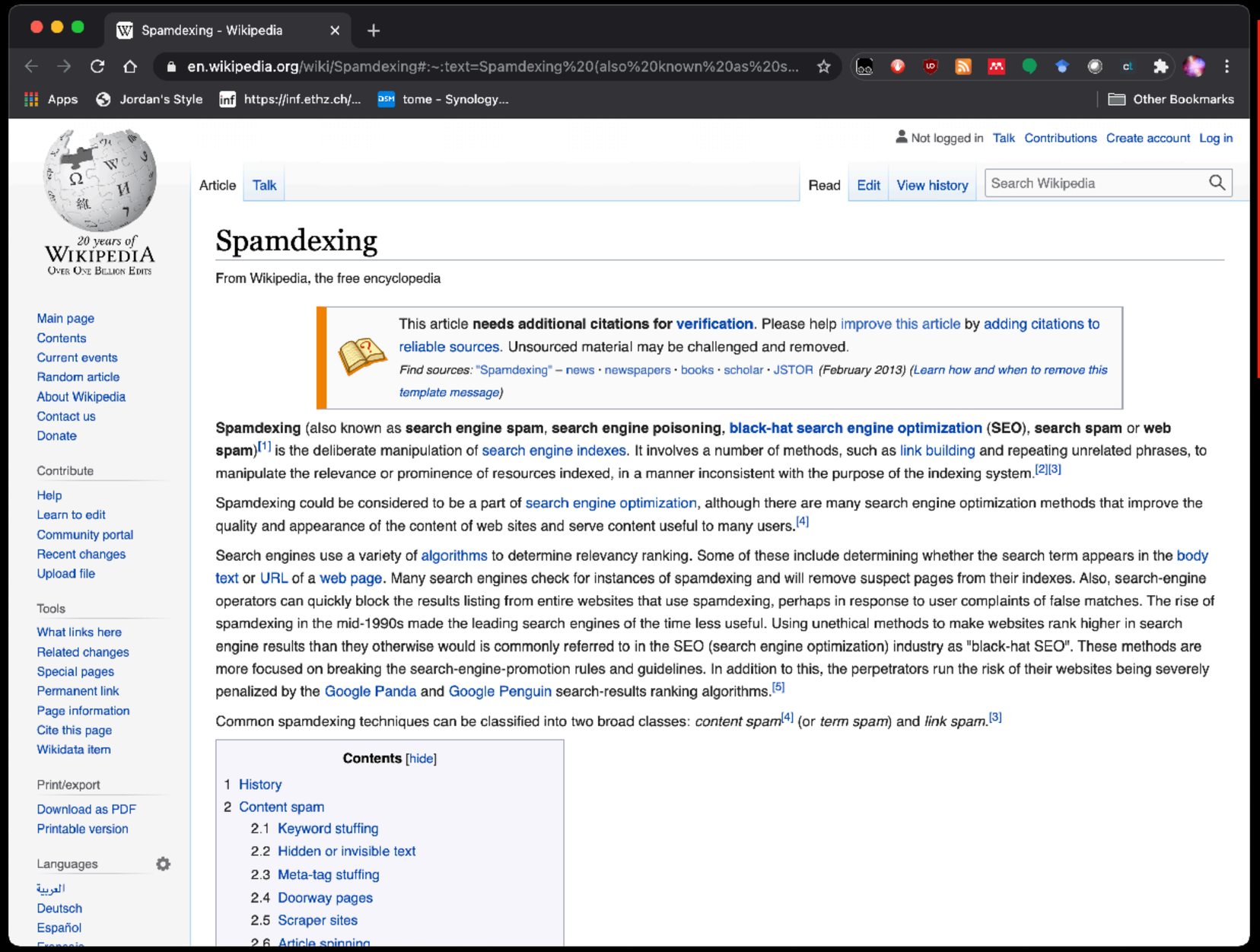
Amanda gorman, Amanda gorman,
Amanda gorman, Amanda gorman,
Amanda gorman, Amanda gorman,
Amanda gorman, Amanda gorman,
Amanda gorman,
Amanda gorman

Amanda gorman, Amanda gorman,
Amanda gorman, Amanda gorman,
Amanda gorman, Amanda gorman,
Amanda gorman, Amanda gorman,
Amanda gorman,
Amanda gorman

Fill your web page with popular keywords

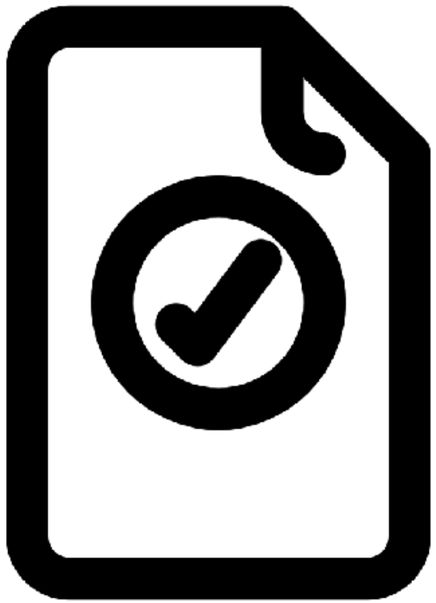
```
    <title>How to create the right meta description &bull; Yoast</title>

    <!-- This site has installed PayPal for WooCommerce v2.1.6 - https://www.angelleye.com/product/woocommerce-paypal-plugin/ -->
<link rel="dns-prefetch" href="//www.googletagmanager.com">
<link rel="preconnect" href="https://www.googletagmanager.com/" crossorigin>

<!-- Google Tag Manager for WordPress by gtm4wp.com -->
<script data-cfasync="false" data-pagespeed-no-defer>//<![CDATA[
    var gtm4wp_datalayer_name = "dataLayer";
    var dataLayer = dataLayer || [];
    var gtm4wp_use_sku_instead        = 1;
    var gtm4wp_id_prefix              = '';
    var gtm4wp_remarketing            = false;
    var gtm4wp_eec                    = 1;
    var gtm4wp_classicec              = 1;
    var gtm4wp_currency               = 'USD';
    var gtm4wp_product_per_impression = 0;
    var gtm4wp_needs_shipping_address = false;
//]]>
</script>
<!-- End Google Tag Manager for WordPress by gtm4wp.com -->
    <!-- This site is optimized with the Yoast SEO Premium plugin v15.7 - https://yoast.com/wordpress/plugins/seo/ -->
    <meta name="description" content="Do you want people to click on your search result? Learn how to write the best meta description. Including 7 characteristics and examples!" />
    <meta name="robots" content="index, follow, max-snippet:-1, max-image-preview:large, max-video-preview:-1" />
    <link rel="canonical" href="https://yoast.com/meta-descriptions/" />
    <meta property="og:locale" content="en_US" />
    <meta property="og:type" content="article" />
    <meta property="og:title" content="How to create the right meta description &bull; Yoast" />
    <meta property="og:description" content="Do you want people to click on your search result? Learn how to write the best meta description. Including 7 characteristics and examples!" />
    <meta property="og:url" content="https://yoast.com/meta-descriptions/" />
    <meta property="og:site_name" content="Yoast" />
    <meta property="article:publisher" content="https://www.facebook.com/yoast" />
    <meta property="article:published_time" content="2020-12-21T12:30:00+00:00" />
    <meta property="article:modified_time" content="2021-01-26T14:12:35+00:00" />
    <meta property="og:image" content="https://yoast.com/app/uploads/2015/01/Meta_description_FI.png" />
    <meta property="og:image:width" content="1200" />
    <meta property="og:image:height" content="628" />
    <meta name="twitter:card" content="summary_large_image" />
    <meta name="twitter:creator" content="@yoast" />
    <meta name="twitter:site" content="@yoast" />
    <meta name="twitter:label1" content="Written by">
    <meta name="twitter:data1" content="Willemien Hallebeek">
    <meta name="twitter:label2" content="Est. reading time">
    <meta name="twitter:data2" content="11 minutes">
    <script type="application/ld+json" class="yoast-schema-graph">
[{"@type":"Organization","@id":"https://yoast.com/#organization","
["https://www.facebook.com/yoast","https://www.instagram.com/yoast/","https://www.linkedin.com/company/1414157/","https://www.youtube.com/yoast","htt
ps://www.pinterest.com/yoast/","https://en.wikipedia.org/wiki/Yoast","https://twitter.com/yoast"],"logo":
{"@type":"ImageObject","@id":"https://yoast.com/#logo","inLanguage":"en-
US","url":"https://yoast.com/app/uploads/2020/09/Yoast_Icon_SocialMedia_500x500.png","width":500,"height":500,"caption":"Yoast"},"image":
{"@id":"https://yoast.com/#logo"},"founder":{"@type":"Person","name":"Joost de Valk","url":"https://yoast.com/about-us/team/joost-de-
```

Can hide meta tags full of common search terms



Contents [hide]

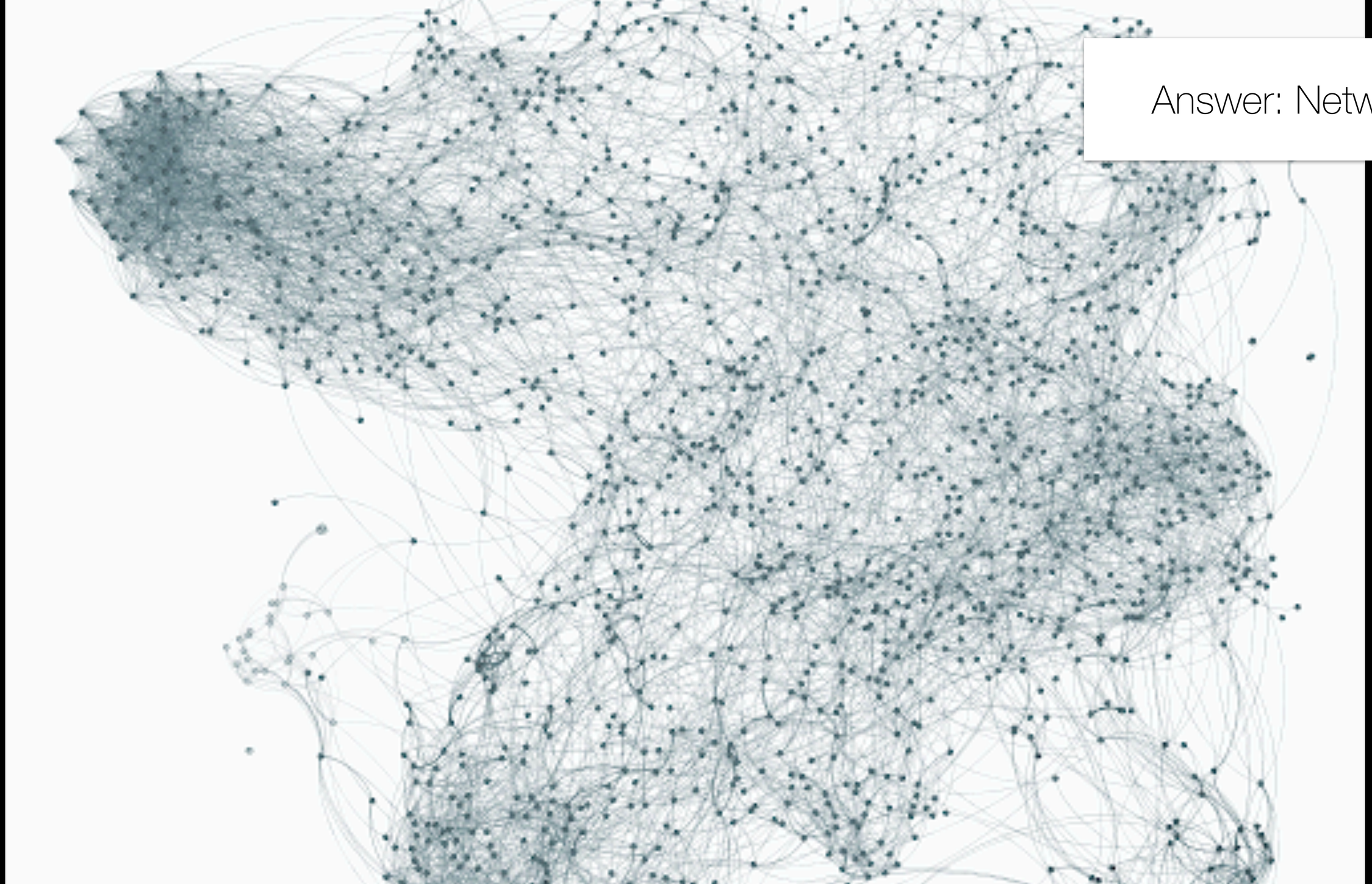
Motivating Question: How do extract insights from Web data?



