

Extracting and Analyzing Graphs

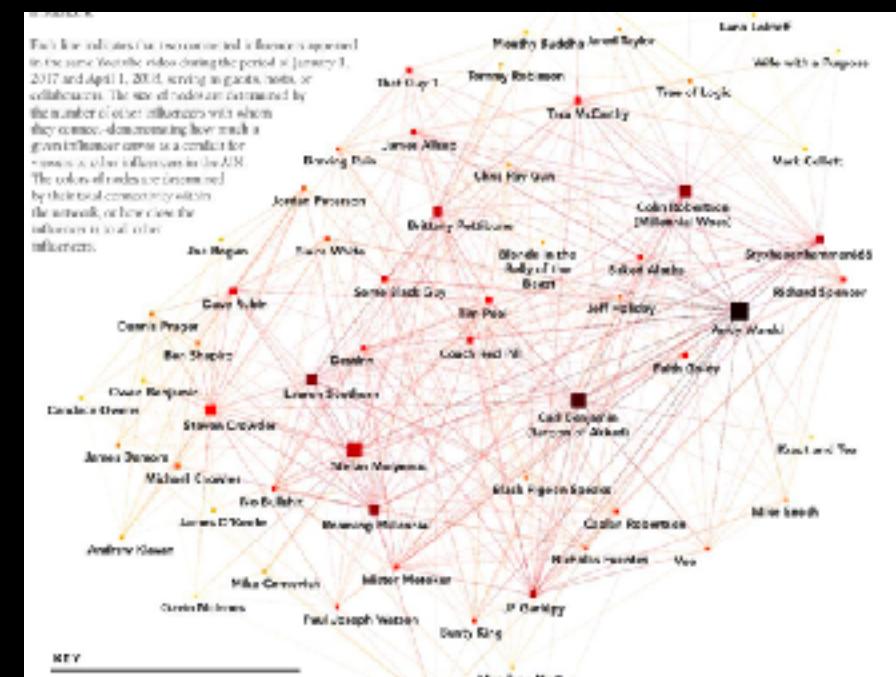
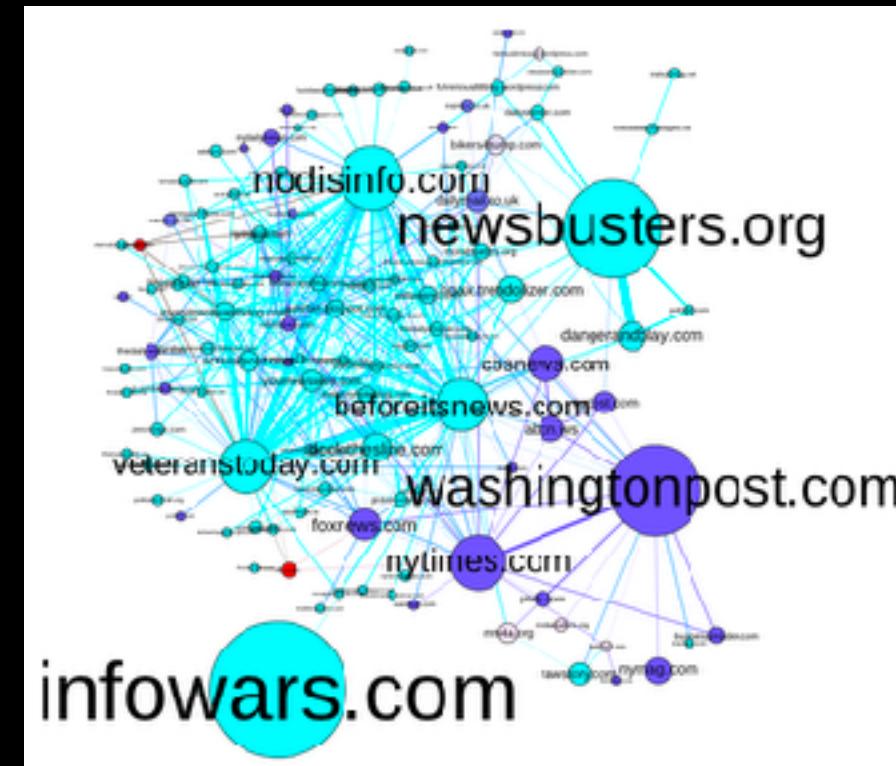
INST414 - Data Science Techniques

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?

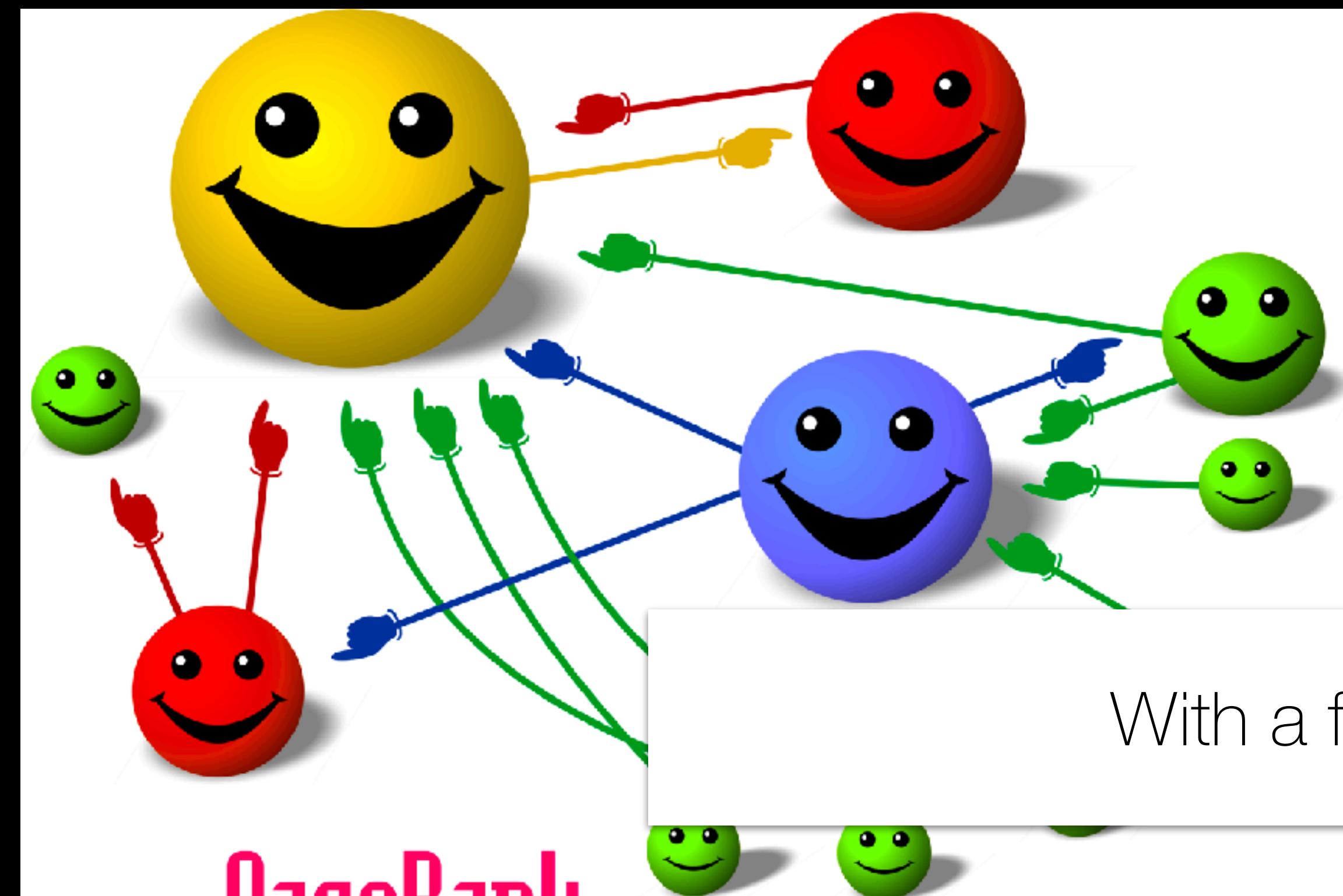
1. Collect and clean large-scale datasets



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

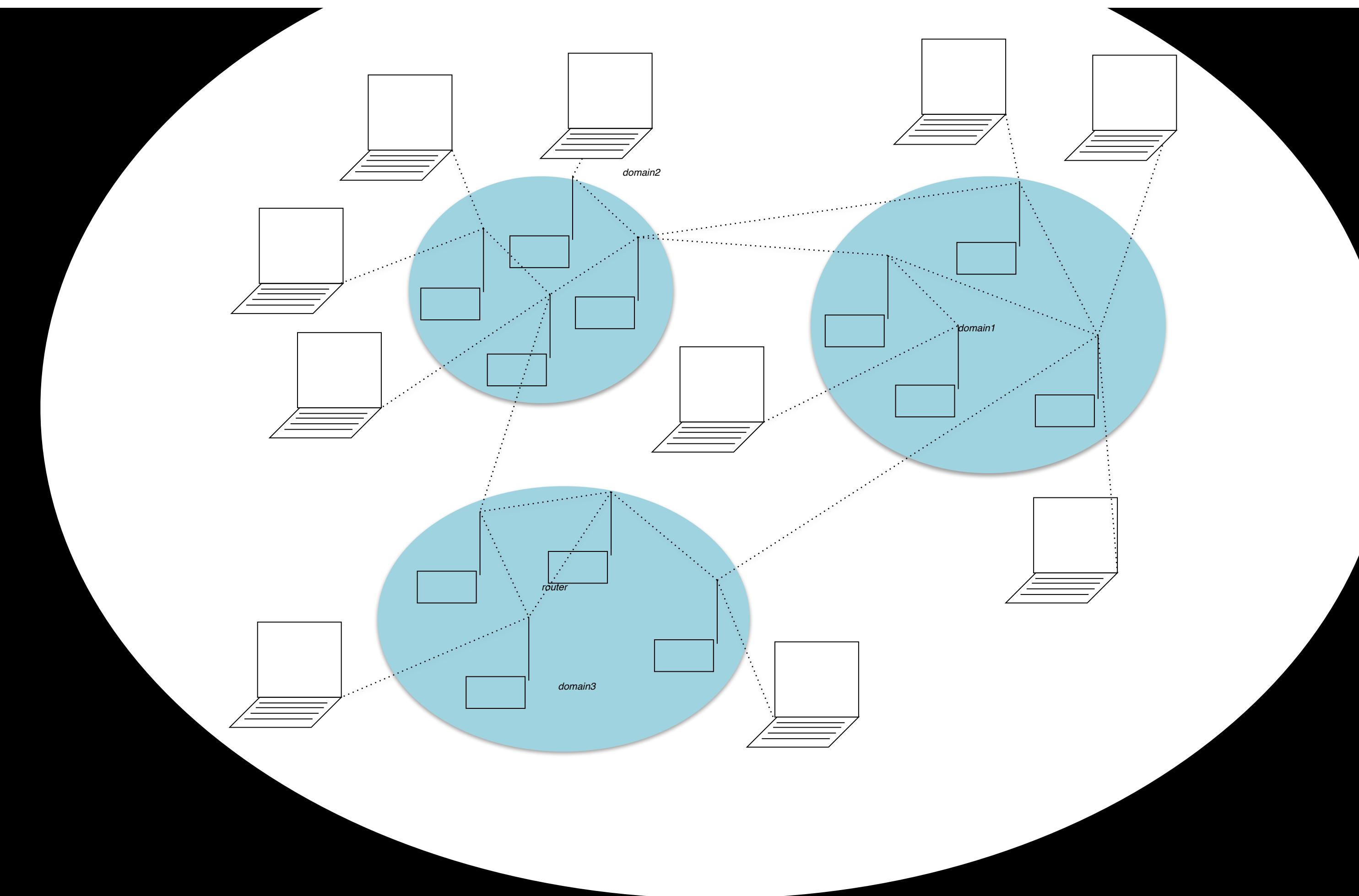
Construct a transition matrix to model a collection of web pages

Differentiate hubs and authorities

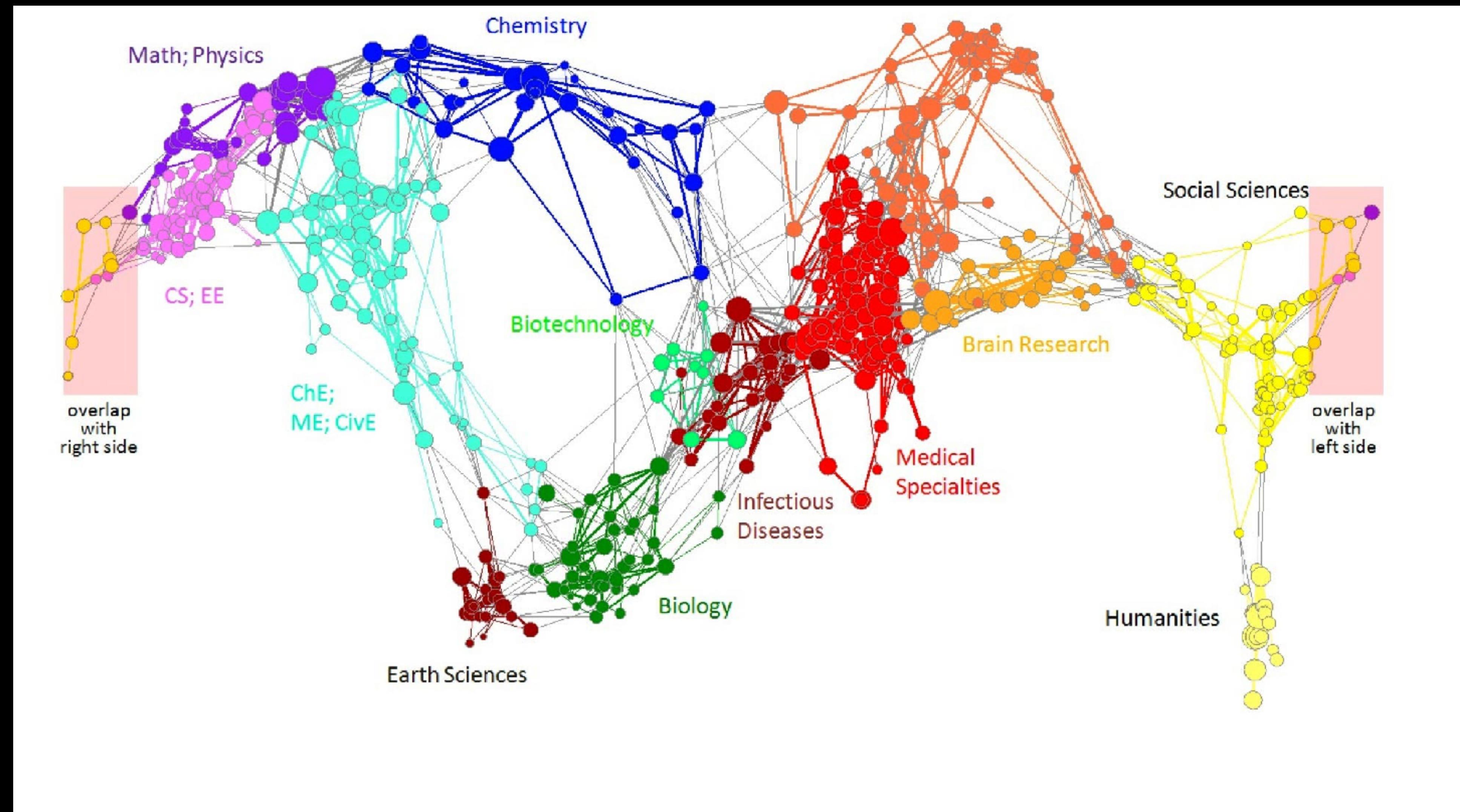
Web data is full of graphs/networks

What are some examples?