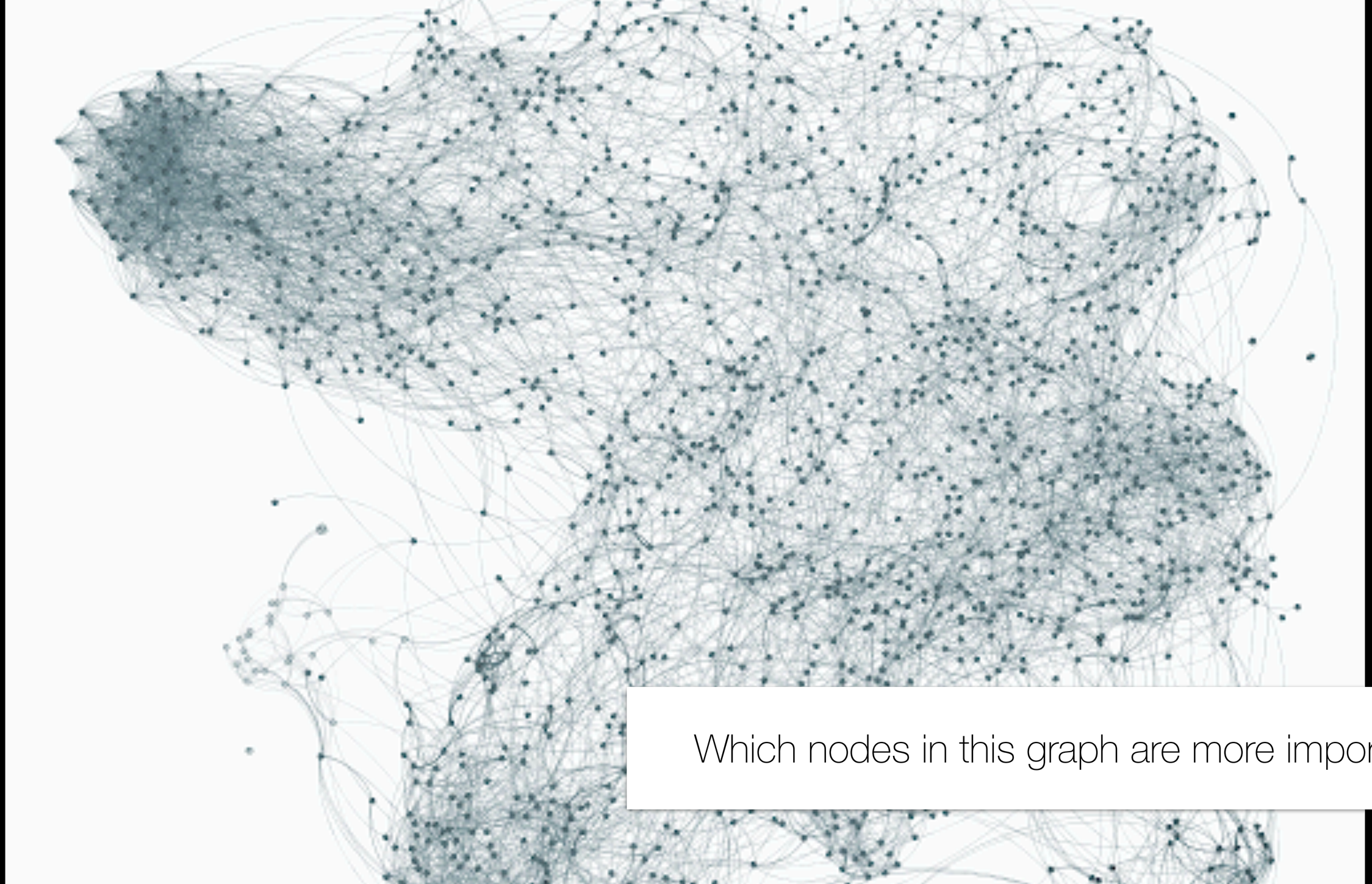
Finding Trustworthy Pages

Finding Relevant Pages

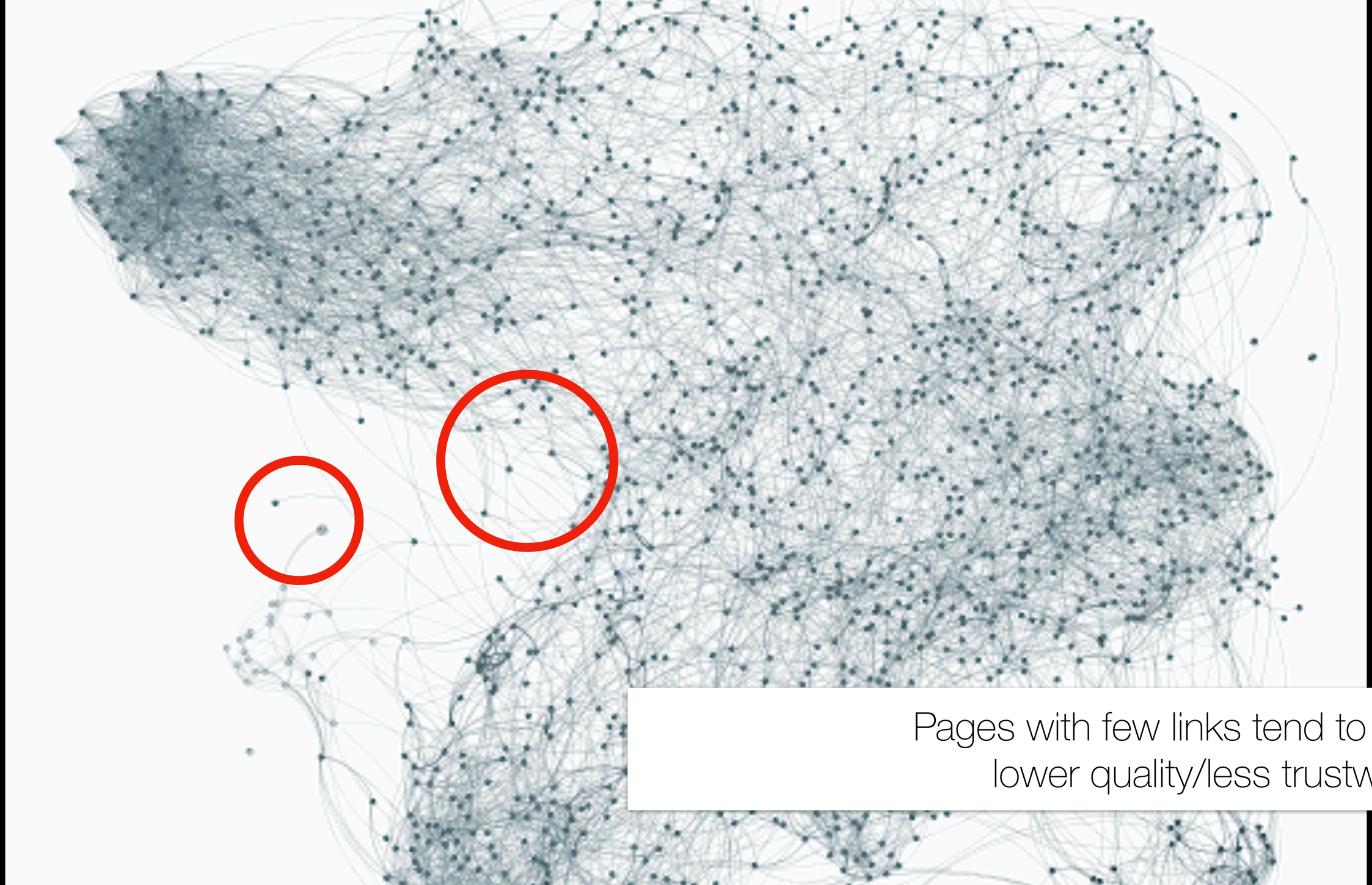
How do we find such pages?



Answer: Networks!

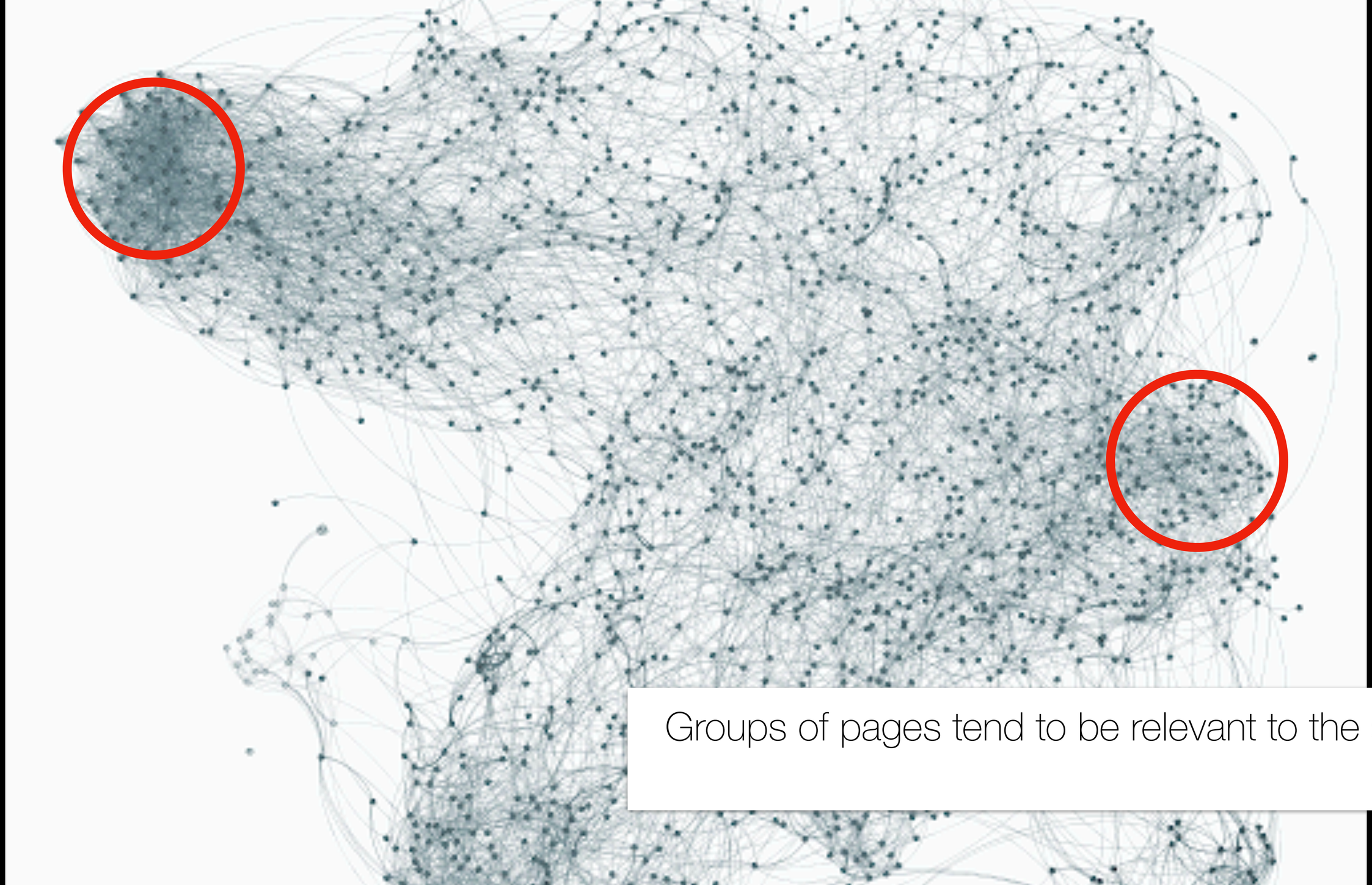


Which nodes in this graph are more important?

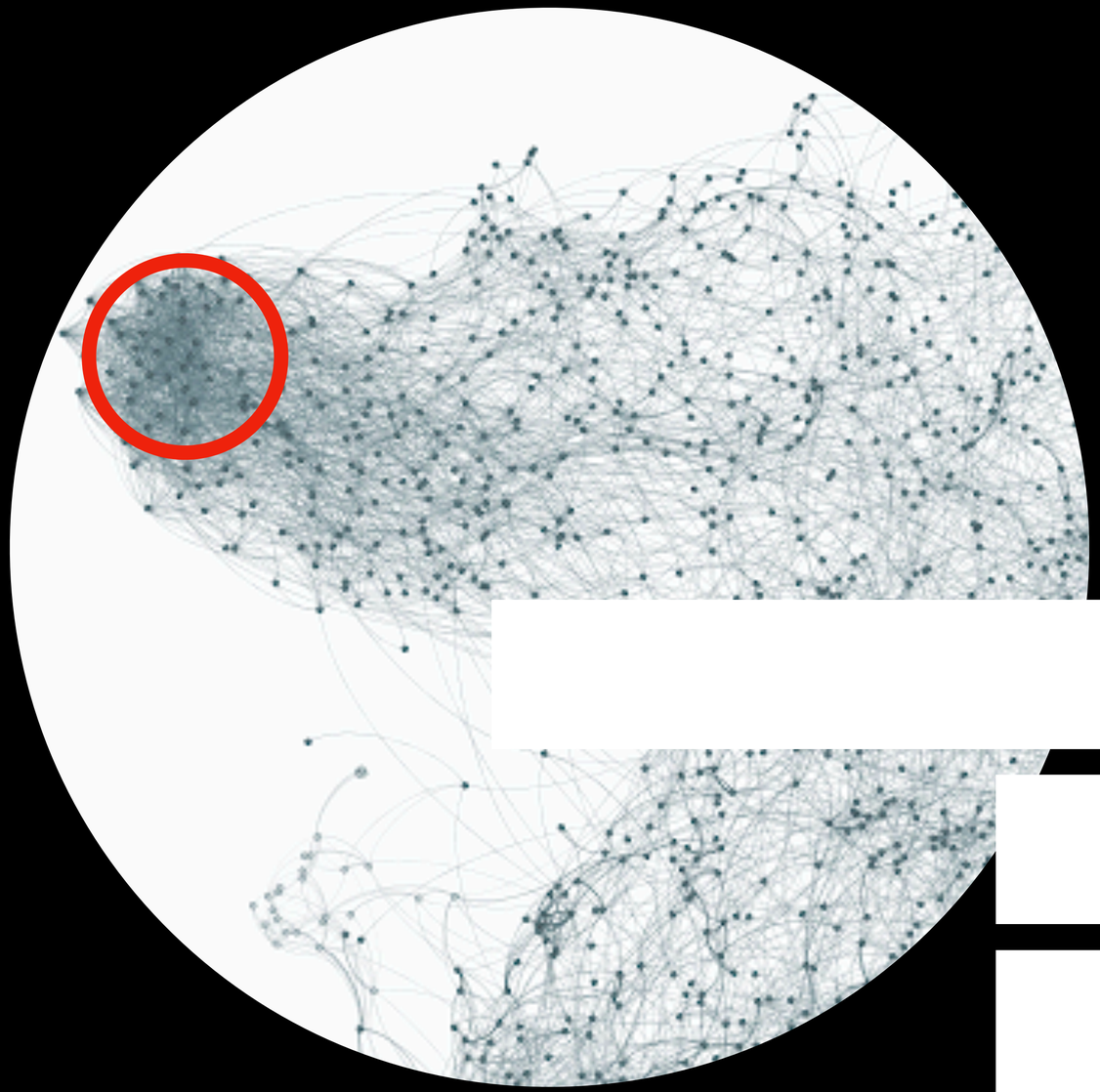


Pages with few links tend to be of lower quality/less trustworthy

Where do you think similar nodes
to this one are in this graph?



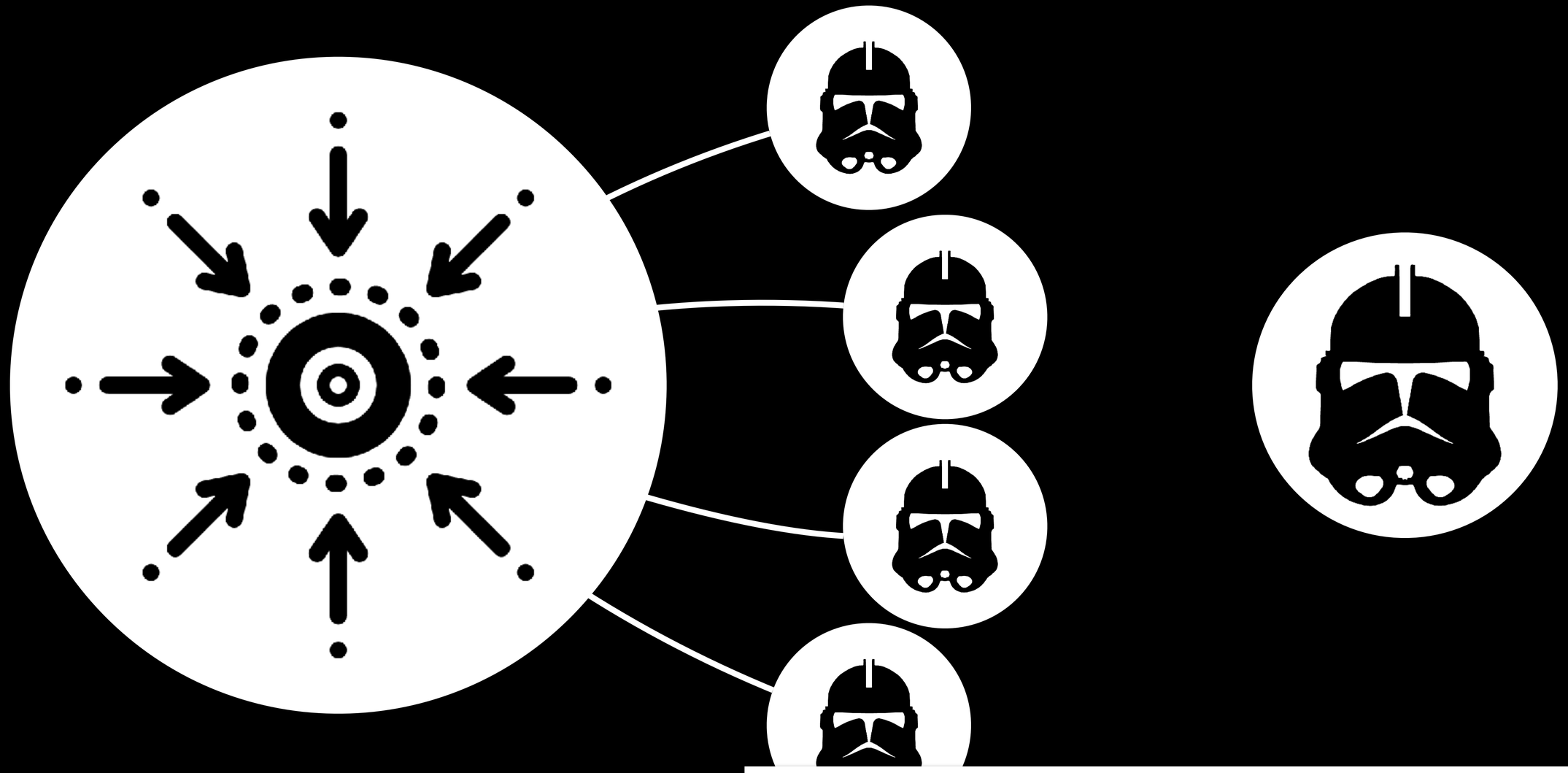
Groups of pages tend to be relevant to the same topic



How do we find pages in these dense clusters?

Pages in these clusters are more “important”

A web surfer is more likely to visit these pages



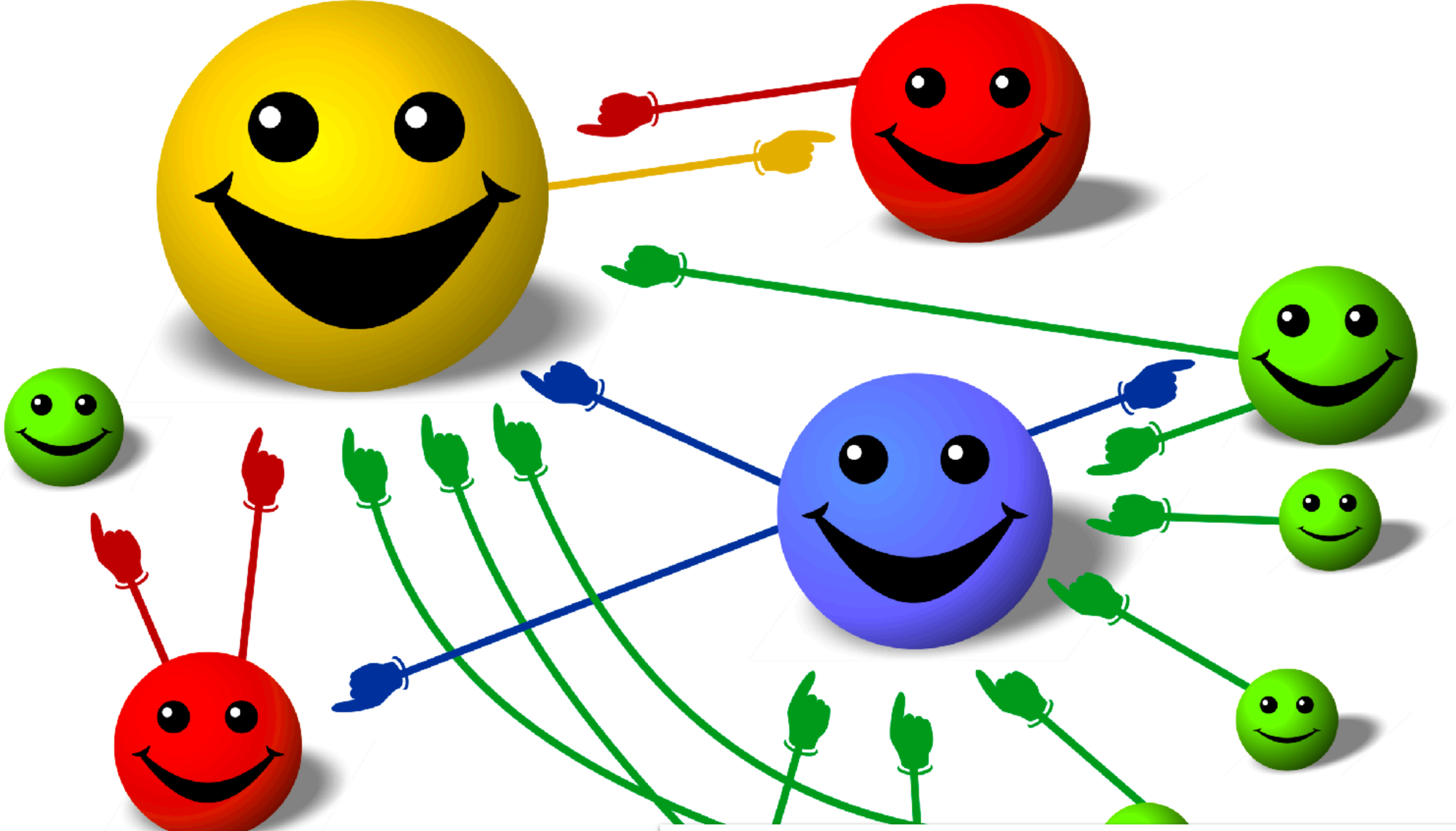
Naive Solution: Rank by Incoming-Link Count (In-Degree)

How do we find important, trustworthy pages?