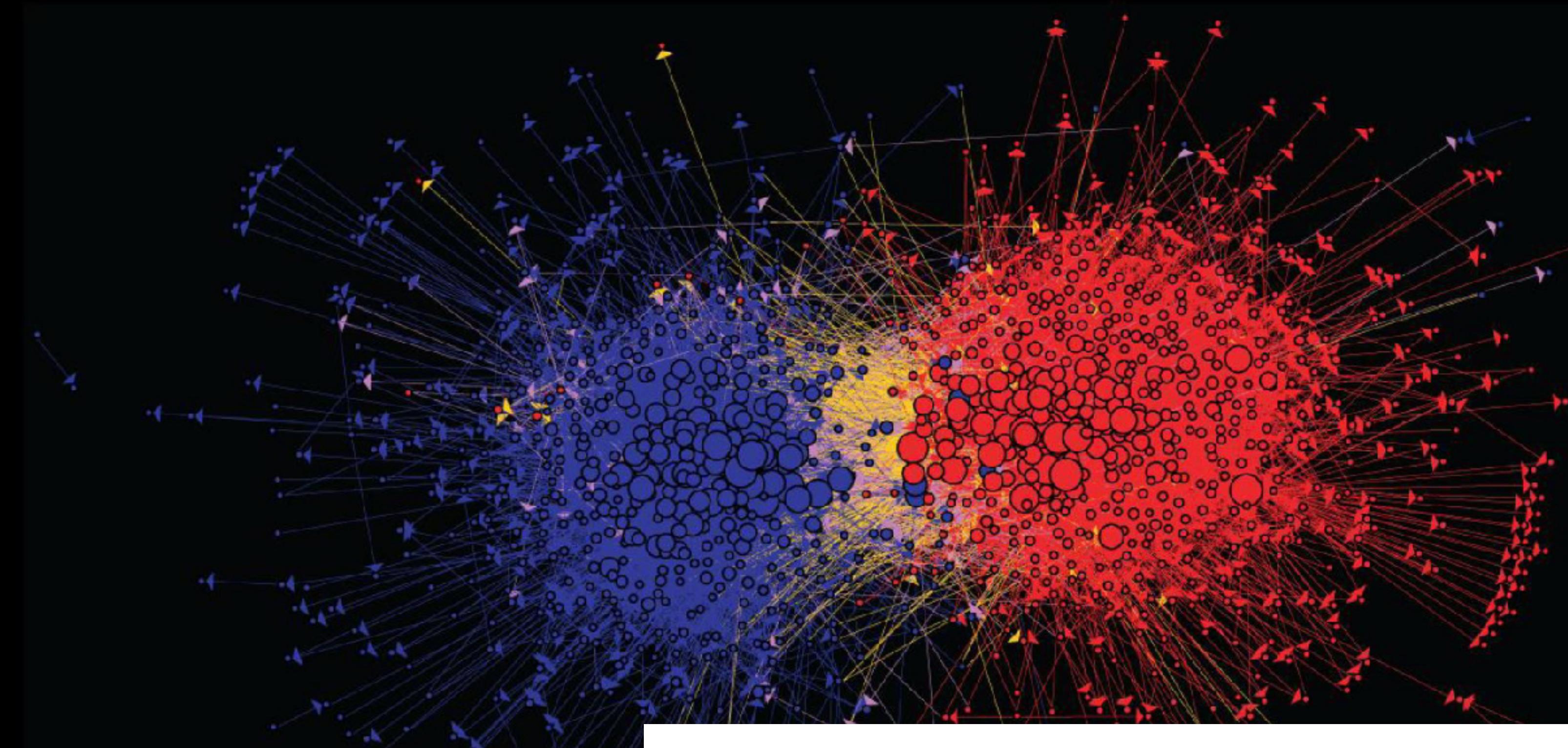
Graphs in the Internet



Graphs in Citations Networks



Graphs in Media Networks



Connections between political blogs

Graphs in Author Networks

SCIENTIFIC AMERICAN

Subscribe

SHARE LATEST

PUBLIC HEALTH | OPINION

How to Fix COVID Contact Tracing

Apps and human tracers both have pros and cons. To be effective, they have to work together

By Elissa Redmiles on December 7, 2020 اعرض هذا باللغة العربية

WIRED

BACKCHANNEL BUSINESS CULTURE MORE ▾ SIGN IN SUBSCRIBE

ELISSA M. REDMILES JOHN KRUMM IDEAS 07.08.2020 08:00 AM

Citizen Science Projects Offer a Model for Coronavirus Apps

Americans don't like when their data is taken—but research shows they would be willing to donate it.

Graphs in Social Media

The image displays two side-by-side screenshots of a Twitter profile for user **Cody Buntain** (@codybuntain). The left screenshot shows the 'Following' tab, while the right one shows the 'Followers' tab. Both screenshots include a sidebar with various icons.

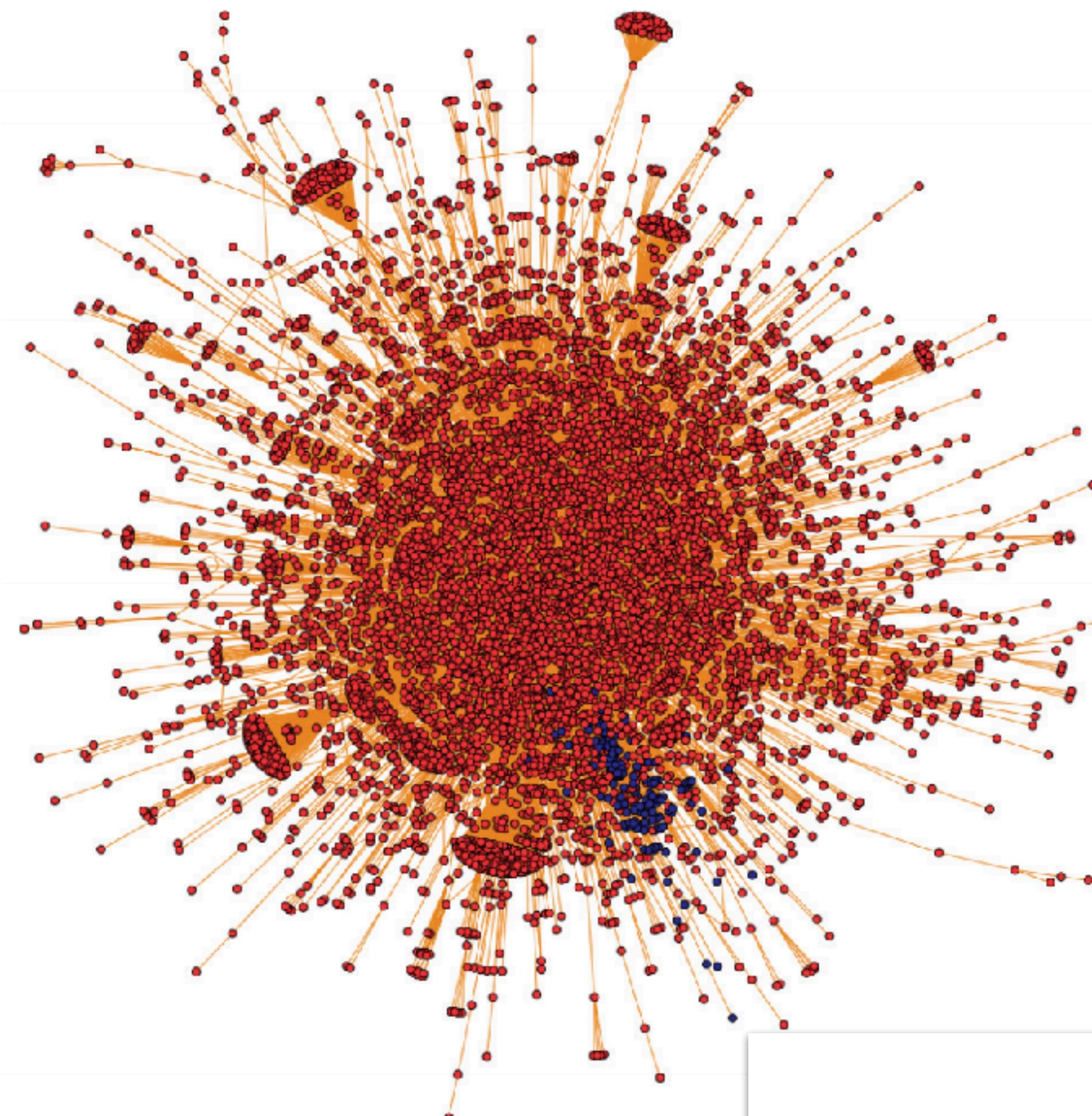
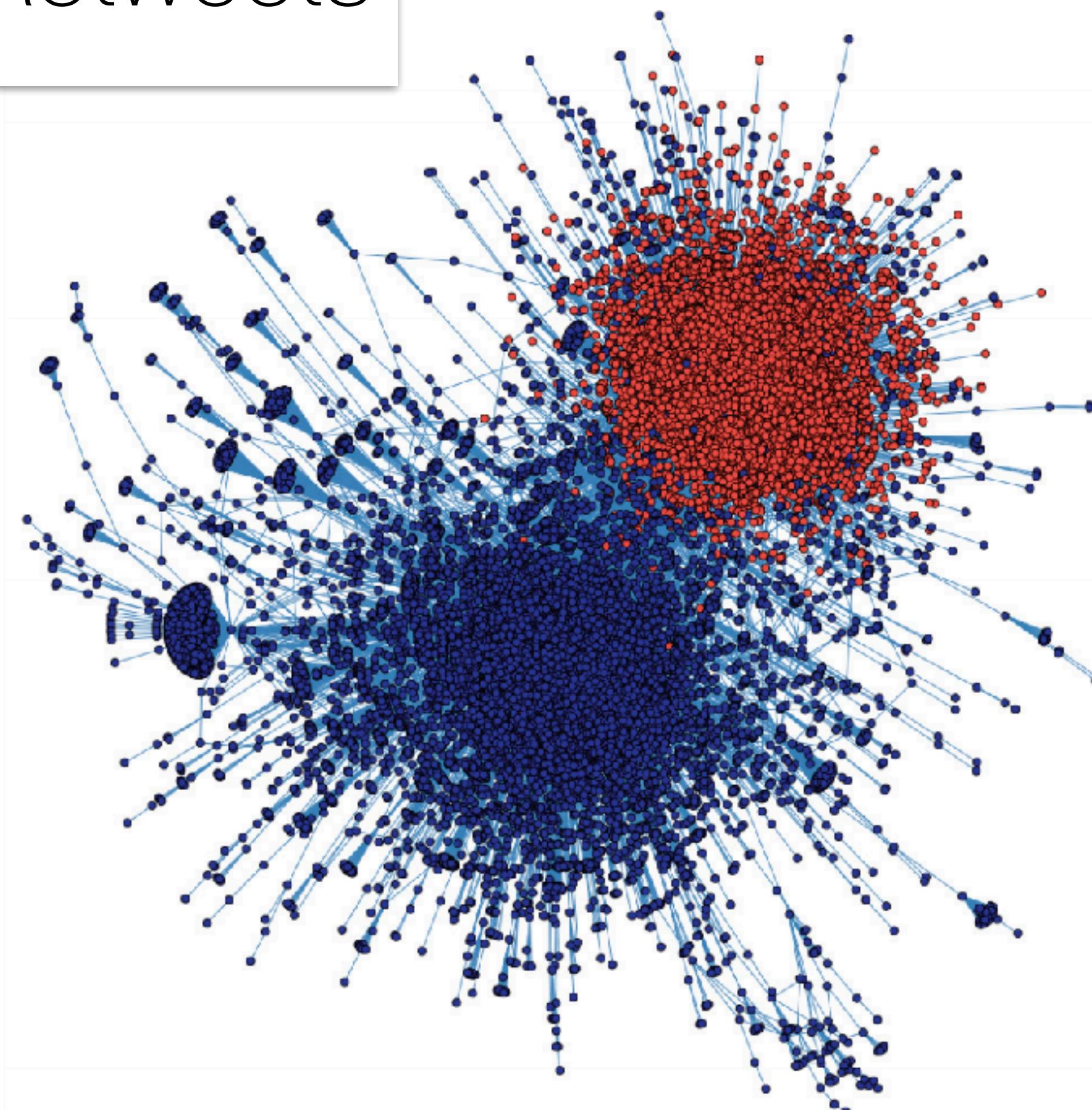
Following Tab (Left):

- Sujoy Kumar Sikdar** (@sujoyks) - Following
- Daniel Votipka** (@drvotipka) - Following
- Kaleigh Rogers** (@KaleighRogers) - Following
- Darren Linvill** (@DarrenLinvill) - Following
- Patrick Warren** (@plwarre) - Following

Followers Tab (Right):

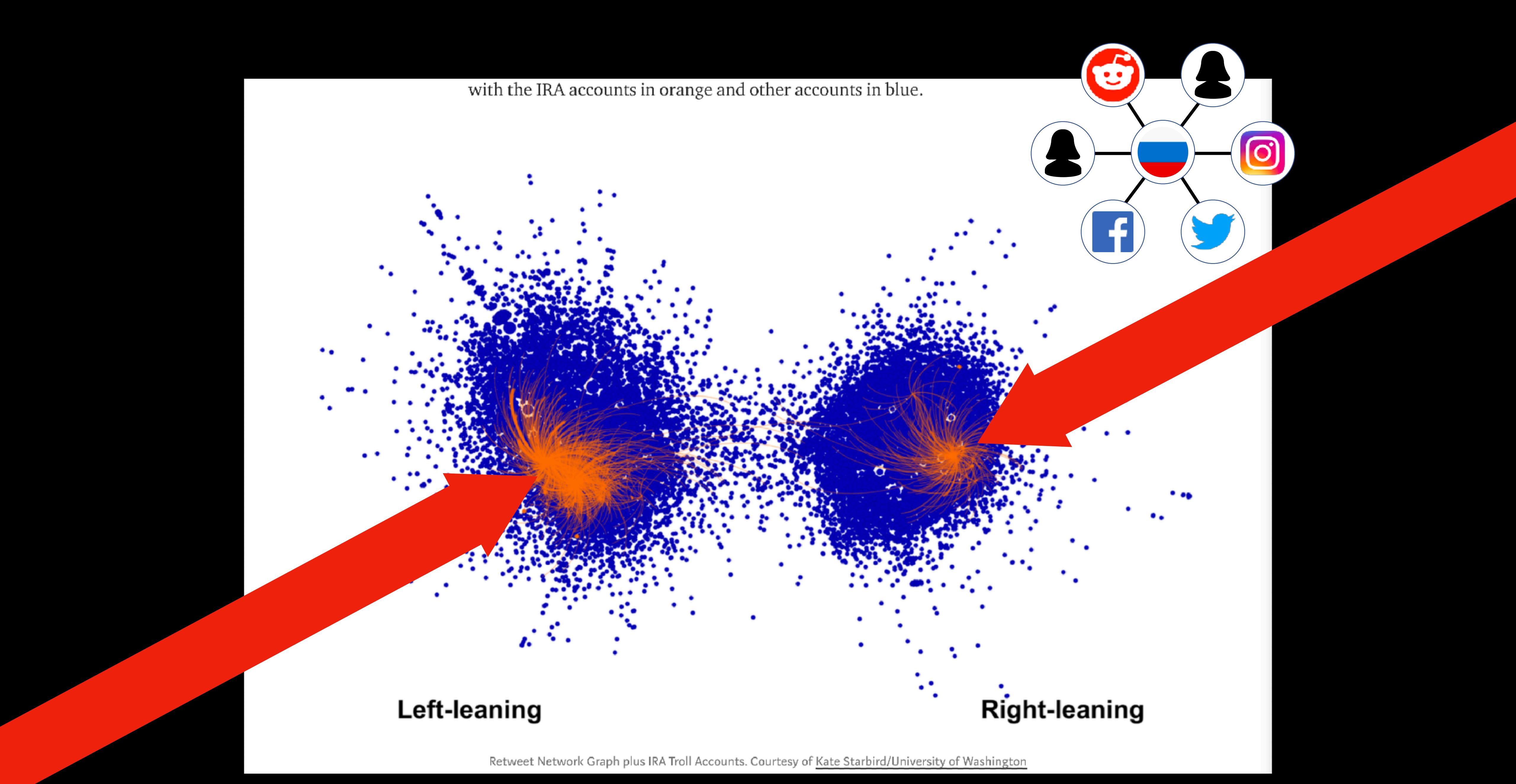
- Sarah Oates** (@media_politics) - Follows you
- Gregory Davis** (@GregoryDavisHNH) - Follows you
- Kaleigh Rogers** (@KaleighRogers) - Follows you
- Alvi Ishmam** (@Alvilshmam) - Follows you
- Don't Jump** (@iHypercube) - Follows you
- Marc Faddoul** (@MarcFaddoul) - Follows you

Retweets

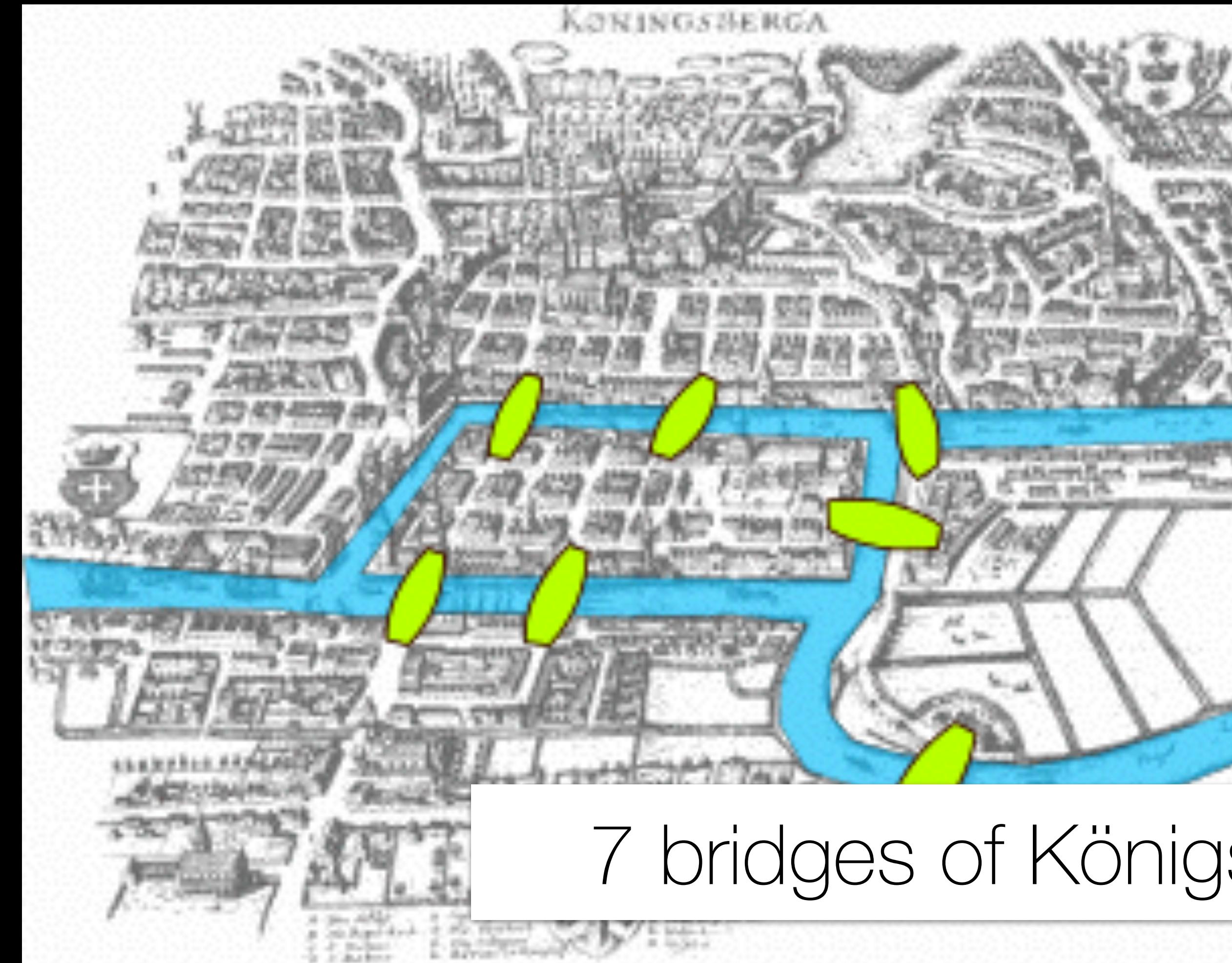


Mentions

Figure 1: The political retweet (left) and mention (right) networks, laid out using a force-directed algorithm with cluster assignments (see § 3.1). Community structure is evident in the retweet network, but less so in the mention network. We 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.