Robust to keyword manipulation

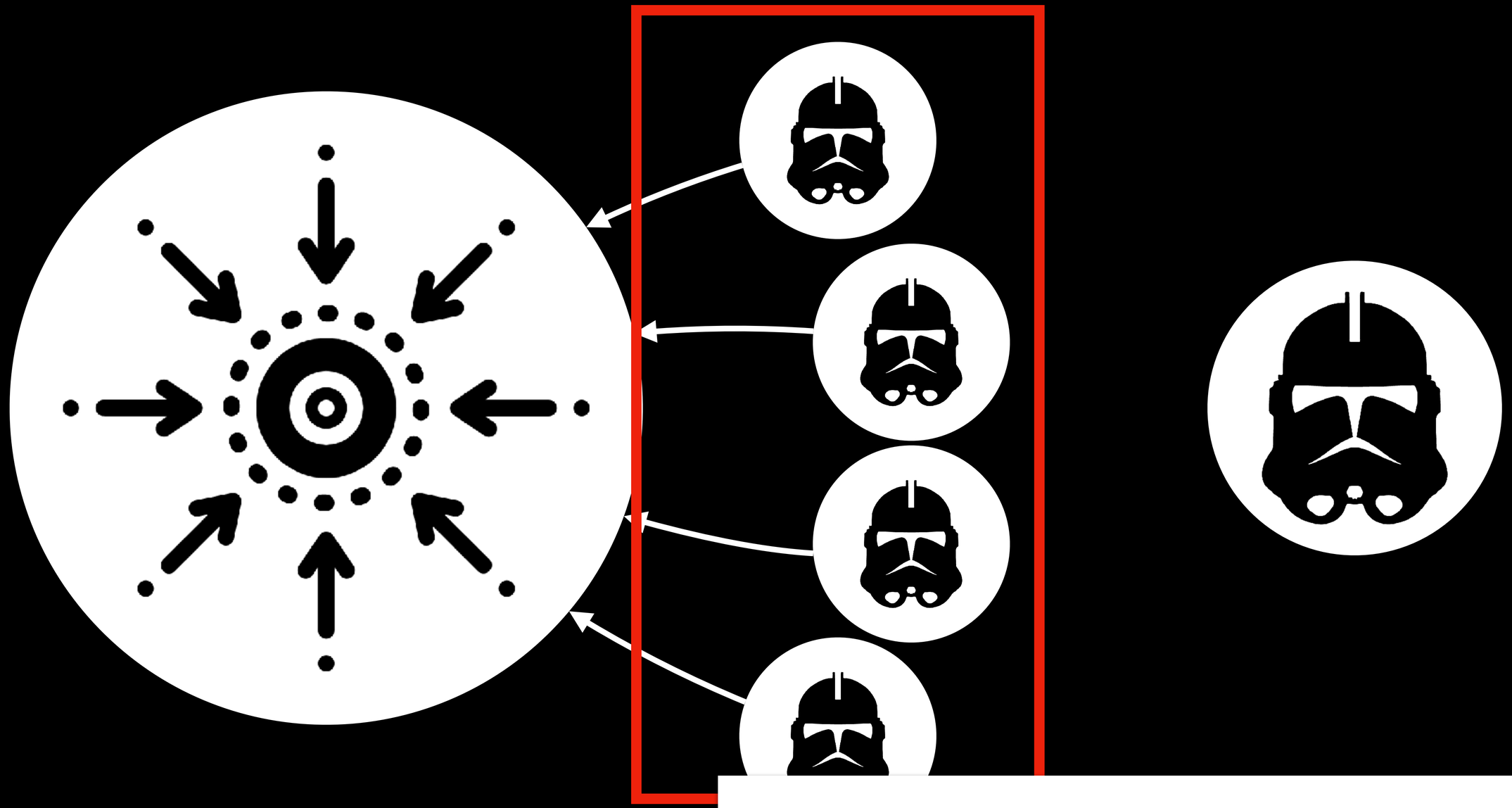
Robust to link manipulation

PageRank, by Larry Page and Sergey Brin

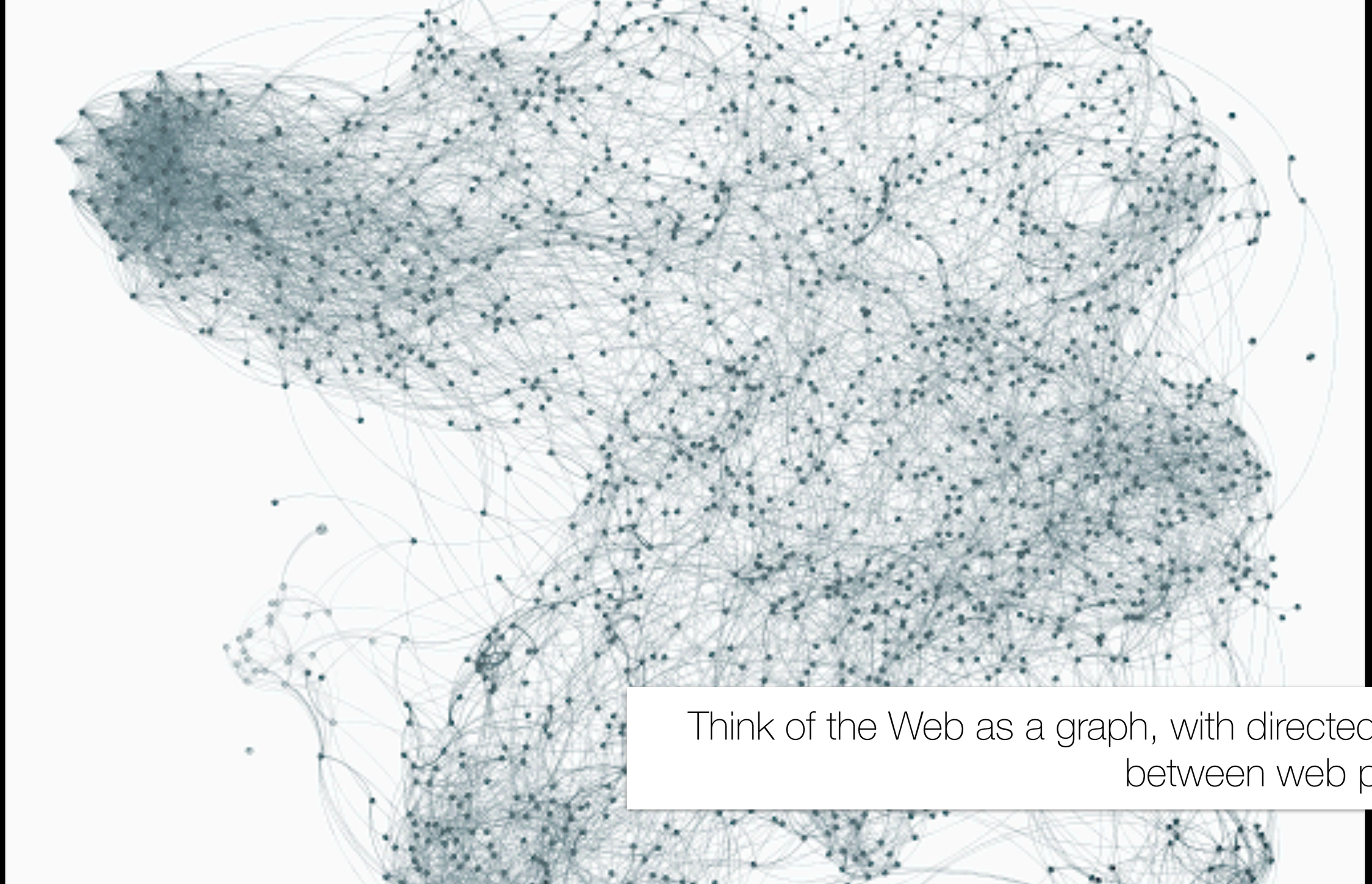


Size of each node is proportional to the sum of node-sizes for all nodes that point to it

PageRank

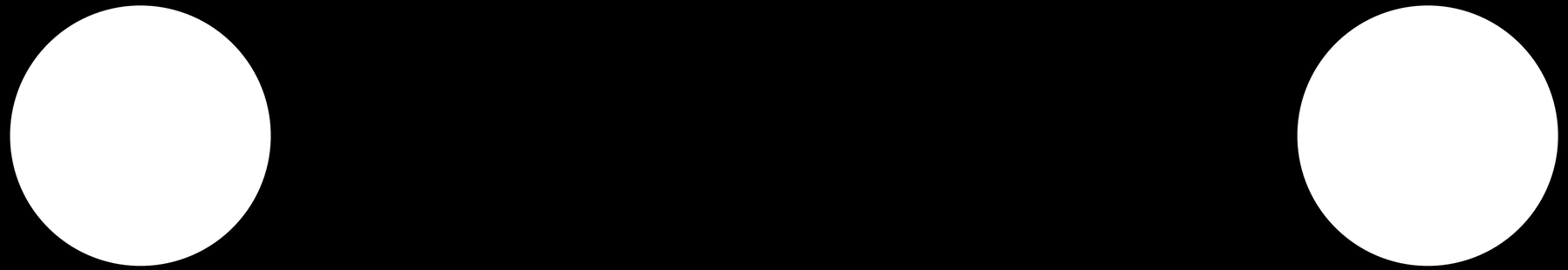


These “fake” nodes will have few incoming links

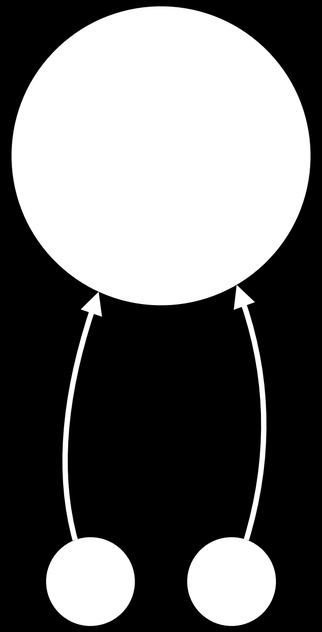


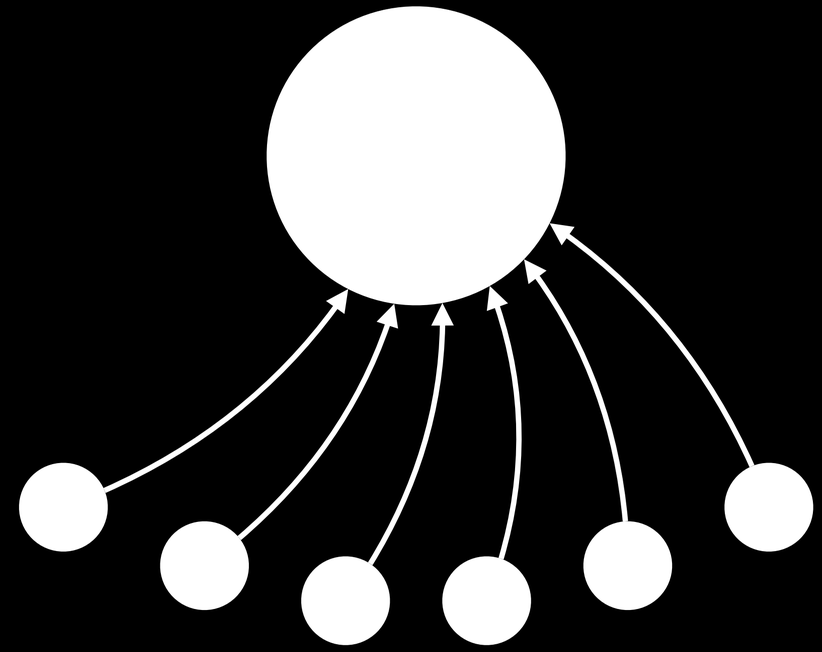
Think of the Web as a graph, with directed links between web pages

Intuition: Pages vote on each other via linking



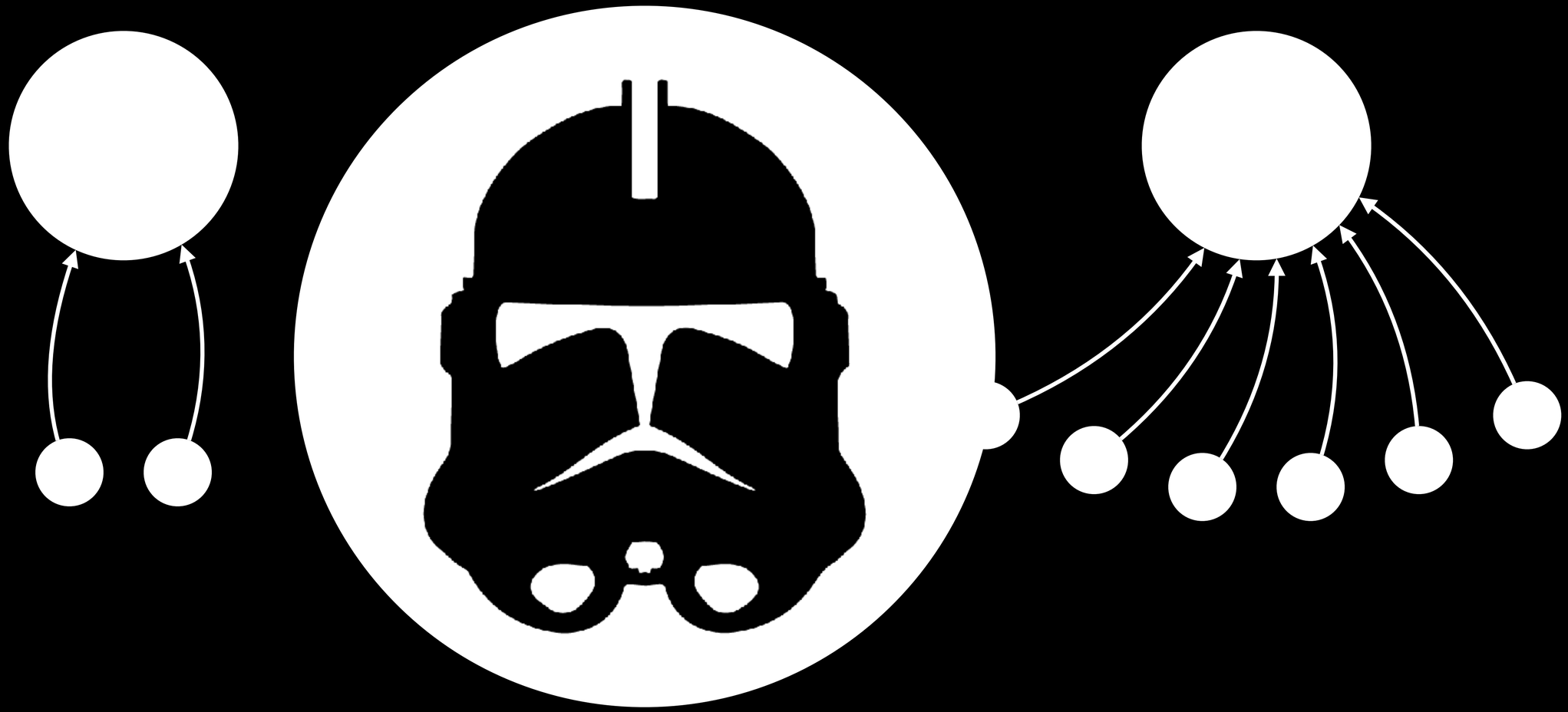
Intuition: Pages vote on each other via linking



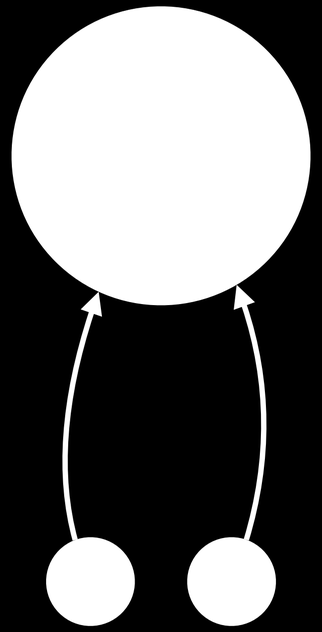


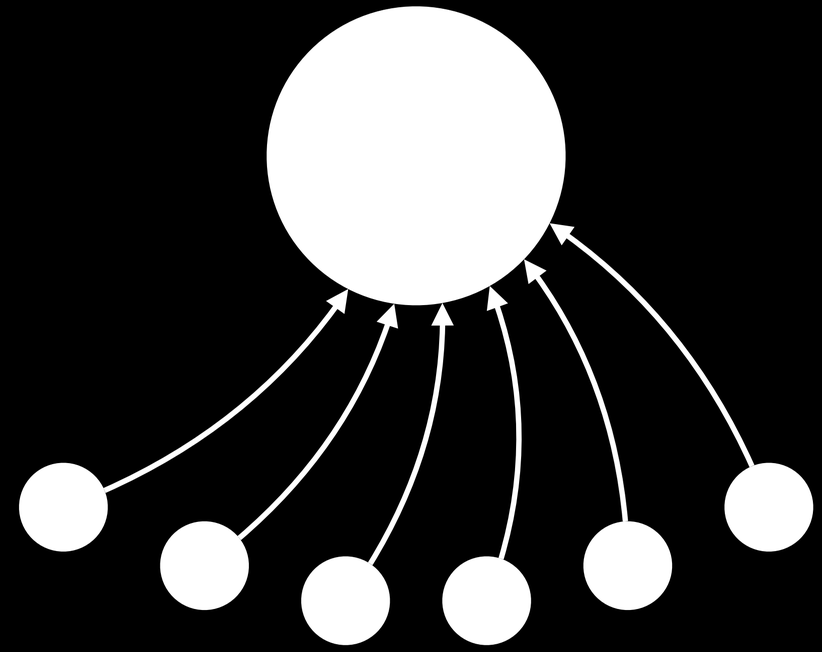
Pages with more in-links are more important