Graphs aren't new



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

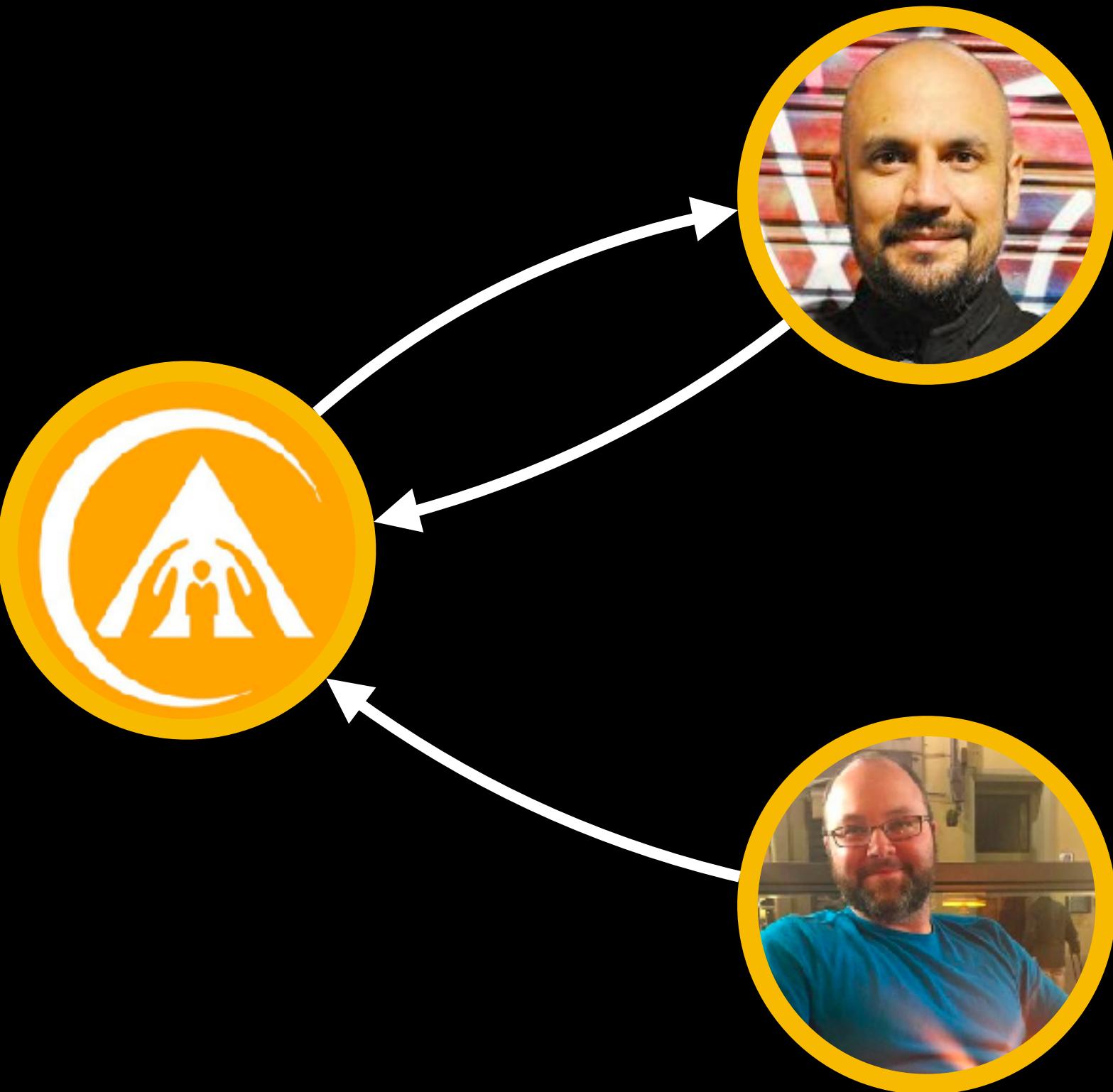
The image displays two side-by-side screenshots of a Twitter profile for user **Cody Buntain** (@codybuntain). The left screenshot shows the 'Following' tab, while the right one shows the 'Followers' tab. Both screenshots include a sidebar with various icons.

Following Tab (Left):

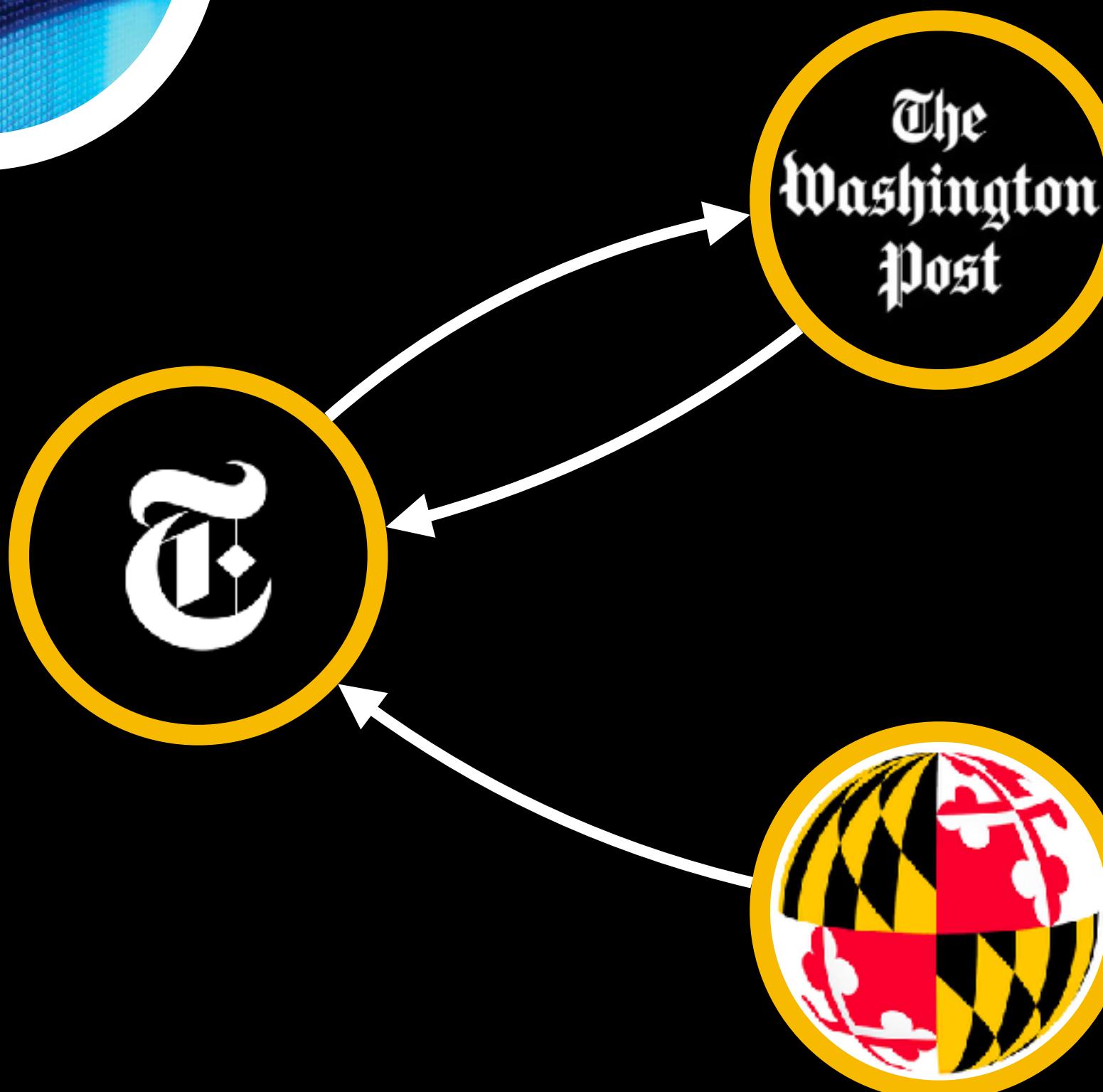
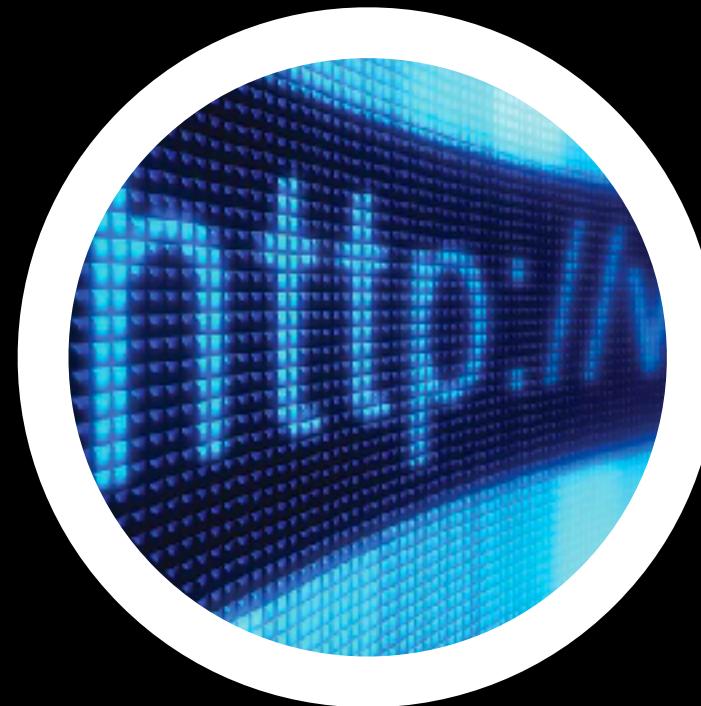
- Sujoy Kumar Sikdar** (@sujoyks) - Following
- Daniel Votipka** (@drvotipka) - Following
- Kaleigh Rogers** (@KaleighRogers) - Following
- Darren Linvill** (@DarrenLinvill) - Following
- Patrick Warren** (@plwarre) - Following

Followers Tab (Right):

- Sarah Oates** (@media_politics) - Follows you
- Gregory Davis** (@GregoryDavisHNH) - Follows you
- Kaleigh Rogers** (@KaleighRogers) - Follows you
- Alvi Ishmam** (@Alvilshmam) - Follows you
- Don't Jump** (@iHypercube) - Follows you
- Marc Faddoul** (@MarcFaddoul) - Follows you



Directed Graph



Washington Post, Breaking News

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

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.