Intuition: Pages vote on each other via linking

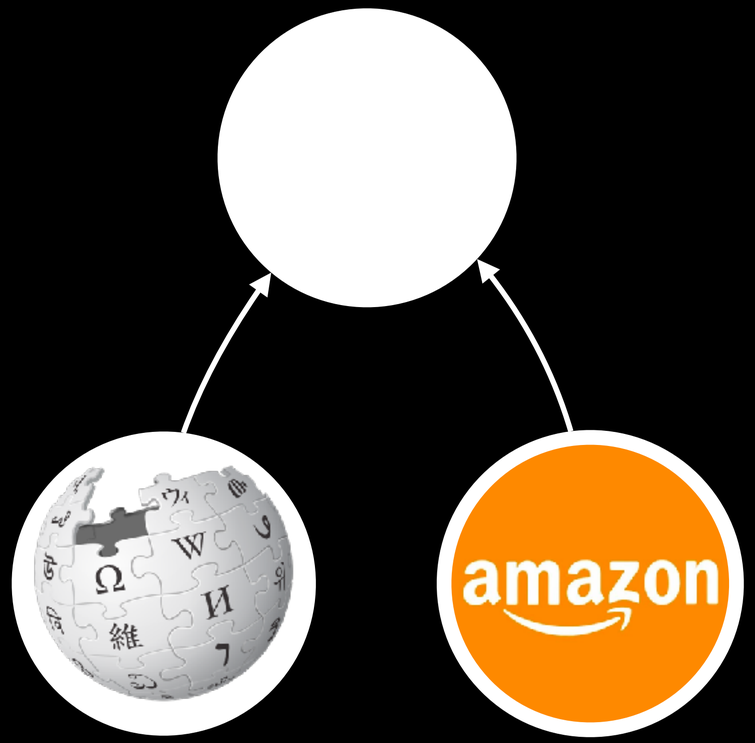


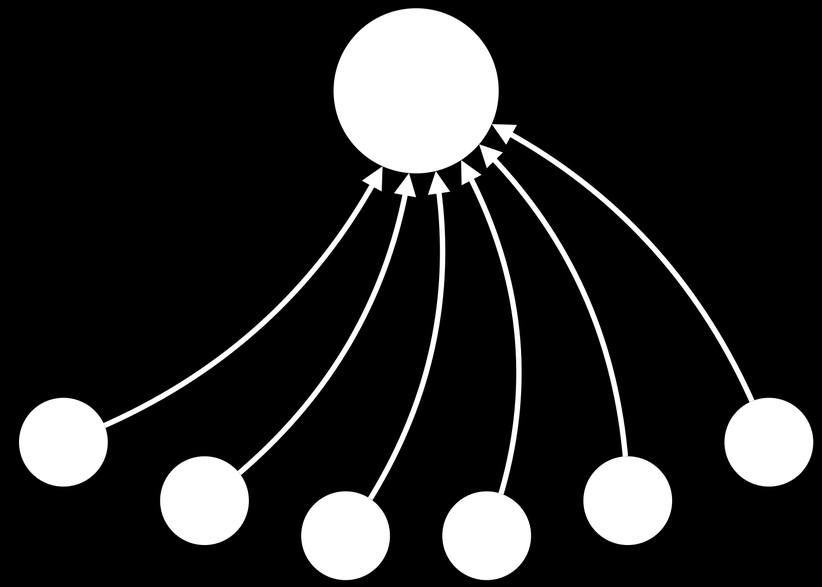
Addendum: Votes from important pages count more



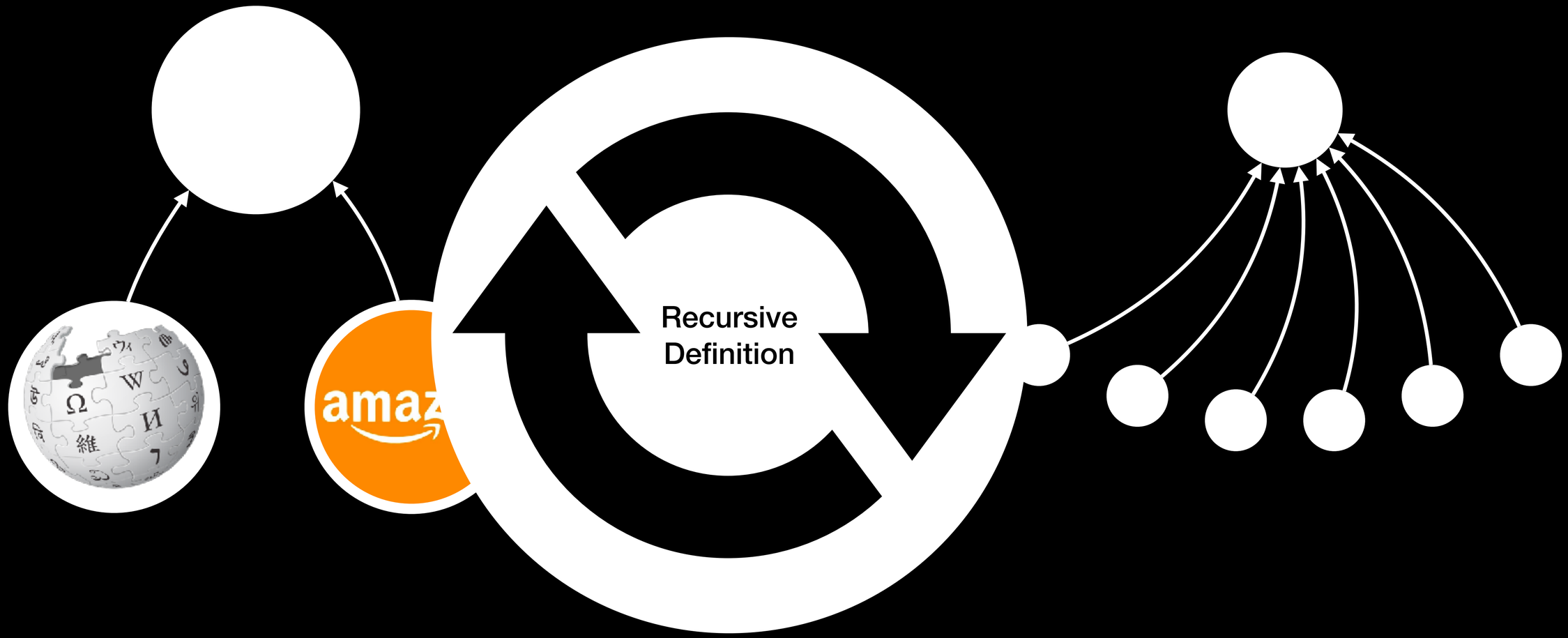


Addendum: Votes from important pages count more





Addendum: Votes from important pages count more



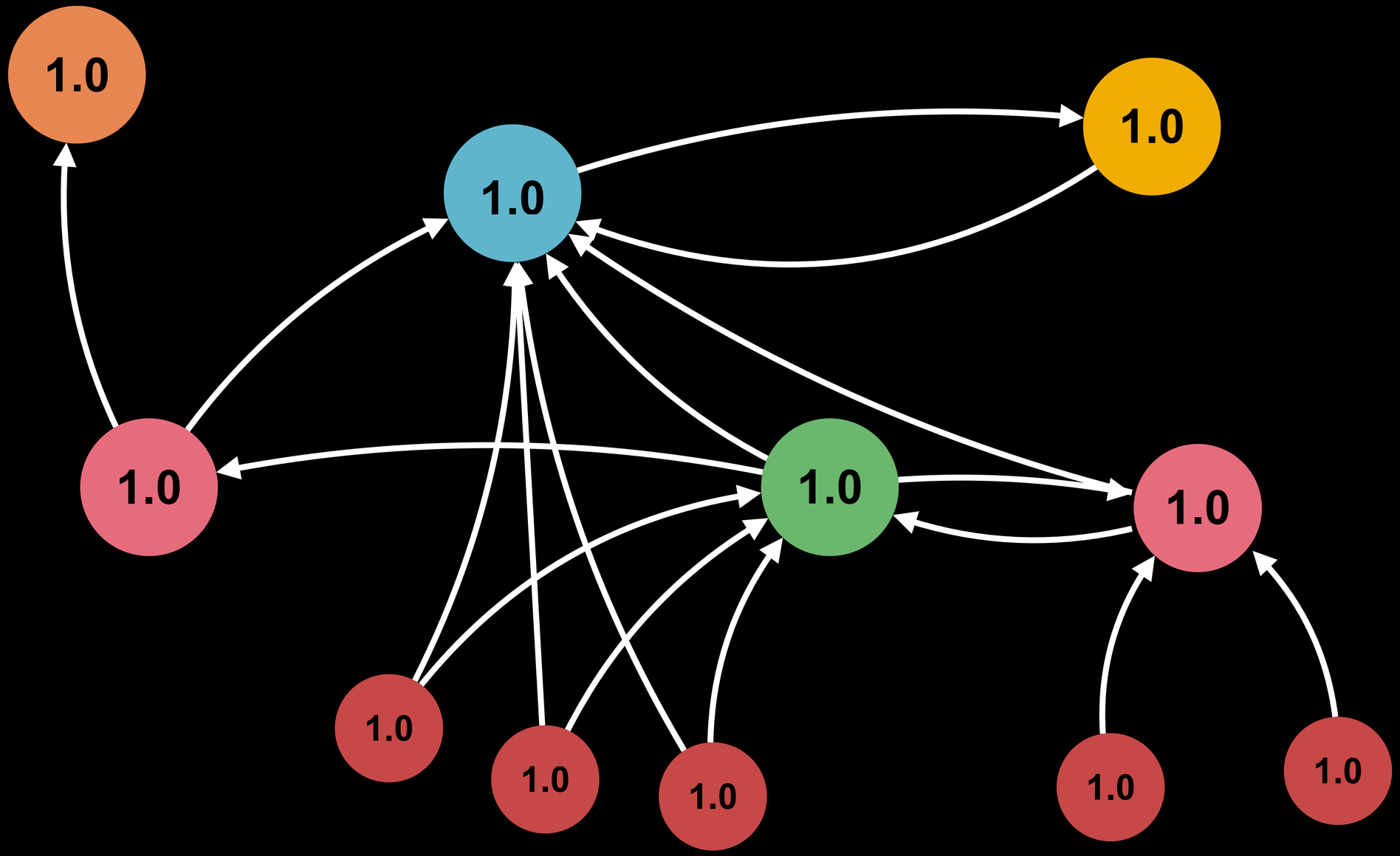
PageRank's Recursive Definition

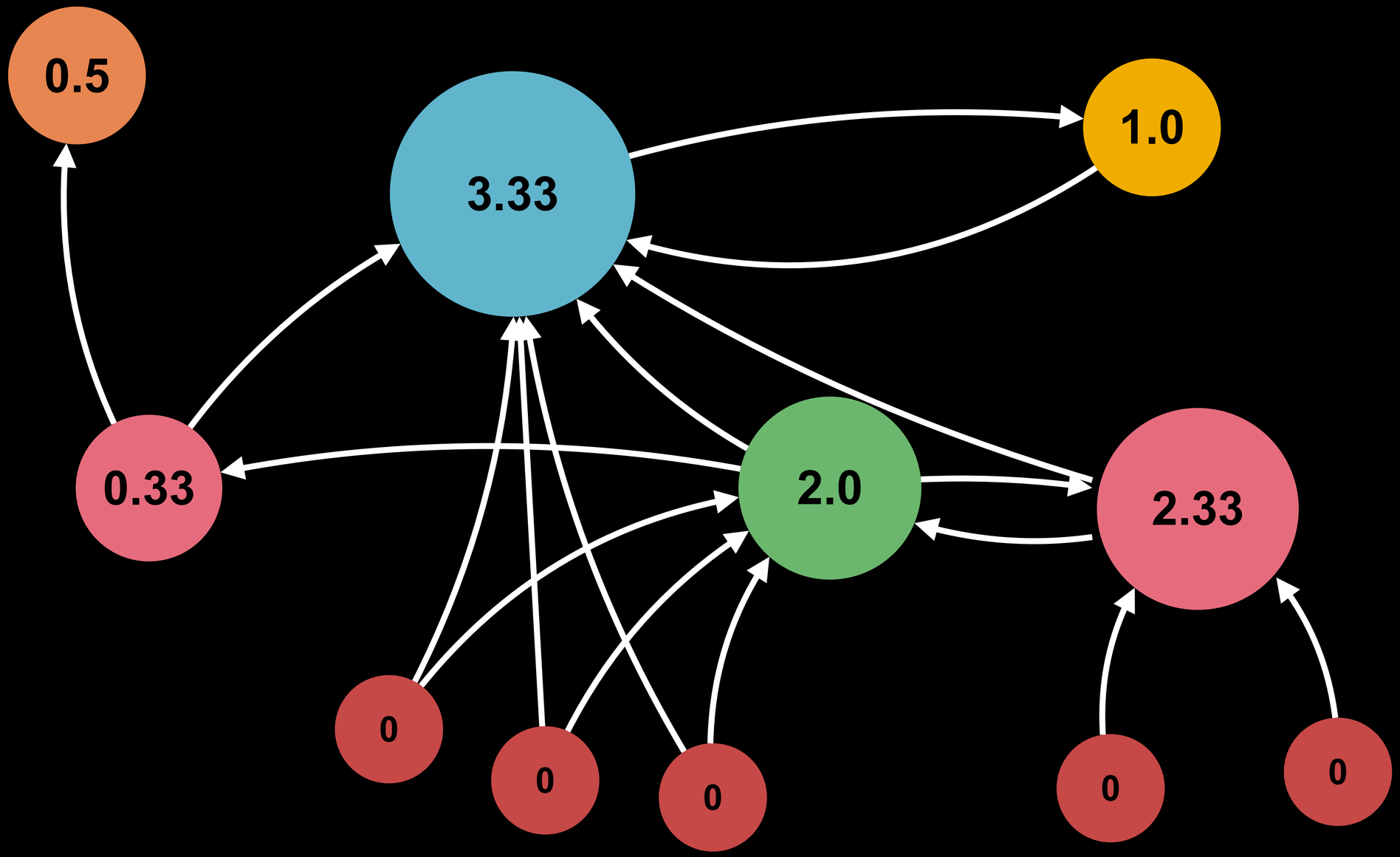
$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d_i}}$$

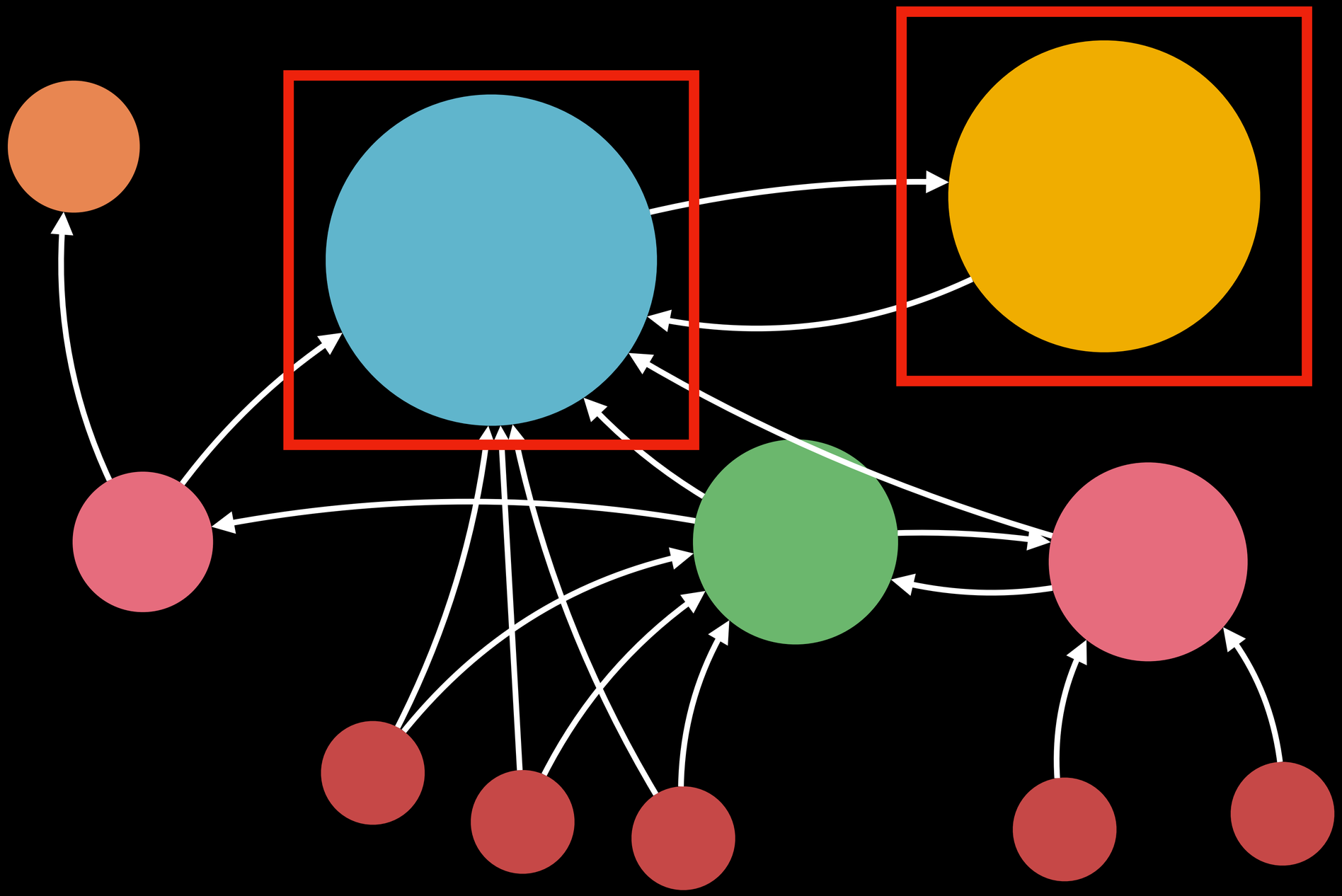
PageRank score r_j for page j

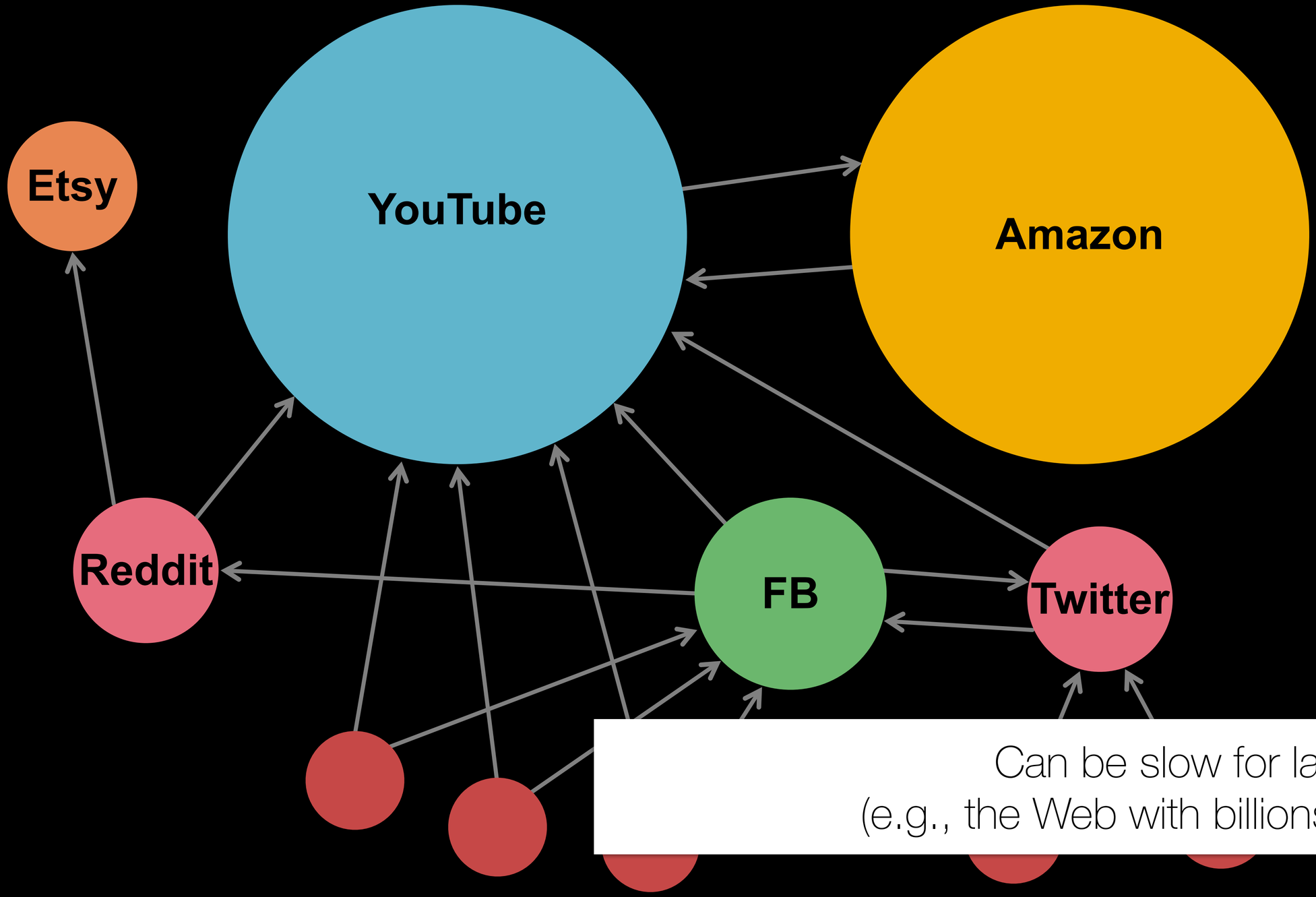
A node's rank can be distributed across its d_j outgoing links

Can be calculated in a distributed fashion via message-passing


1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0


0.5
3.33
1.0
0.33
2.0
2.33
0
0
0
0
0





Can be slow for large graphs
(e.g., the Web with billions of pages)

What insights might this algorithm provide in other network contexts?

This Lecture's Learning Objectives

Construct graphs from web structures

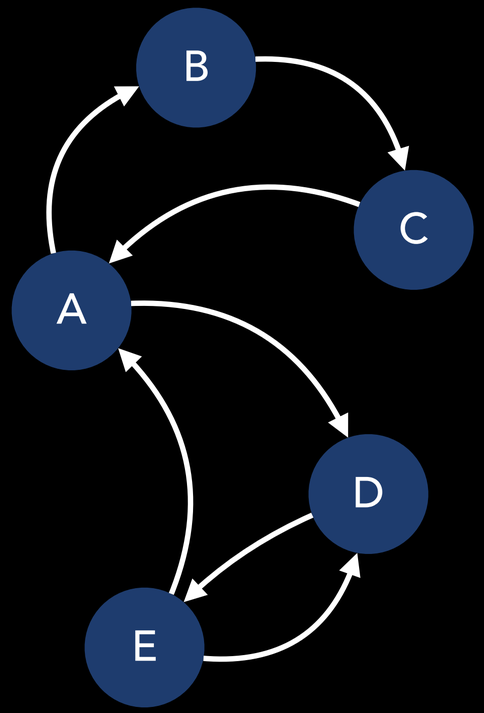
Describe at least two ways how one may try to exploit search engines