Cody Buntain

Edit Profile



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

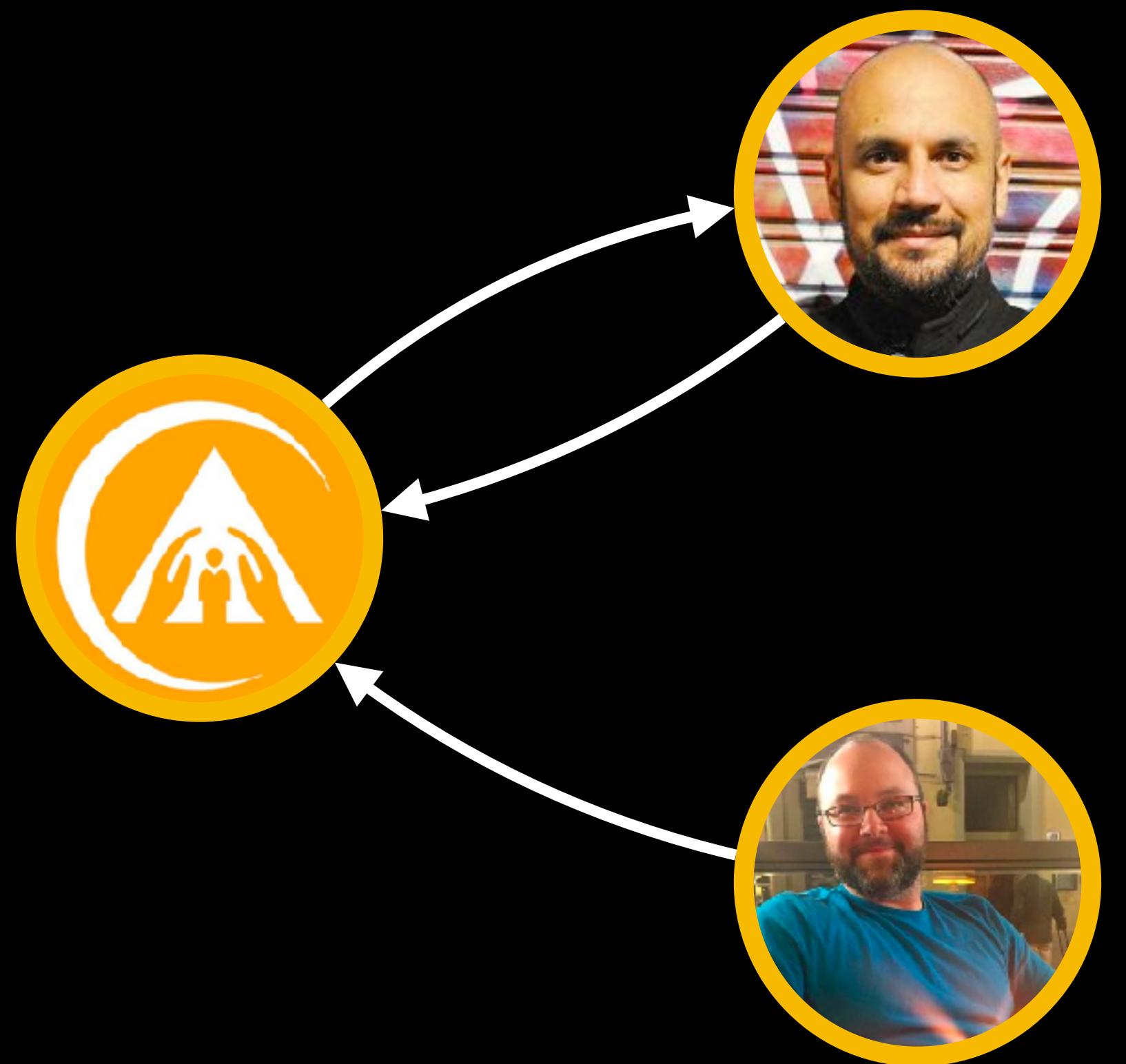


Christine Moulder

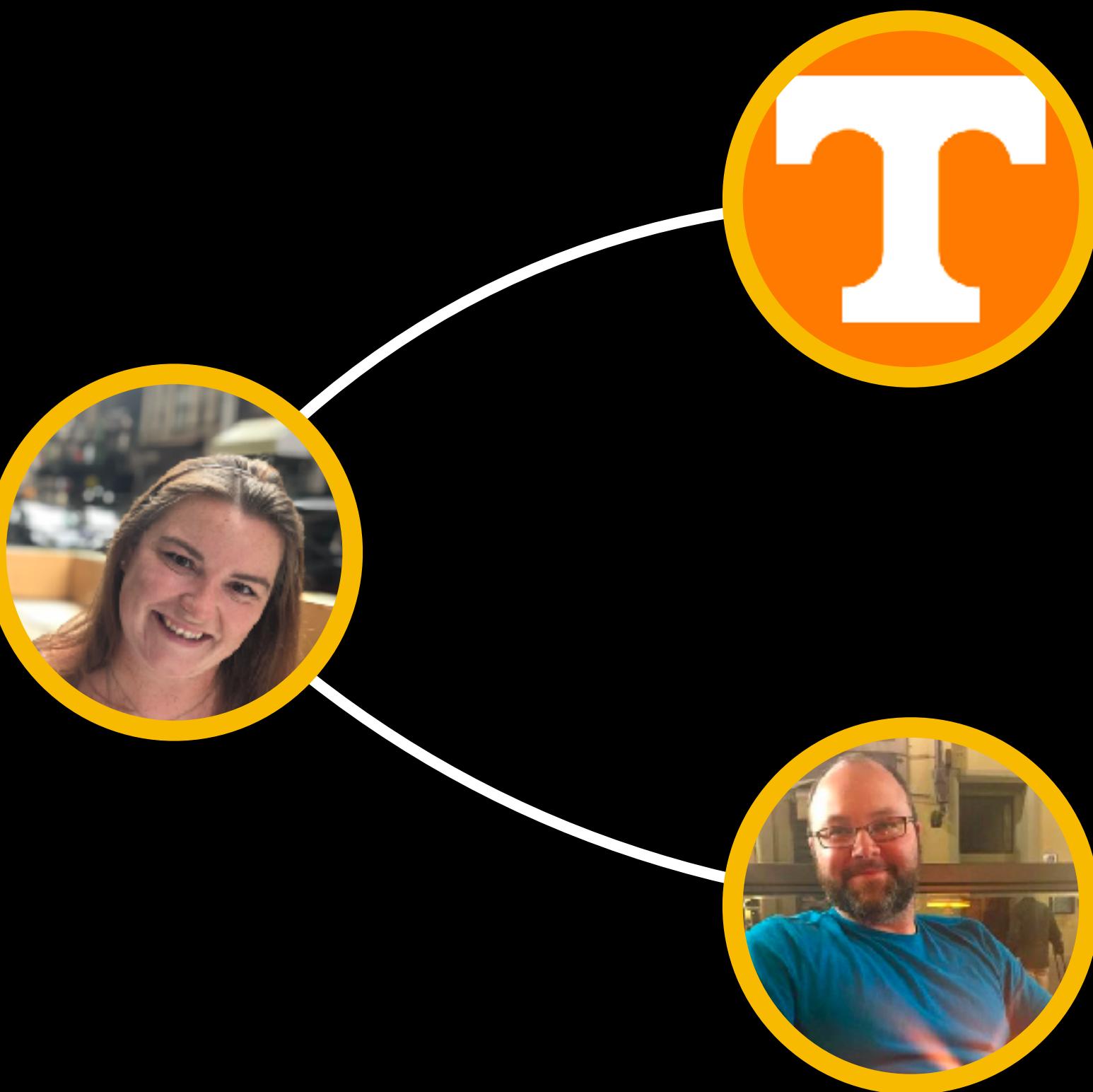
78 mutual friends

Friends





Directed Graph



Undirected Graph



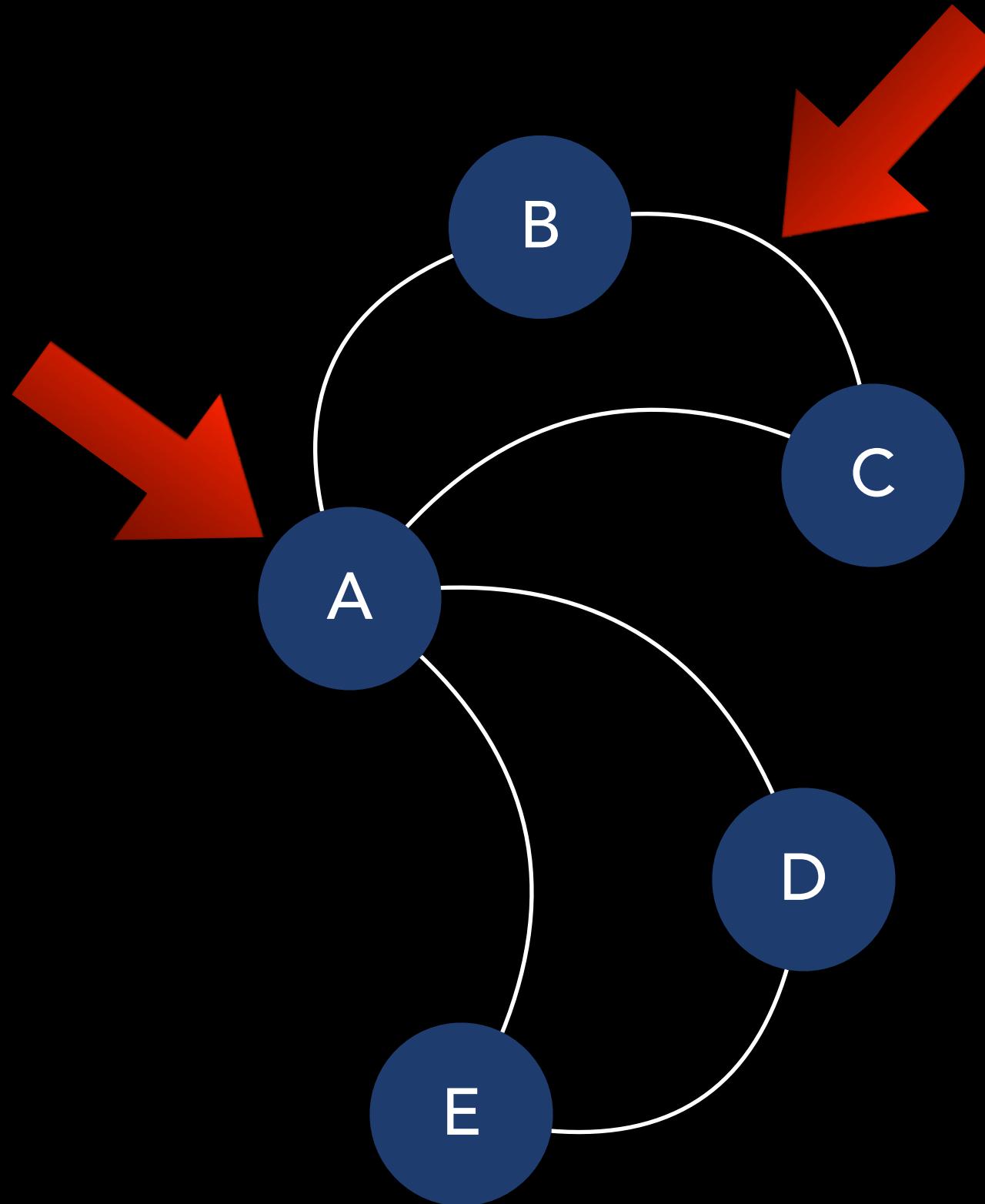
Directed
Graph

Not Reciprocal

Undirected Graph

WHAT IS A NETWORK?

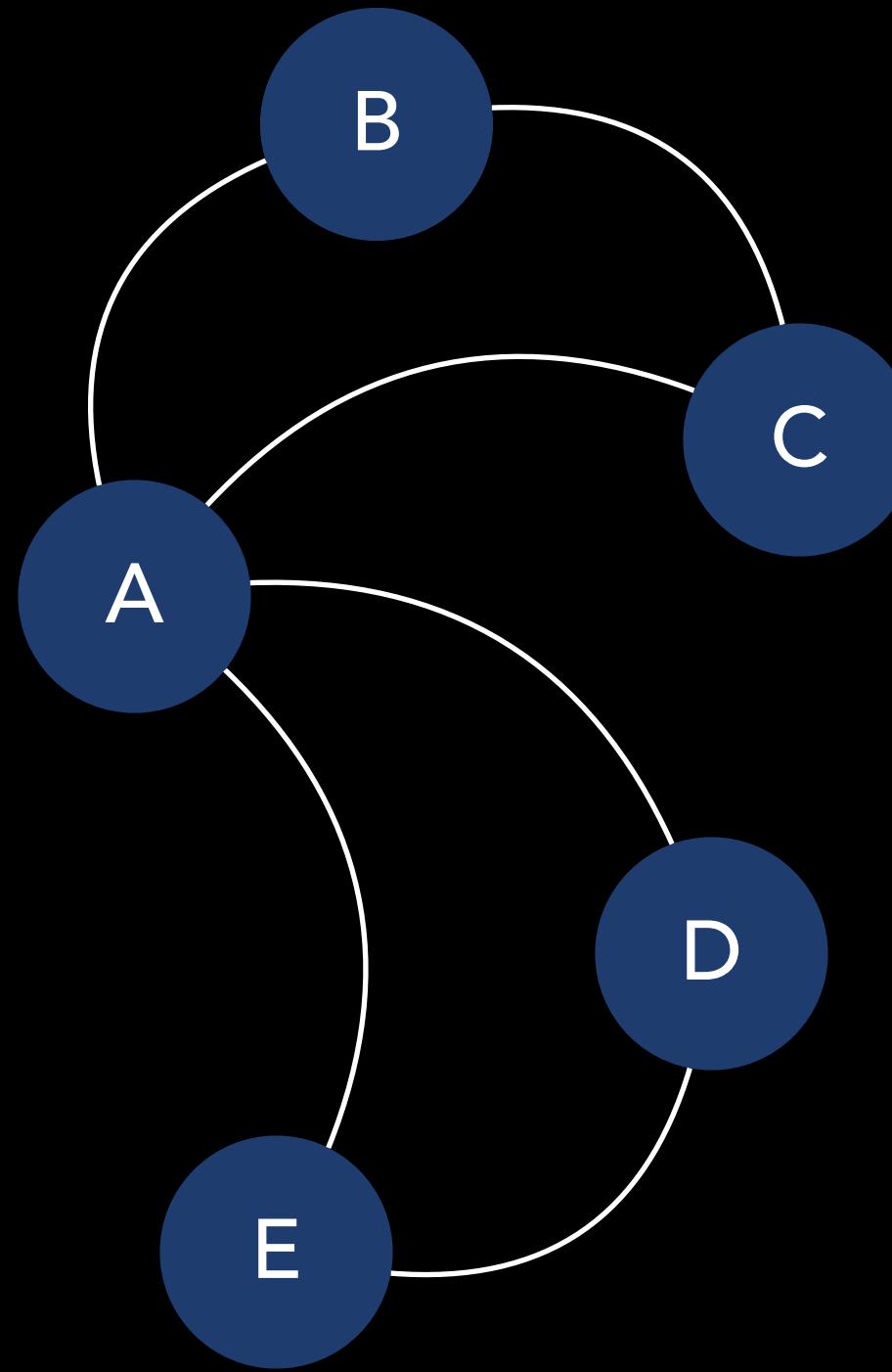
- A “network” or “graph” is:
 - A collection of nodes and edges
 - Nodes - objects or entities
 - Edges - connections between nodes



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

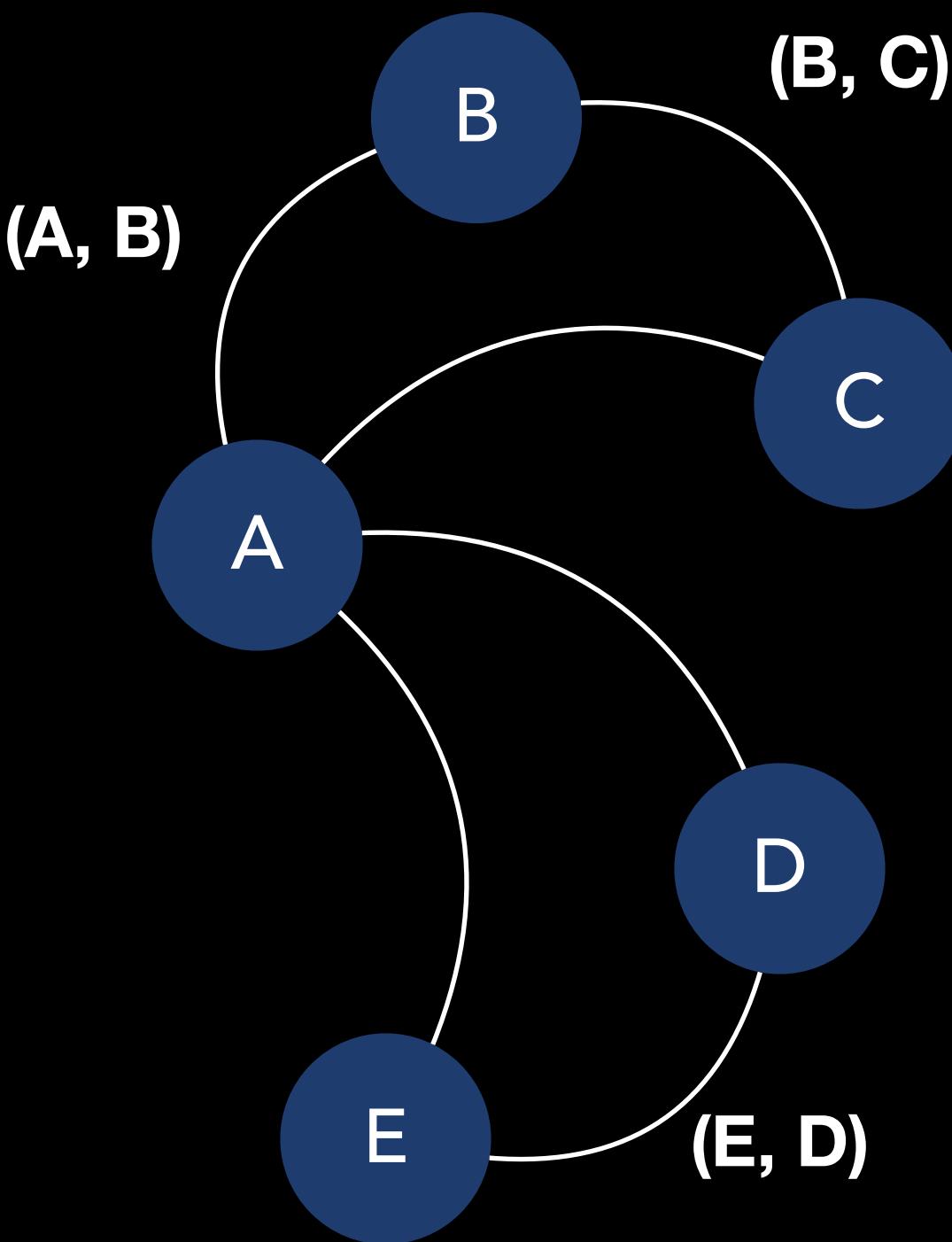
WHAT IS A NETWORK?

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



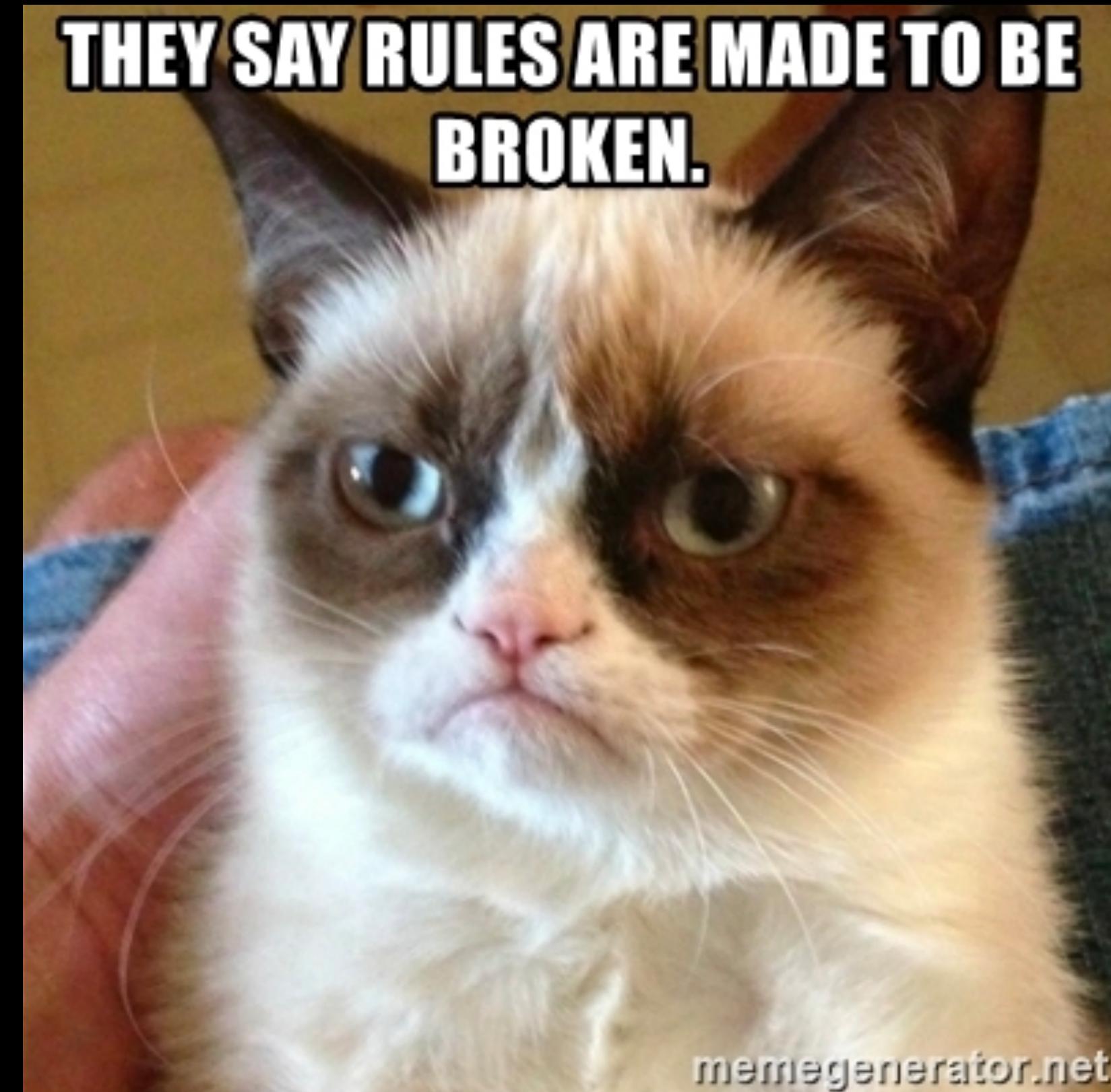
WHAT IS A NETWORK?