Explain how PageRank captures "importance" in a graph structure

Construct a transition matrix to model a collection of web pages

Differentiate hubs and authorities

An Alternative PageRank Definition

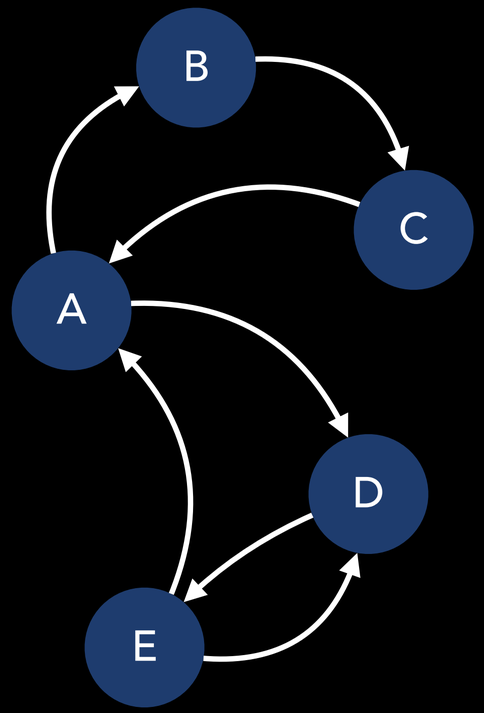


Adjacency Matrix *A* =

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

Let M be a stochastic adjacency matrix, $M_{ij} = 1/d_i$

An Alternative PageRank Definition



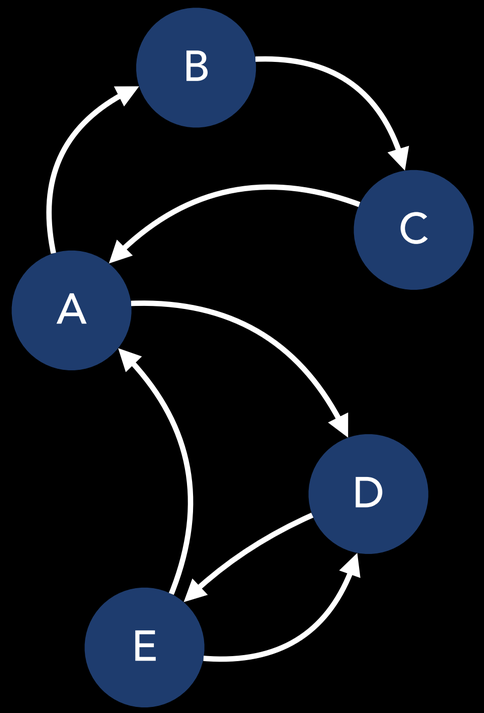
Transition Matrix *M* =

	A	B	C	D	E
A	0	0	0	0	1/2
B	1/2	0	1	0	0
C	0	1	0	0	0
D	1/2	0	0	0	1/2
E	0	0	0	1	0

	1	1	1	1	1

Let *M* be a stochastic adjacency matrix, $M_{ij} = 1/d_i$

An Alternative PageRank Definition



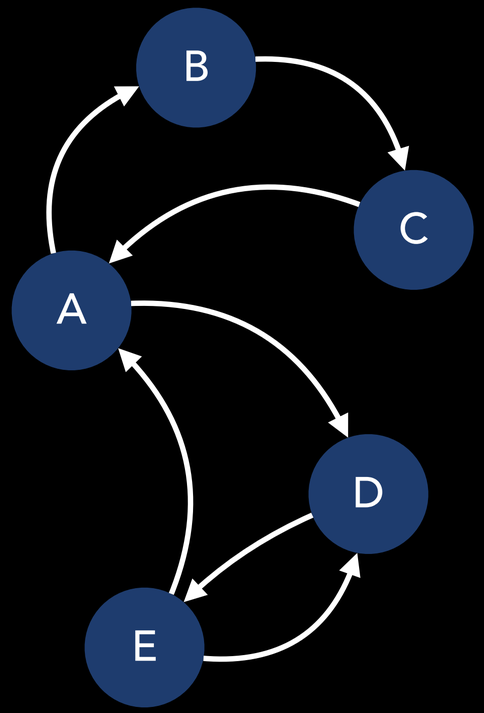
Transition Matrix *M* =

	A	B	C	D	E
A	0	0	0	0	1/2
B	1/2	0	1	0	0
C	0	1	0	0	0
D	1/2	0	0	0	1/2
E	0	0	0	1	0

Define *r* to be vector of PageRank scores

$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d_i}}$$

An Alternative PageRank Definition



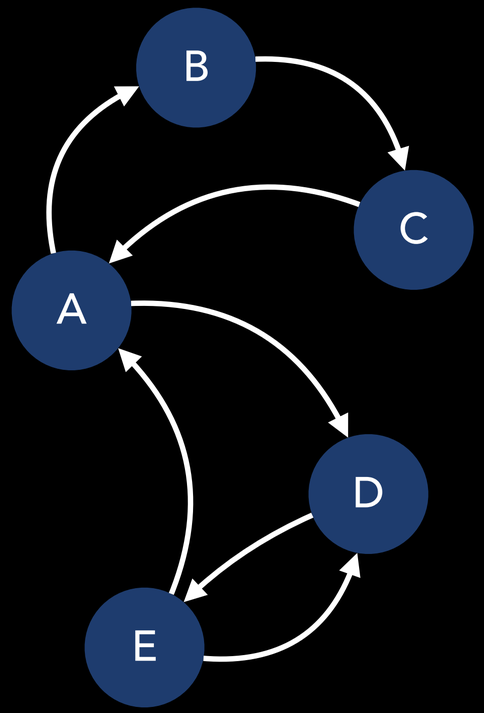
Transition Matrix *M* =

	A	B	C	D	E
A	0	0	0	0	1/2
B	1/2	0	1	0	0
C	0	1	0	0	0
D	1/2	0	0	0	1/2
E	0	0	0	1	0

Define *r* to be vector of PageRank scores

r(ank) = M r

An Alternative PageRank Definition



Transition Matrix *M* =

	A	B	C	D	E
A	0	0	0	0	1/2
B	1/2	0	1	0	0
C	0	1	0	0	0
D	1/2	0	0	0	1/2
E	0	0	0	1	0

Vector *r* is an eigenvector of the transition matrix *M*

$$r(ank) = M\,r$$

$$\lambda r = M\,r$$

An Alternative PageRank Definition

PageRank scores then are given in the largest eigenvector of *M*

Can be found via Power Iteration

$r^{(0)} = [1/N, \ldots, 1/N]^T$

$r^{(t+1)} = M\, r^t$

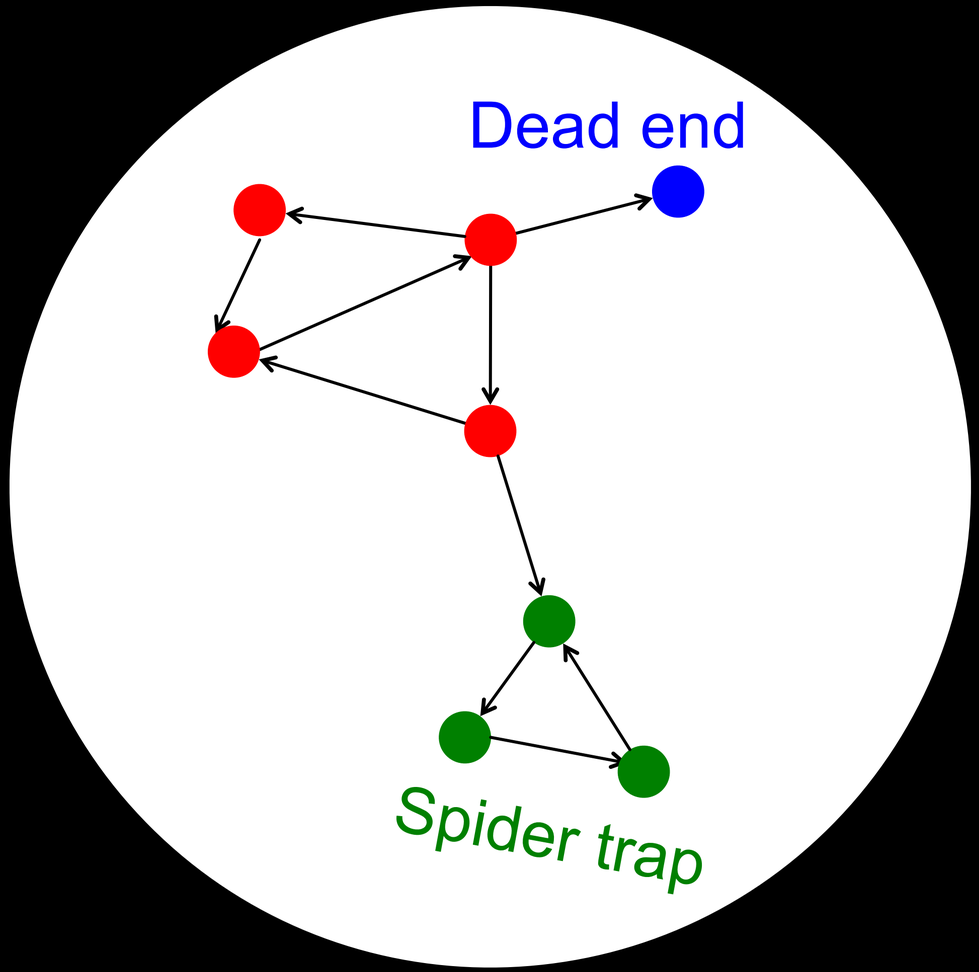
Transition Matrix *M* =

	A	B	C	D	E
A	0	0	0	0	1/2
B	1/2	0	1	0	0
C	0	1	0	0	0
D	1/2	0	0	0	1/2
E	0	0	0	1	0

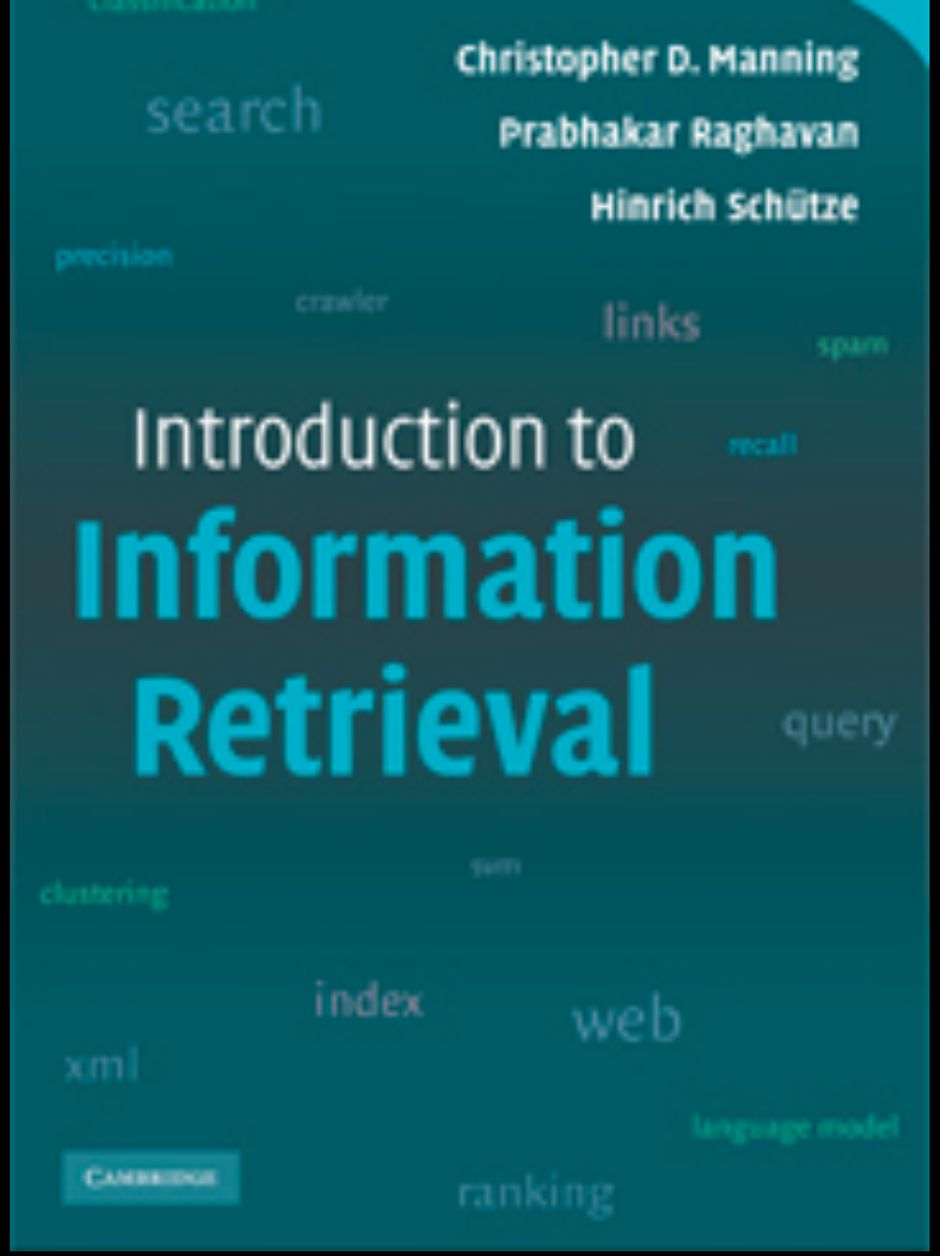
Vector *r* is an eigenvector of the transition matrix *M*

r(ank) = M r

This formulation works well for strongly connected graphs



Many extensions to PageRank to deal with these irregular structures



Query: Amanda Gorman

Amanda Gorman

From Wikipedia, the free encyclopedia

Amanda S. C. Gorman[1] (born 1998) is an American poet and activist. Her work focuses on issues of oppression, feminism, race, and marginalization, as well as the African diaspora. Gorman was the first person to be named National Youth Poet Laureate. She published the poetry book *The One for Whom Food Is Not Enough* in 2015. In 2021, she delivered her poem "The Hill We Climb" at the inaugu U.S. President Joe Biden. Her inaugu poem generated international acclaim stimulated her two books to reach best

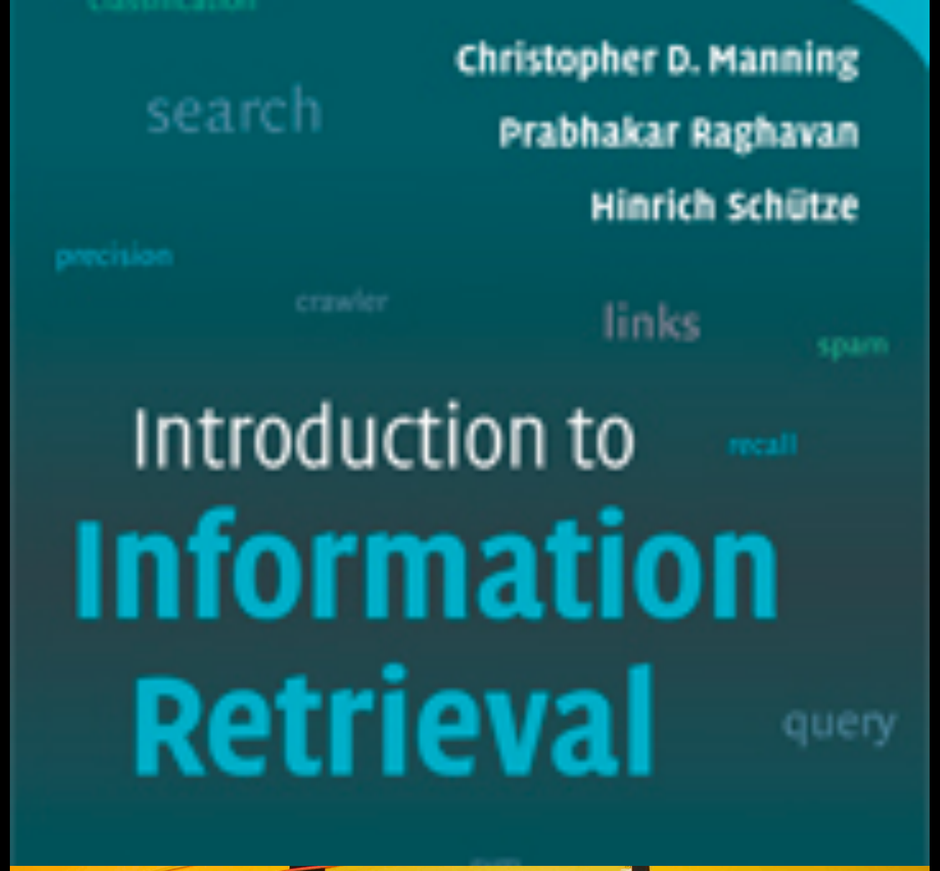
Wordsmith. Change-maker.

Amanda Gorman is the youngest inaugural poet in U.S. history, as well as an award-winning writer and *cum laude* graduate of Harvard University, where she studied Sociology. She has written for the *New York Times* and has three books forthcoming with Penguin Random House.

Born and raised in Los Angeles, she b writing at only a few years of age. No have won her invitations to the Obama

Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorman,
Amanda gorm
Amanda gorm

Query: Amanda Gorman



Largest *r*

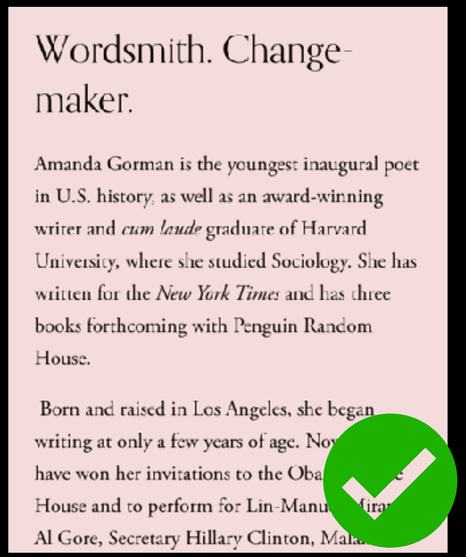
Next *r*

Small *r*

Amanda Gorman

From Wikipedia, the free encyclopedia

Amanda S. C. Gorman[1] (born 1998) is an American poet and activist. Her work focuses on issues of oppression, feminism, race, and marginalization, as well as the African diaspora. Gorman was the first person to be named National Youth Poet Laureate. She published the poetry book *The One for Whom Food Is Not Enough* in 2015. In 2021, she delivered her poem "The Hill We Climb" at the inauguration of U.S. President Joe Biden. Her inauguration poem generated international acclaim stimulated her two books to reach b status, and earned her a professiona management contract.



Wordsmith. Change-maker.

Amanda Gorman is the youngest inaugural poet in U.S. history, as well as an award-winning writer and *cum laude* graduate of Harvard University, where she studied Sociology. She has written for the *New York Times* and has three books forthcoming with Penguin Random House.

Born and raised in Los Angeles, she began writing at only a few years of age. No have won her invitations to the Oba House and to perform for Lin-Manu Al Gore, Secretary Hillary Clinton, Mal

...

Open Problems with PageRank

Measures generic popularity

Only one measure of importance

Can be manipulated (though harder)

Open Problems with PageRank

Measures generic popularity

Only one measure of importance

Can be manipulated (though harder)

Open Problems with PageRank

Measures generic popularity

Only one measure of importance

Can be manipulated (though harder)

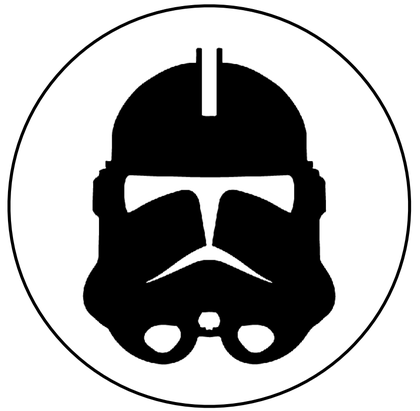
Open Problems with PageRank

Measures generic popularity

Only one measure of importance

MarketWatch

Can be manipulated (though harder)



Measures generic popularity

dmoz

User-specific history vector

User-specific history vector

User-specific history vector

Can be solved with topic-specific PageRank

Only one measure of importance

MarketWatch

Hubs: Aggregator of many authorities

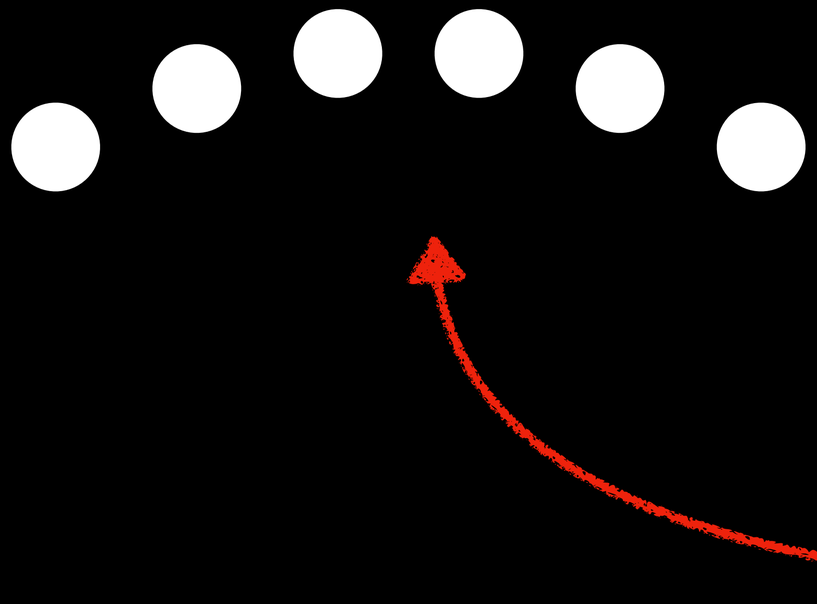
Authorities: Topic-specific source

HITS Algorithm separates the two types

Separate sources into **hubs** and **authorities**

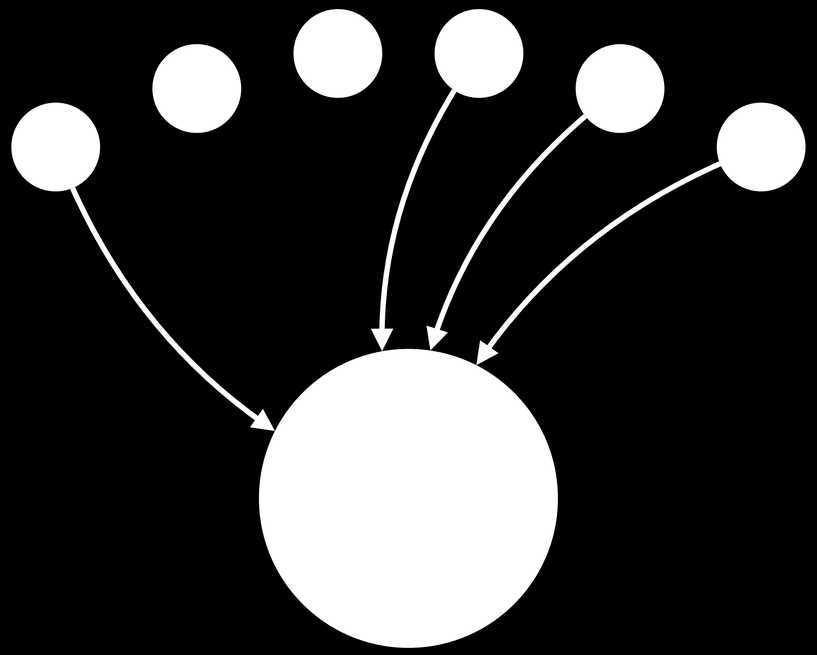
This Lecture's Learning Objectives

- Construct graphs from web structures
- Describe at least two ways how one may try to exploit search engines
- Explain how PageRank captures "importance" in a graph structure
- Construct a transition matrix to model a collection of web pages
- Differentiate hubs and authorities



Given a query *q*…

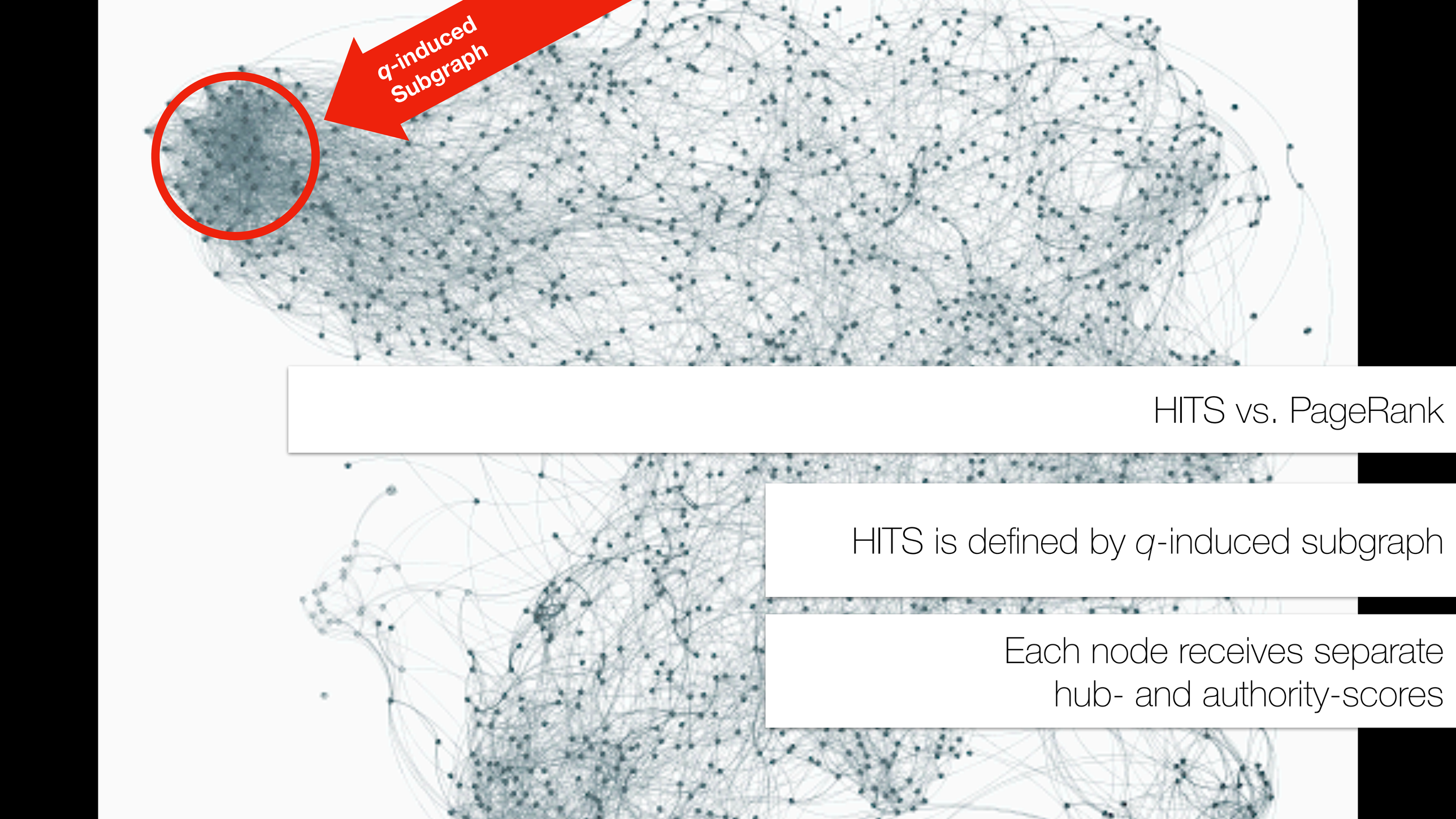
Find pages that include *q*



Many of these pages may link to a common site

This “common site” has high “authority”

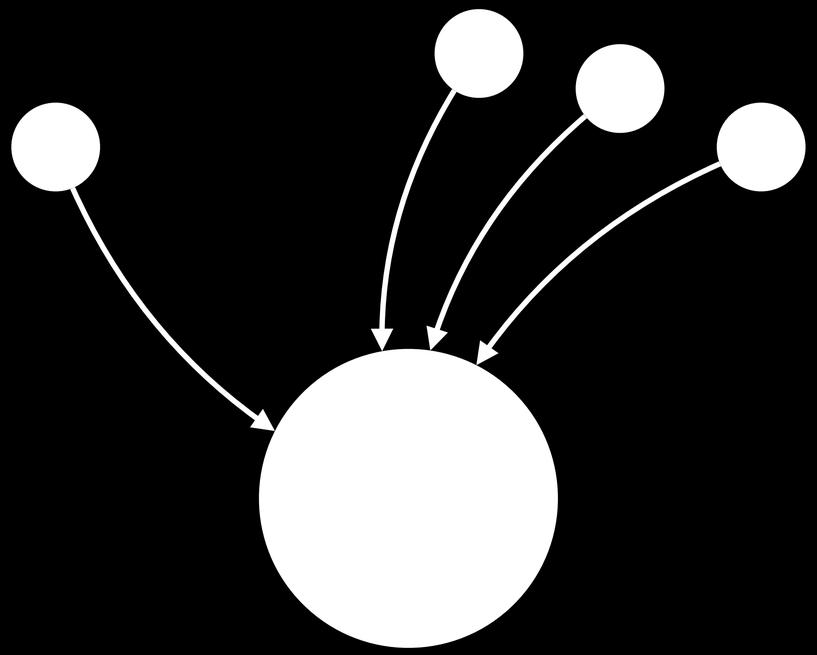
Pages that refer you to this common site are “hubs”



HITS vs. PageRank

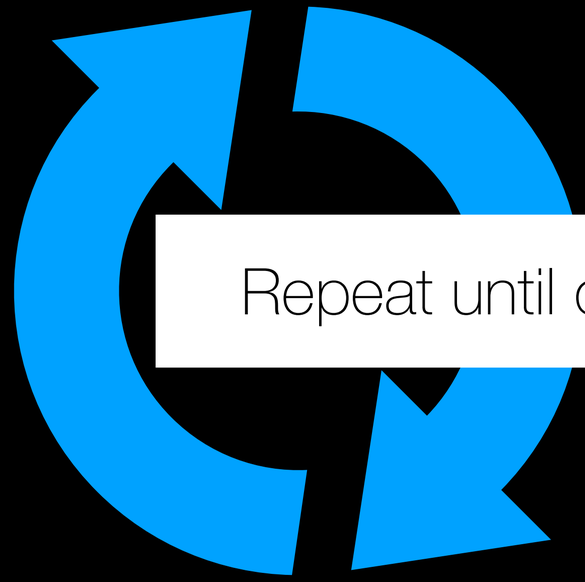
HITS is defined by *q*-induced subgraph

Each node receives separate
hub- and authority-scores



Hub Score: Sum of adjacent authority scores

Authority Score: Sum of adjacent hub scores



Repeat until convergence

This Lecture's Learning Objectives

- Construct graphs from web structures
- Describe at least two ways how one may try to exploit search engines
- Explain how PageRank captures "importance" in a graph structure
- Construct a transition matrix to model a collection of web pages
- Differentiate hubs and authorities

Questions?

Prof. Cody Buntain | @codybuntain | cbuntain@umd.edu
Director, Information Ecosystems Lab