- Mathematical Definition:
 - $G(V, E)$
 - V – Vertices (i.e., nodes)
 - E – Edges
 - Tuples of vertices (v_1, v_2)
 - $E \subset V \times V$



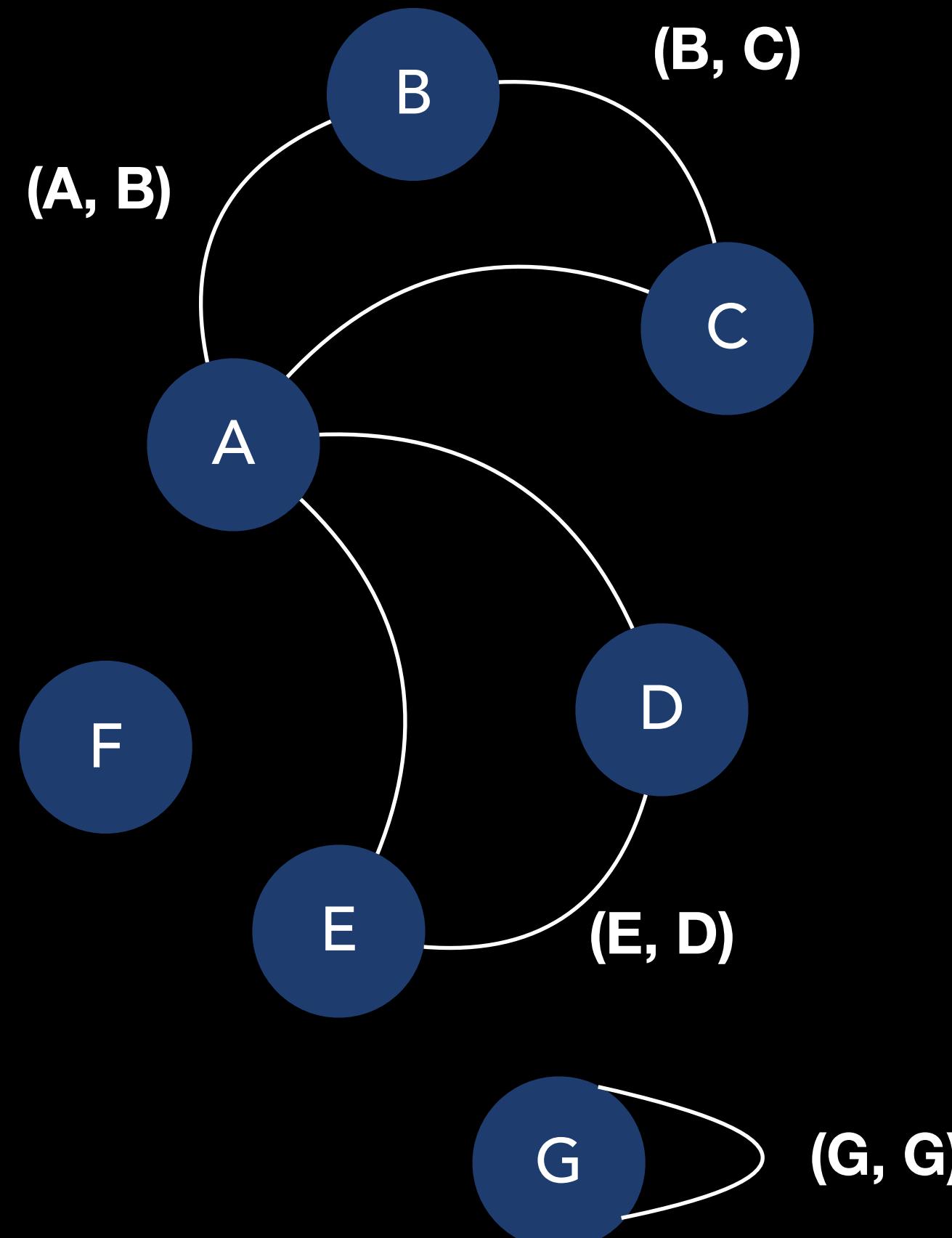
RULE OF THUMB ABOUT CONSISTENCY

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



WHAT IS A NETWORK? - VOCABULARY

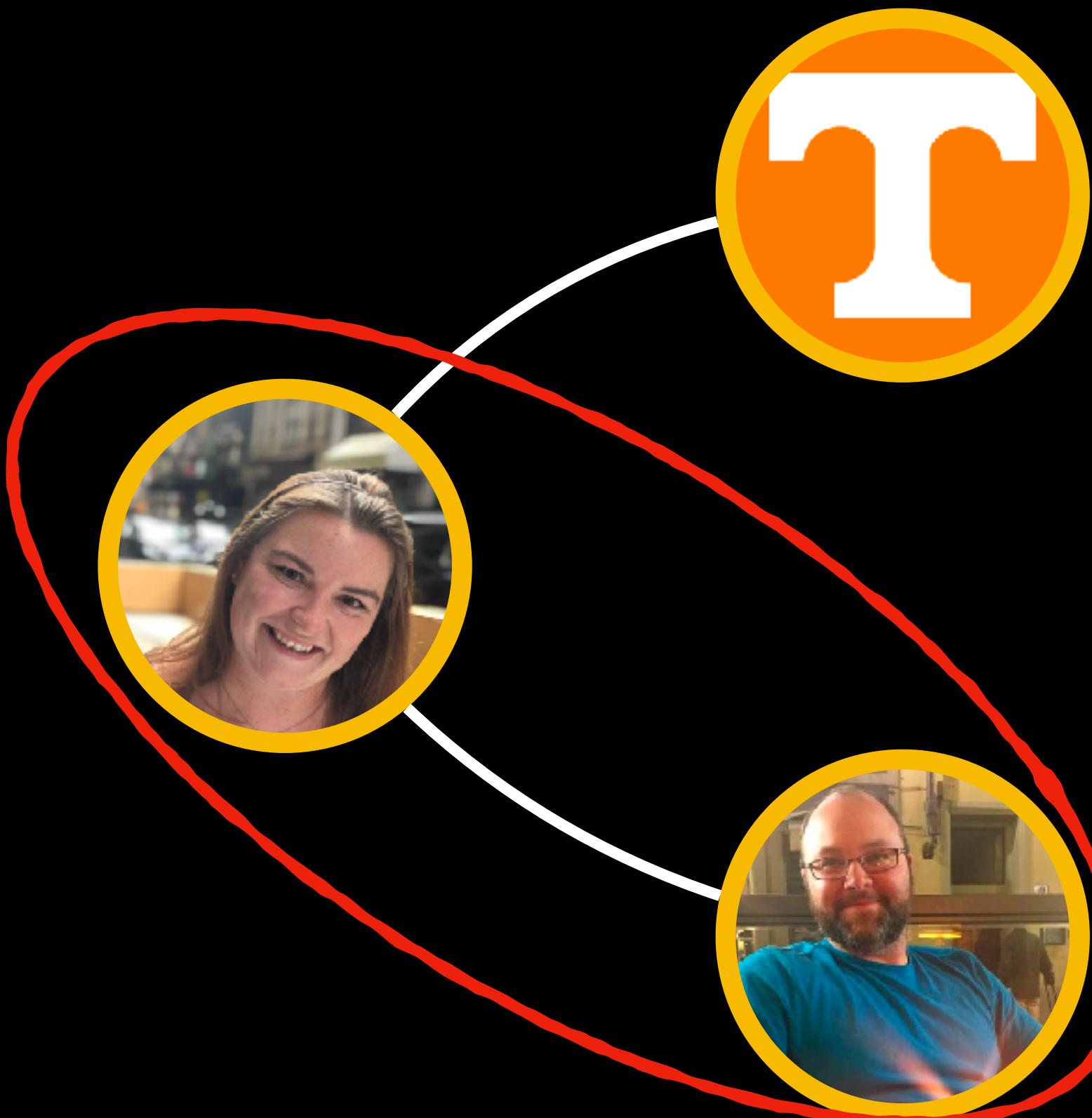
- A and B are neighbors
- F has no incident edges
 - No neighbors
 - I.e., a “singleton”
 - Nodes can be connected to themselves
 - I.e., a “loop”



Can also integrate “weights” into graphs

UNWEIGHTED AND WEIGHTED GRAPHS

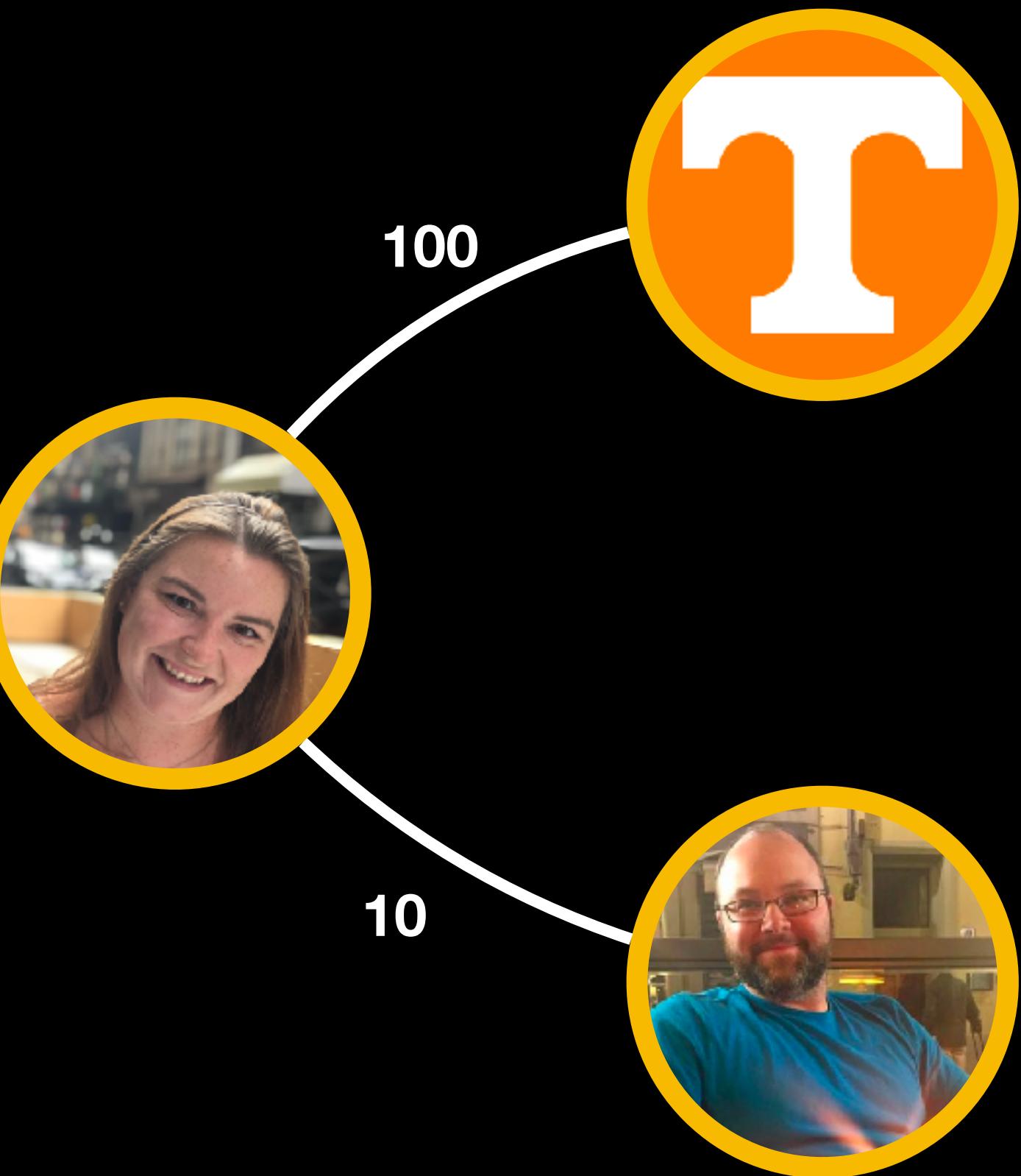
- Edges:
 - Tuple (pair) of vertices
 - All edges mean the same thing



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 Ties¹

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?

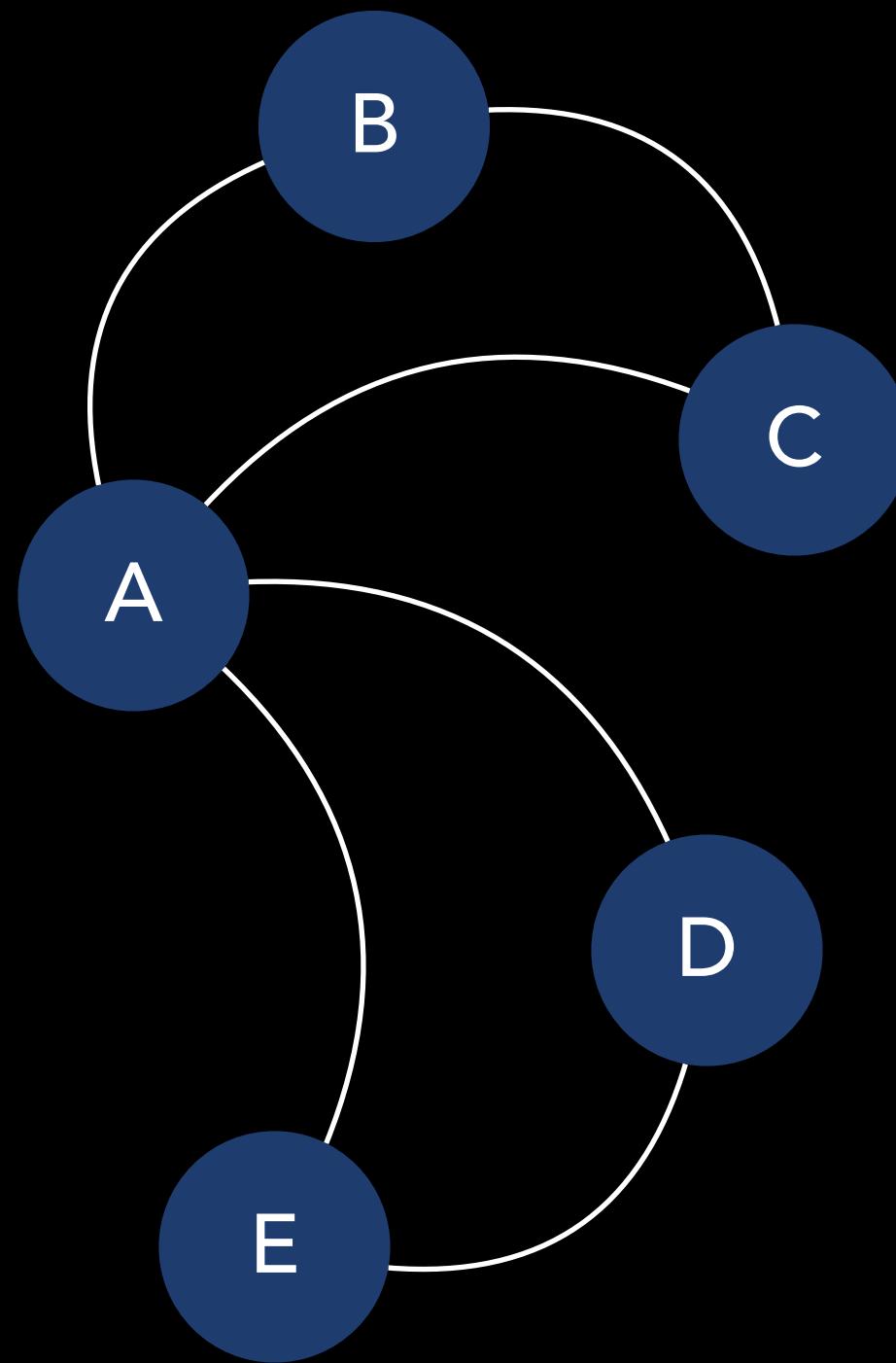
Importance of domain expertise

Representing Graphs

Representing and storing graphs

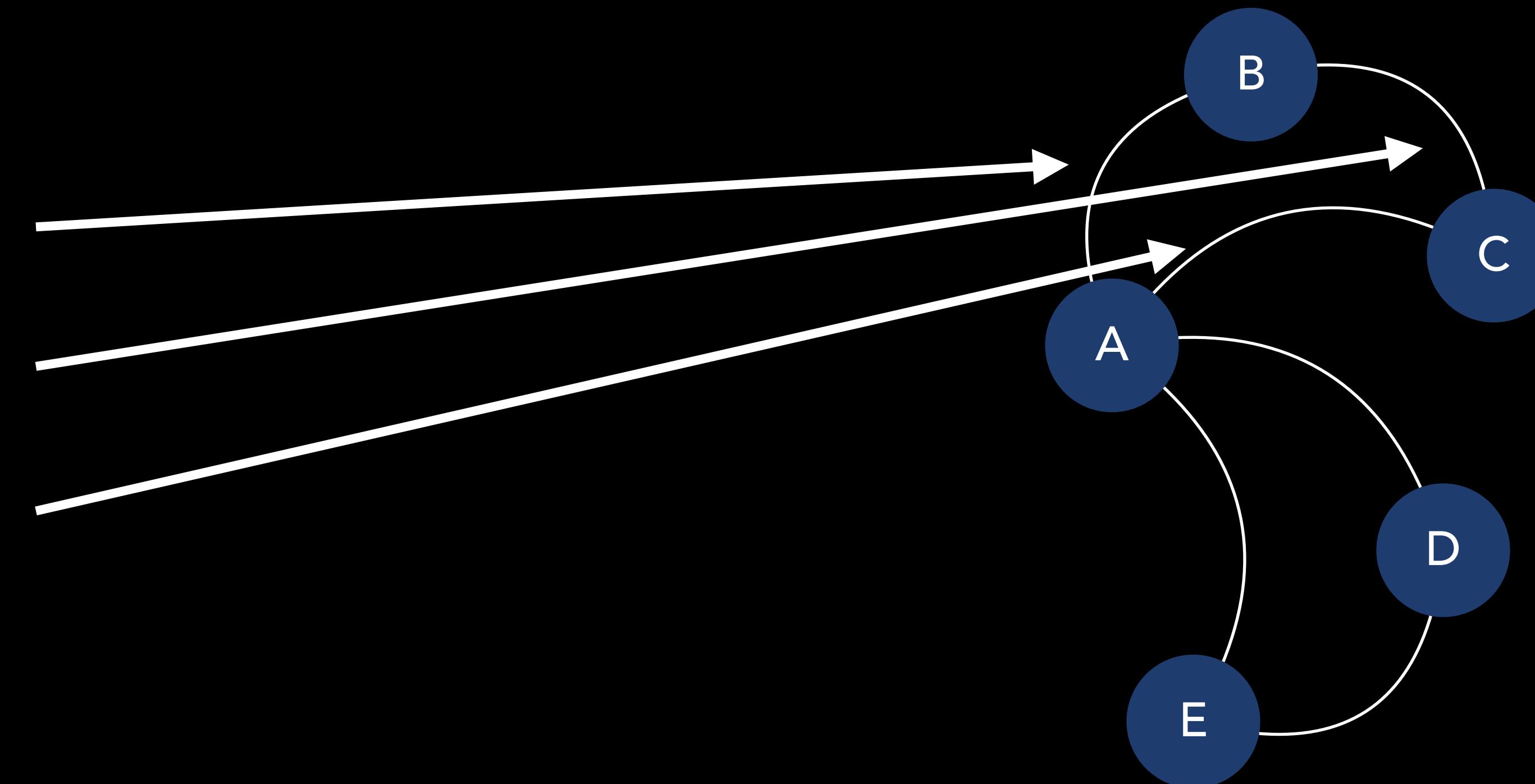
REPRESENTING GRAPHS - EDGE LISTS

- Mathematical Definition:
 $G(V, E)$
- E – Edges
 - Tuples of vertices (v_1, v_2)
 - Edge Lists:
 - A list of all edge tuples in the graph



REPRESENTING GRAPHS - EDGE LISTS

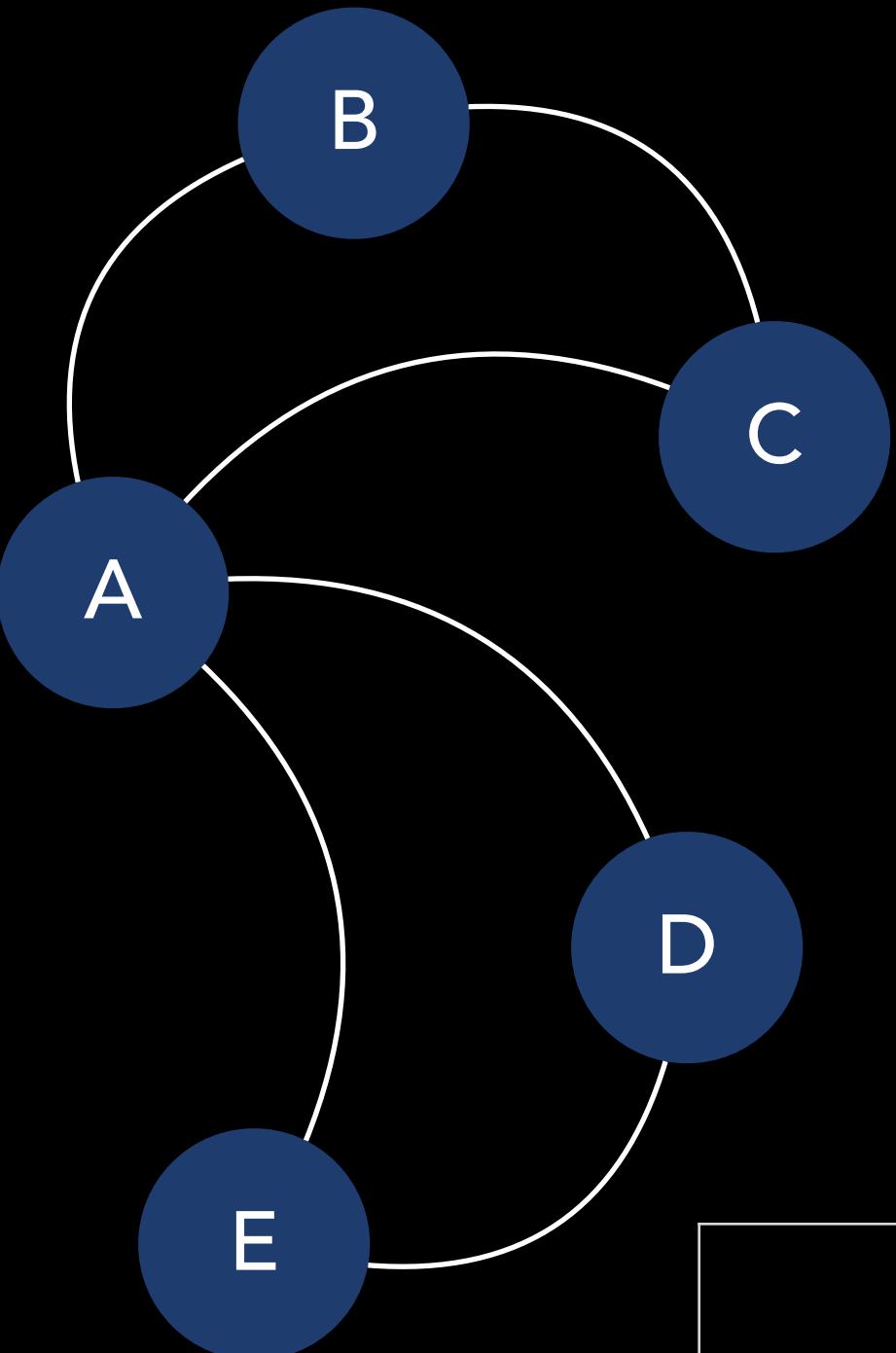
- A,B
- B,C
- A,C
- A,E
- A,D
- E,D



REPRESENTING GRAPHS

- ADJACENCY MATRIX

- Alternate Mathematical Definition: $G(V, E)$
- $N = |V|$, number of nodes
- Matrix $A \in \mathbb{R}^{N \times N}$
 - “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

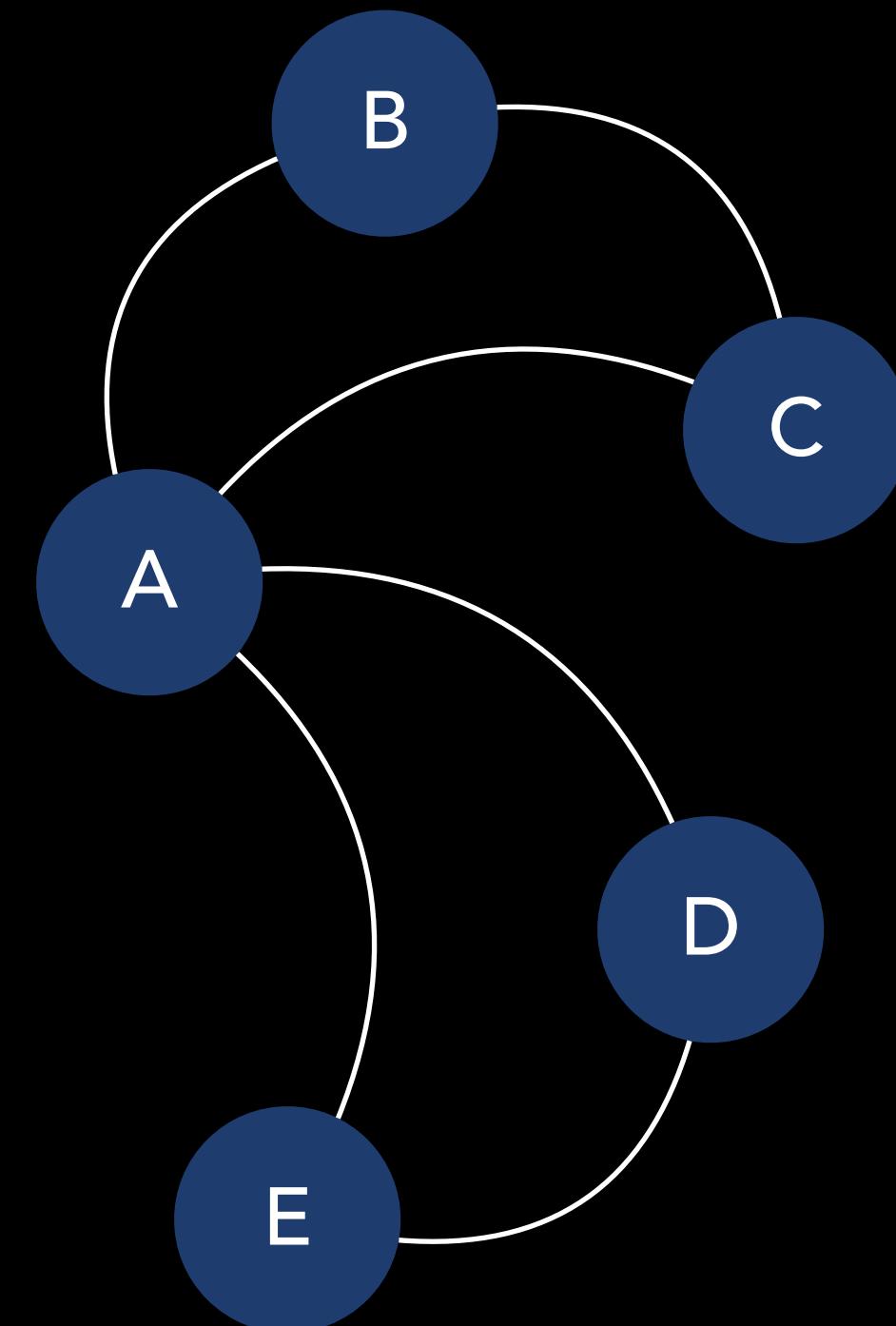


	A	B	C	D	E
A	0	1	1	1	1
B	1	0	1	0	0
C	1	1	0	0	0
D	1	0	0	0	1
E	1	0	0	1	0

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



NetworkX

Network Analysis in Python

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](#)

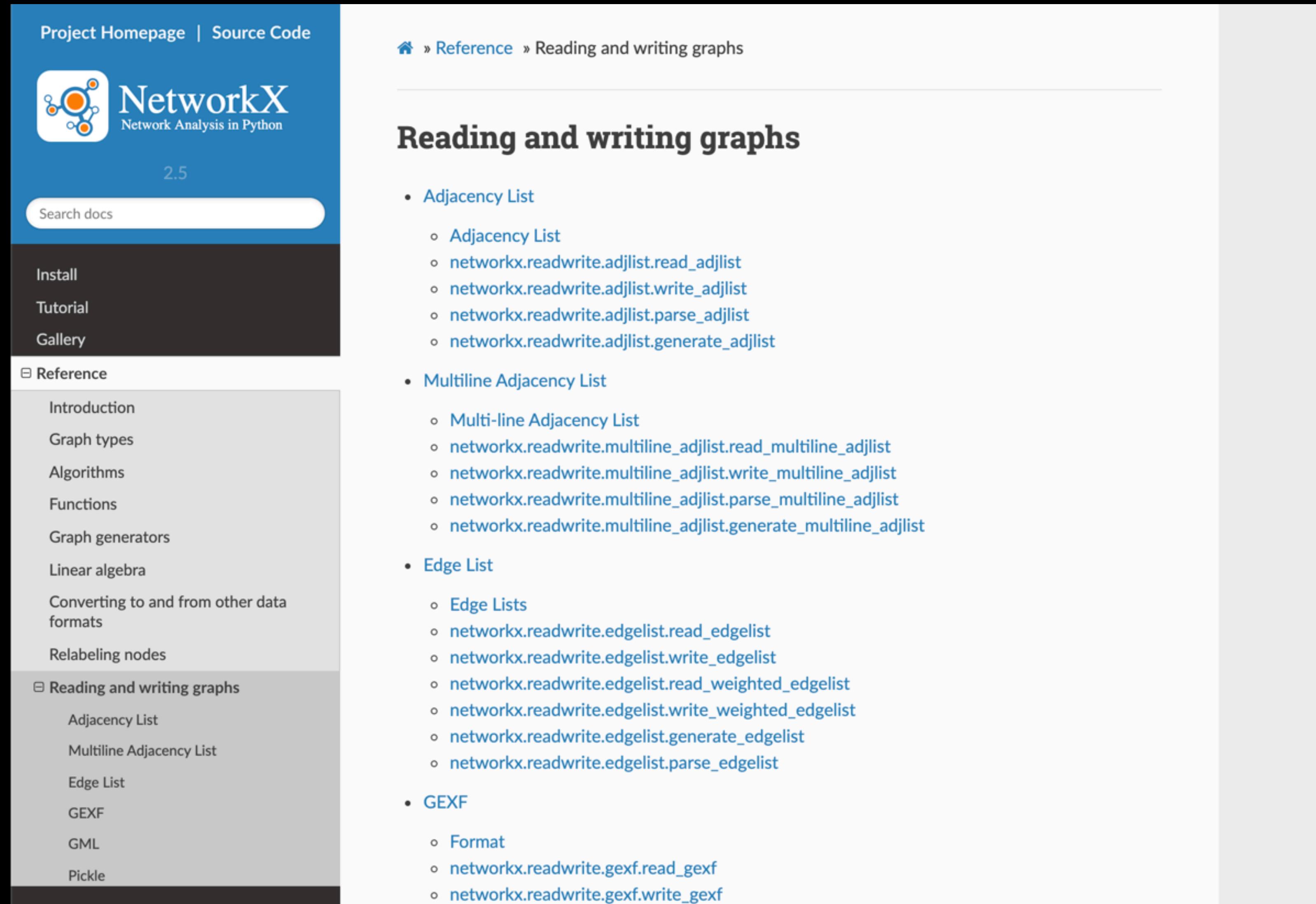
Software for complex networks

- 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



The image shows two screenshots of the NetworkX documentation. The left screenshot is the homepage, featuring a blue header with the NetworkX logo and version 2.5, a search bar, and a sidebar with links like 'Install', 'Tutorial', 'Gallery', and 'Reference'. The right screenshot is a detailed page under 'Reference' titled 'Reading and writing graphs', listing various file formats and their corresponding Python functions for reading and writing.

Project Homepage | Source Code

NetworkX
Network Analysis in Python

2.5

Search docs

Install

Tutorial

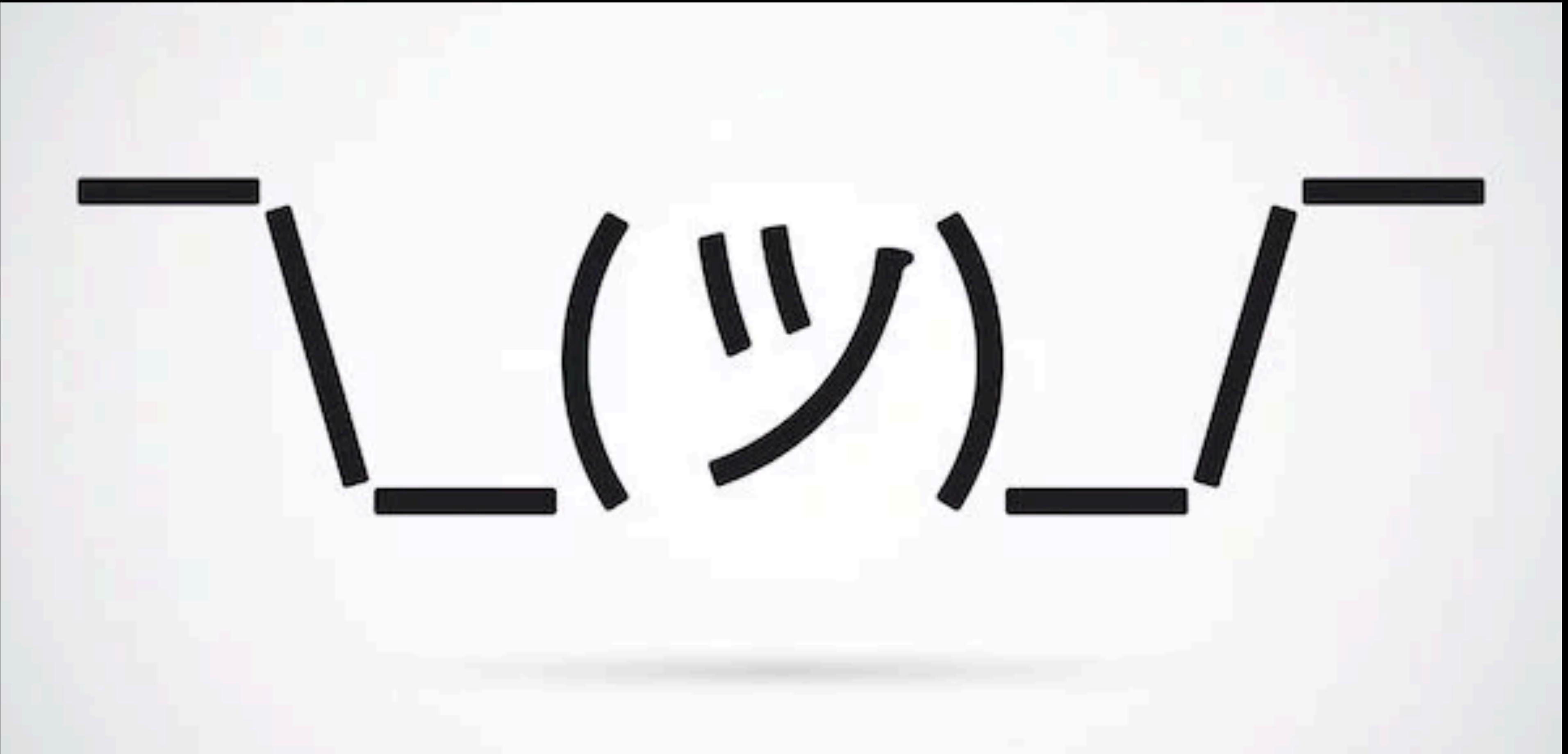
Gallery

Reference

- Introduction
- Graph types
- Algorithms
- Functions
- Graph generators
- Linear algebra
- Converting to and from other data formats
- Relabeling nodes

Reading and writing graphs

- Adjacency List
 - Adjacency List
 - `networkx.readwrite.adjlist.read_adjlist`
 - `networkx.readwrite.adjlist.write_adjlist`
 - `networkx.readwrite.adjlist.parse_adjlist`
 - `networkx.readwrite.adjlist.generate_adjlist`
- Multiline Adjacency List
 - Multi-line Adjacency List
 - `networkx.readwrite.multiline_adjlist.read_multiline_adjlist`
 - `networkx.readwrite.multiline_adjlist.write_multiline_adjlist`
 - `networkx.readwrite.multiline_adjlist.parse_multiline_adjlist`
 - `networkx.readwrite.multiline_adjlist.generate_multiline_adjlist`
- Edge List
 - Edge Lists
 - `networkx.readwrite.edgelist.read_edgelist`
 - `networkx.readwrite.edgelist.write_edgelist`
 - `networkx.readwrite.edgelist.read_weighted_edgelist`
 - `networkx.readwrite.edgelist.write_weighted_edgelist`
 - `networkx.readwrite.edgelist.generate_edgelist`
 - `networkx.readwrite.edgelist.parse_edgelist`
- GEXF
 - Format
 - `networkx.readwrite.gexf.read_gexf`
 - `networkx.readwrite.gexf.write_gexf`
- GML
- Pickle



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?

Best of the Web — Since 1994

TRUSTED

[Directory](#) | [Submit Site](#) | [Blog](#) | [Sign In](#)

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](#)

Home Services
[HVAC](#) [Movers](#) [Appliance Repair](#) [Pest Control](#)
[Plumbers](#)

Marketing
[Software](#) [Industry Specific](#) [SEO](#) [Companies](#)
[Social Media](#) [Local Marketing](#)

Science
[Biology](#) [Physics](#) [Technology](#) [Social Sciences](#)

Regional
[United States](#) [Canada](#) [Australia](#) [United Kingdom](#)

Family [Real Estate](#) [Gardening](#)

Attorneys and Firms
[Injury](#) [Divorce](#) [Attorneys](#) [Bankruptcy](#)
[Planning](#) [Criminal Defense](#)

Entertainment
[Movies](#) [Maps](#) [Education](#) [Libraries](#)

Sports
[Football](#) [Basketball](#) [Baseball](#) [Soccer](#) [Basketball](#)



[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

Early Methods: Curated Web directories

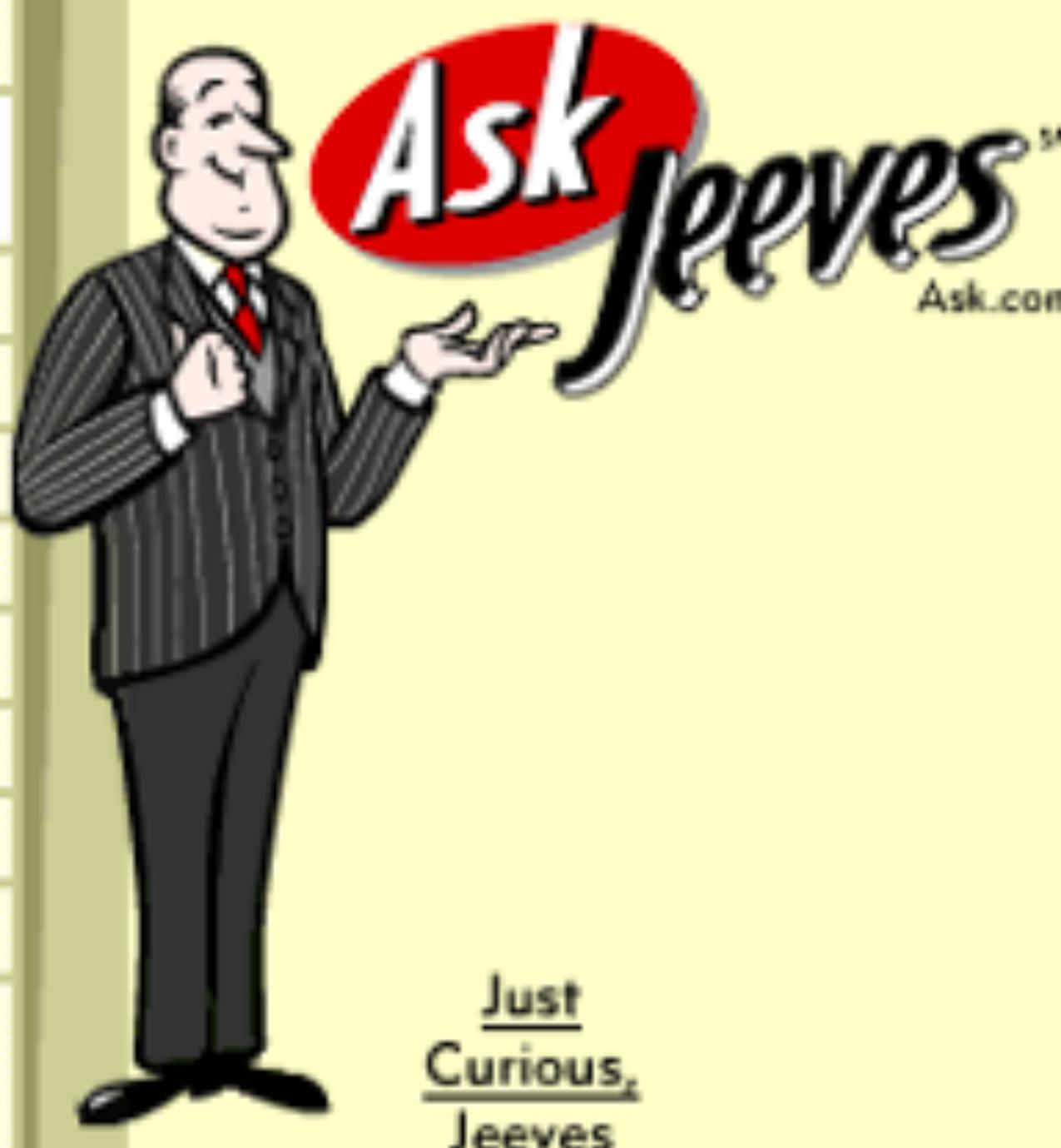
Every business owner or operator understands the power of online reviews, right? According to Podium, online reviews are becoming ever more important to how we as consumers do business, with 93% of consumers saying that online reviews do

• Home • About • Help • Corporate Services

May I Suggest:

- PERSONAL JEEVES
- ANSWER POINT
- MONEY
- TRAVEL
- HEALTH
- COMPUTERS
- ENTERTAINMENT
- HOME & FAMILY
- SHOPPING

ASK JEEVES
FOR **Kids!**



Just
Curious,
Jeeves

PERSONAL
Jeeves
is here!
→

Have a
Question?
Just type it
in and click **Ask!**

Ask!

Most Recent Questions About **Business**:

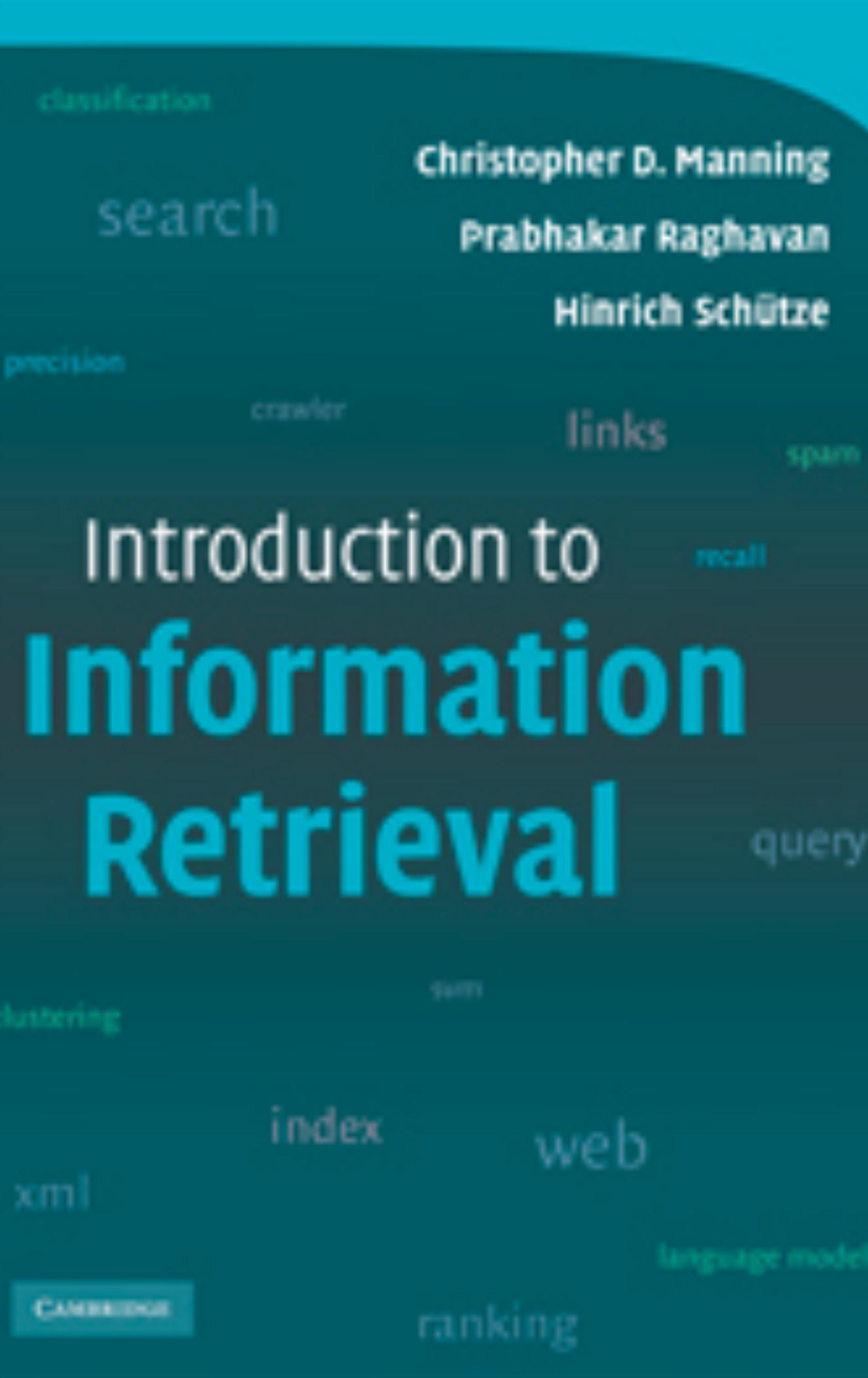
Where can I find the Web site
for the company British
Telecom?

Ask!

What are people asking RIGHT NOW? →

• Make Jeeves Your Homepage • Ask Jeeves U.K. • Adver

Then: Information Retrieval and Search Engines



Introduction to Information Retrieval



Query: Amanda Gorman

Amanda Gorman

From Wikipedia, the free encyclopedia

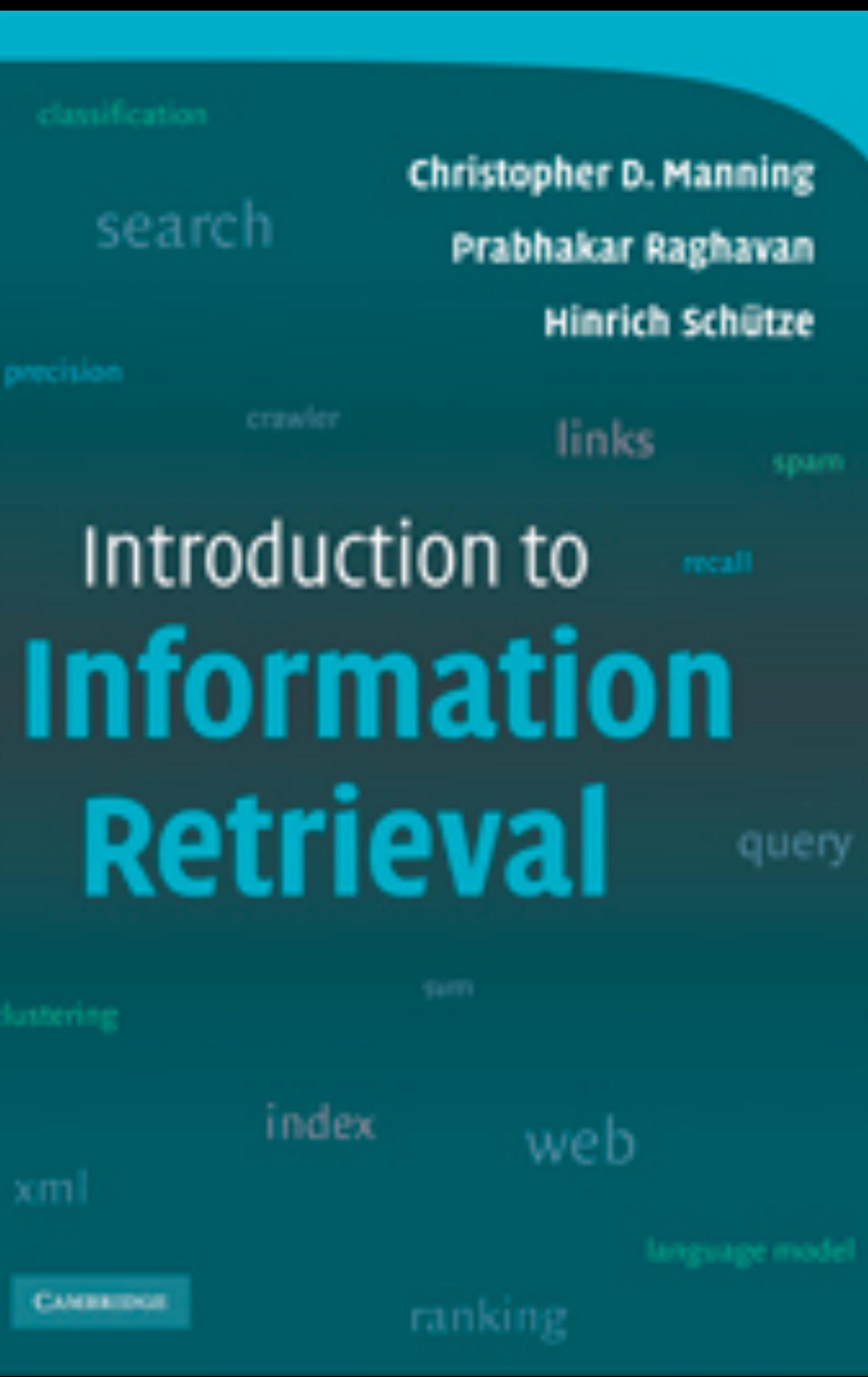
Amanda S. C. Gorman^[1] (born 1998) is an American poet and activist. Her work focuses on issues of oppression, feminism, race, and marginalization, as well as the African diaspora. Gorman was the first person to be named National Youth Poet Laureate. She published the poetry book *The One for Whom Food 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



Introduction to Information Retrieval



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



We offer
online stores!
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





We offer
**online
stores!**

Amanda gorman, Amanda gorman,
Amanda gorman,
Amanda gorman



We offer
**online
stores!**

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

```
7 <title>How to create the right meta description &bull; Yoast</title>
8
9
10 <!-- This site has installed PayPal for WooCommerce v2.1.6 - https://www.angelleye.com/product/woocommerce-paypal-plugin/ -->
11 <link rel="dns-prefetch" href="//www.googletagmanager.com">
12 <link rel="preconnect" href="https://www.googletagmanager.com/" crossorigin>
13
14 <!-- Google Tag Manager for WordPress by gtm4wp.com -->
15 <script data-cfasync="false" data-pagespeed-no-defer>//<![CDATA[
16     var gtm4wp_datalayer_name = "dataLayer";
17     var dataLayer = dataLayer || [];
18     var gtm4wp_use_sku_instead      = 1;
19     var gtm4wp_id_prefix           = '';
20     var gtm4wp_remarketing         = false;
21     var gtm4wp_eec                 = 1;
22     var gtm4wp_classicec          = 1;
23     var gtm4wp_currency            = 'USD';
24     var gtm4wp_product_per_impression = 0;
25     var gtm4wp_needs_shipping_address = false;
26 //]]>
27 </script>
28 <!-- End Google Tag Manager for WordPress by gtm4wp.com -->
29 <!-- This site is optimized with the Yoast SEO Premium plugin v15.7 - https://yoast.com/wordpress/plugins/seo/ -->
30 <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!" />
31 <meta name="robots" content="index, follow, max-snippet:-1, max-image-preview:large, max-video-preview:-1" />
32 <link rel="canonical" href="https://yoast.com/meta-descriptions/" />
33 <meta property="og:locale" content="en_US" />
34 <meta property="og:type" content="article" />
35 <meta property="og:title" content="How to create the right meta description &bull; Yoast" />
36 <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!" />
37 <meta property="og:url" content="https://yoast.com/meta-descriptions/" />
38 <meta property="og:site_name" content="Yoast" />
39 <meta property="article:publisher" content="https://www.facebook.com/yoast" />
40 <meta property="article:published_time" content="2020-12-21T12:30:00+00:00" />
41 <meta property="article:modified_time" content="2021-01-26T14:12:35+00:00" />
42 <meta property="og:image" content="https://yoast.com/app/uploads/2015/01/Meta_description_FI.png" />
43 <meta property="og:image:width" content="1200" />
44 <meta property="og:image:height" content="628" />
45 <meta name="twitter:card" content="summary_large_image" />
46 <meta name="twitter:creator" content="@yoast" />
47 <meta name="twitter:site" content="@yoast" />
48 <meta name="twitter:label1" content="Written by" />
49 <meta name="twitter:data1" content="Willemien Hallebeek" />
50 <meta name="twitter:label2" content="Est. reading time" />
51 <meta name="twitter:data2" content="11 minutes" />
52 <script type="application/ld+json" class="yoast-schema-graph">
[{"@type": "Organization", "@id": "https://yoast.com/#organization", "@name": "Yoast", "@url": "https://www.yoast.com", "sameAs": ["https://www.facebook.com/yoast", "https://www.instagram.com/yoast/", "https://www.linkedin.com/company/1414157/", "https://www.youtube.com/yoast", "https://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-valk"}]
```

Can hide meta tags full of common search terms

Spamdexing - Wikipedia

en.wikipedia.org/wiki/Spamdexing#~:text=Spamdexing%20(also%20known%20as%20s... Other Bookmarks

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Spamdexing

From Wikipedia, the free encyclopedia

This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.

Find sources: "Spamdexing" – news · newspapers · books · scholar · JSTOR (February 2013) (Learn how and when to remove this template message)

Spamdexing (also known as search engine spam, search engine poisoning, black-hat search engine optimization (SEO), search spam or web spam)^[1] is the deliberate manipulation of search engine indexes. It involves a number of methods, such as link building and repeating unrelated phrases, to manipulate the relevance or prominence of resources indexed, in a manner inconsistent with the purpose of the indexing system.^{[2][3]}

Spamdexing could be considered to be a part of search engine optimization, although there are many search engine optimization methods that improve the quality and appearance of the content of web sites and serve content useful to many users.^[4]

Search engines use a variety of algorithms to determine relevancy ranking. Some of these include determining whether the search term appears in the body text or URL of a web page. Many search engines check for instances of spamdexing and will remove suspect pages from their indexes. Also, search-engine operators can quickly block the results listing from entire websites that use spamdexing, perhaps in response to user complaints of false matches. The rise of spamdexing in the mid-1990s made the leading search engines of the time less useful. Using unethical methods to make websites rank higher in search engine results than they otherwise would is commonly referred to in the SEO (search engine optimization) industry as "black-hat SEO". These methods are more focused on breaking the search-engine-promotion rules and guidelines. In addition to this, the perpetrators run the risk of their websites being severely penalized by the Google Panda and Google Penguin search-results ranking algorithms.^[5]

Common spamdexing techniques can be classified into two broad classes: content spam^[4] (or term spam) and link spam.^[3]

Contents [hide]

- 1 History
- 2 Content spam
 - 2.1 Keyword stuffing
 - 2.2 Hidden or invisible text
 - 2.3 Meta-tag stuffing
 - 2.4 Doorway pages
 - 2.5 Scraper sites
 - 2.6 Article spinning

1 History

2 Content spam

- 2.1 Keyword stuffing
- 2.2 Hidden or invisible text
- 2.3 Meta-tag stuffing
- 2.4 Doorway pages
- 2.5 Scraper sites
- 2.6 Article spinning
- 2.7 Machine translation
- 2.8 Pages with no information related to page title

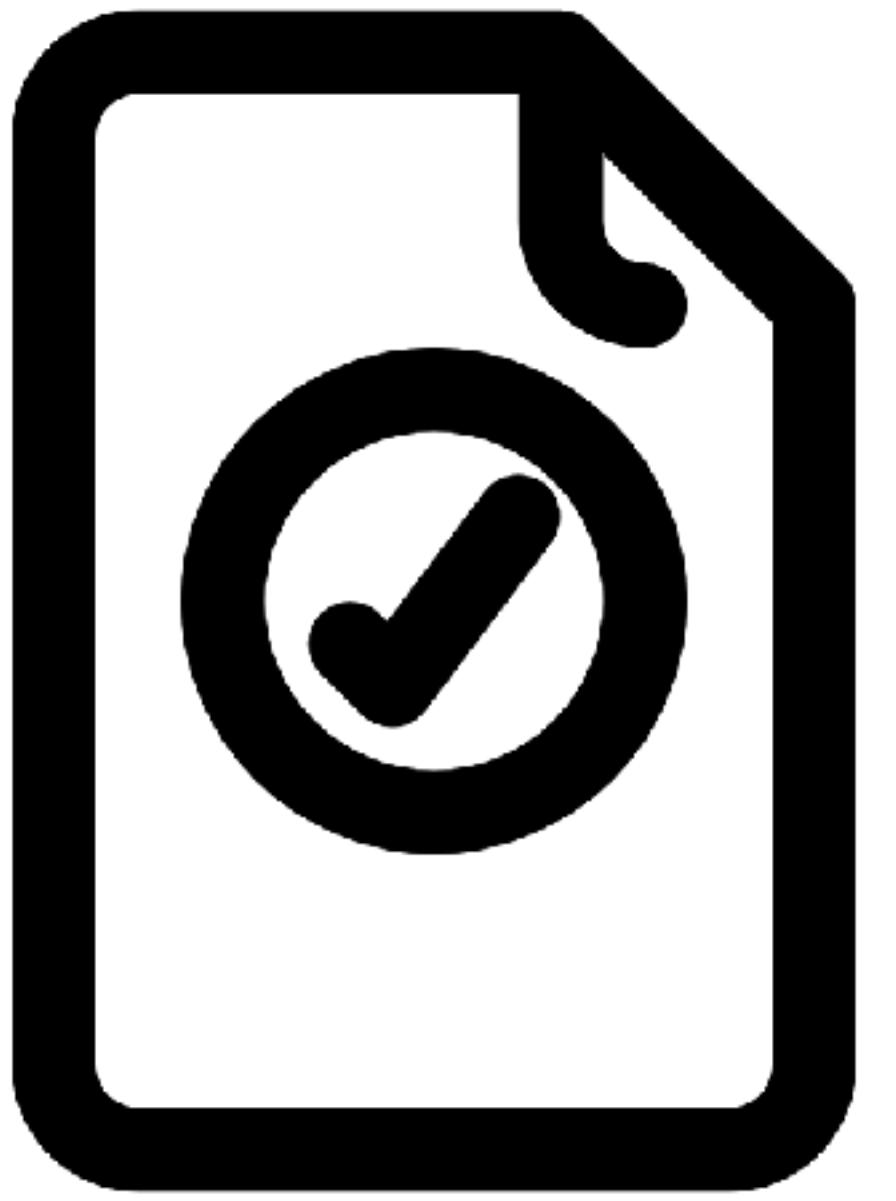
3 Link spam

- 3.1 Link farms
- 3.2 Private blog networks
- 3.3 Hidden links
- 3.4 Sybil attack
- 3.5 Spam blogs
- 3.6 Guest blog spam
- 3.7 Buying expired domains
- 3.8 Cookie stuffing
- 3.9 Using world-writable pages
 - 3.9.1 Spam in blogs
 - 3.9.2 Comment spam
 - 3.9.3 Wiki spam
 - 3.9.4 Referrer log spamming
 - 3.9.5 Countermeasures

4 Other types

- 4.1 Mirror websites
- 4.2 URL redirection
- 4.3 Cloaking

Motivating Question: How do extract insights from Web data?



Finding Trustworthy
Pages

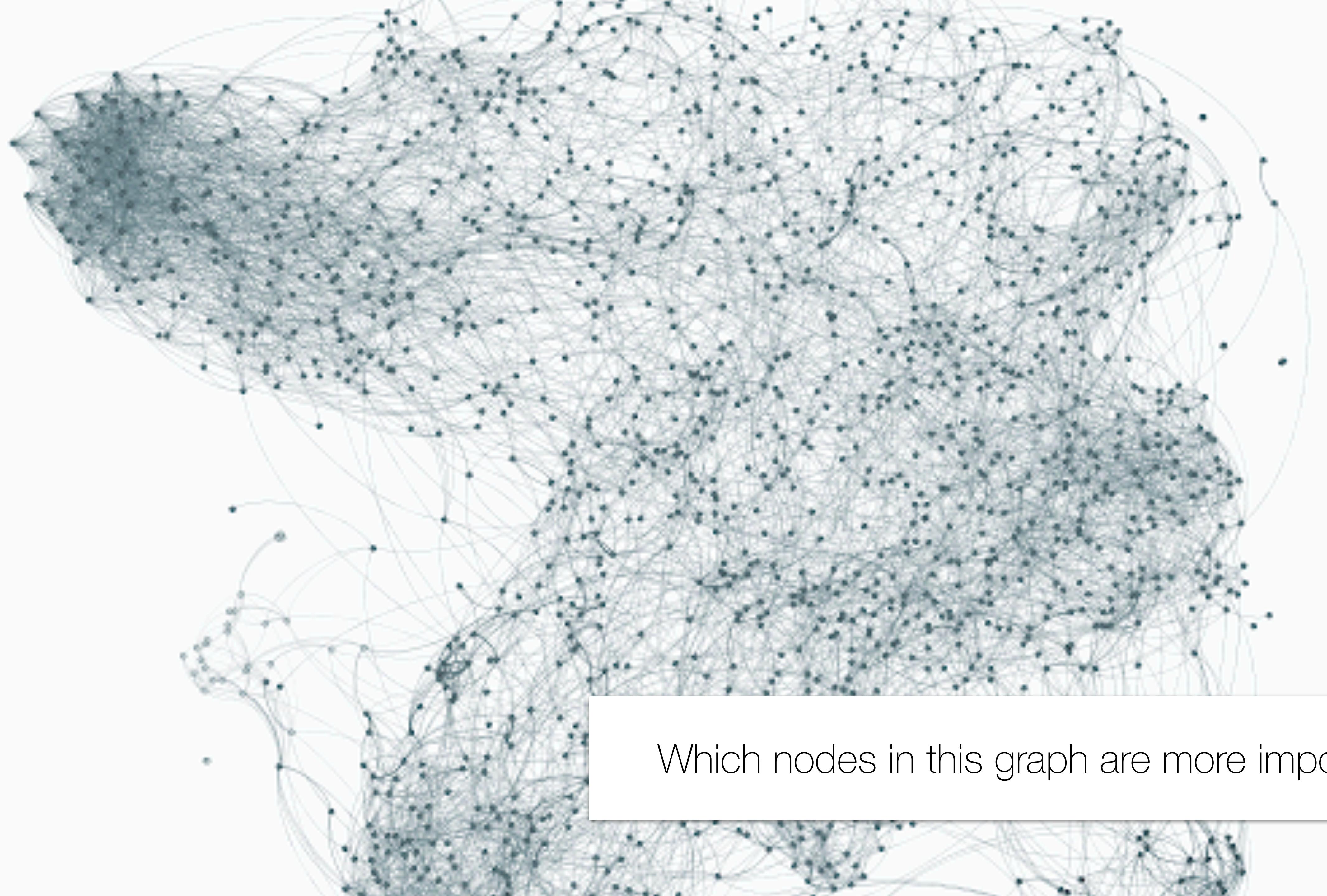


Finding Relevant
Pages

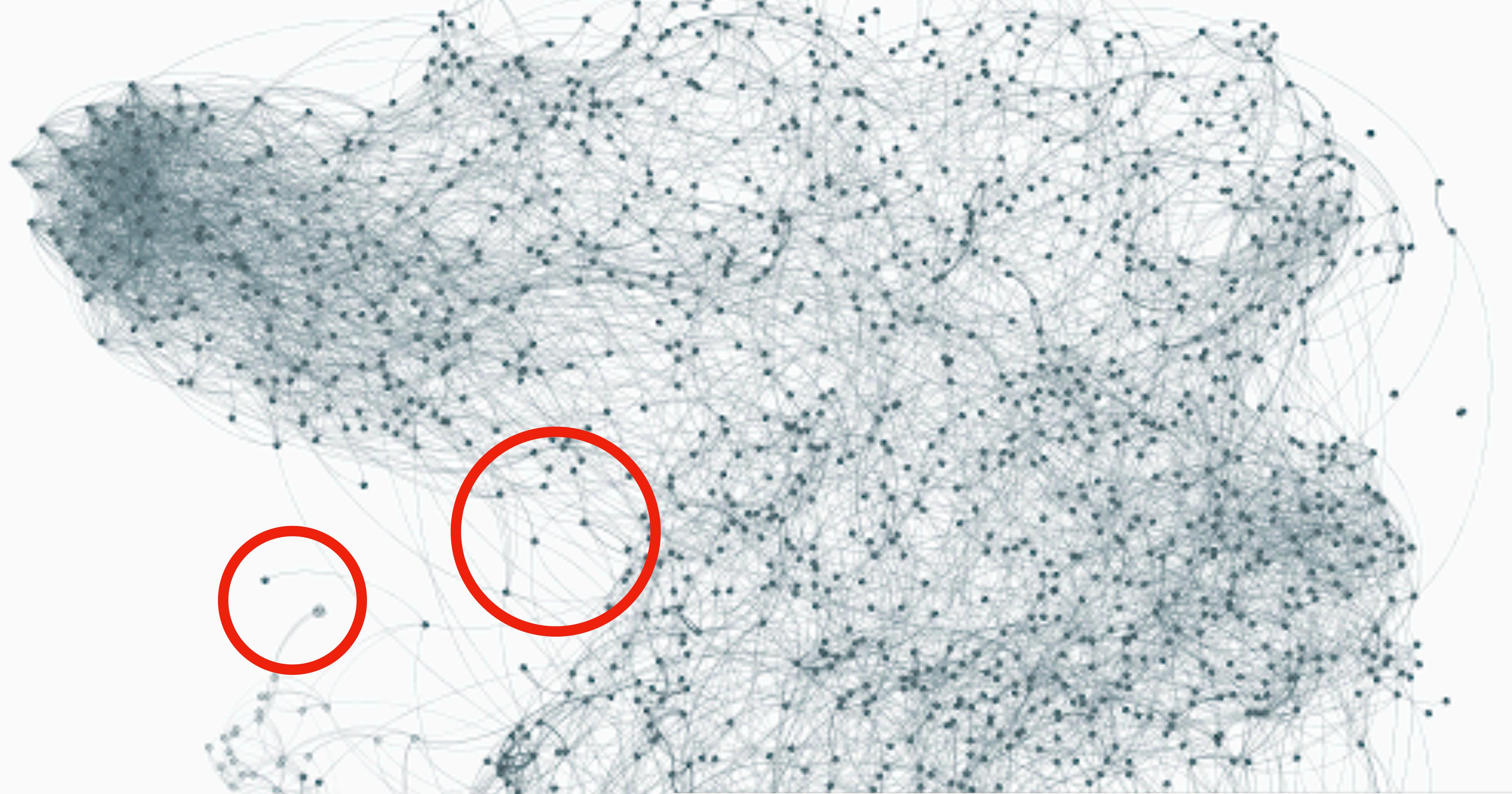
How do we find such pages?

A complex network graph consisting of numerous small black dots representing nodes, connected by a dense web of thin grey lines representing edges. The nodes are clustered into several large, irregular groups of varying sizes, with some groups appearing more densely packed than others. The overall structure is organic and interconnected.

Answer: Networks!



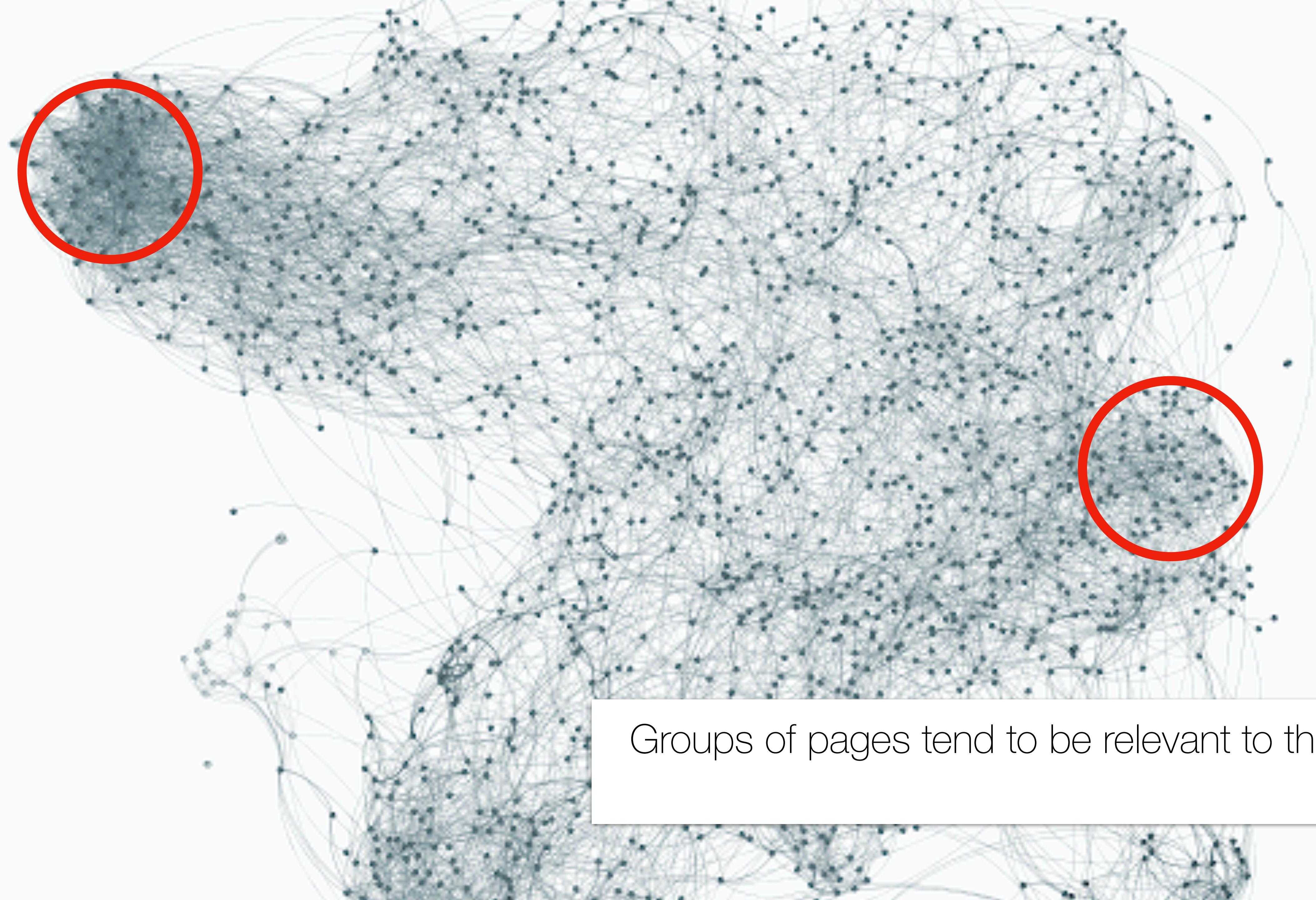
Which nodes in this graph are more important?



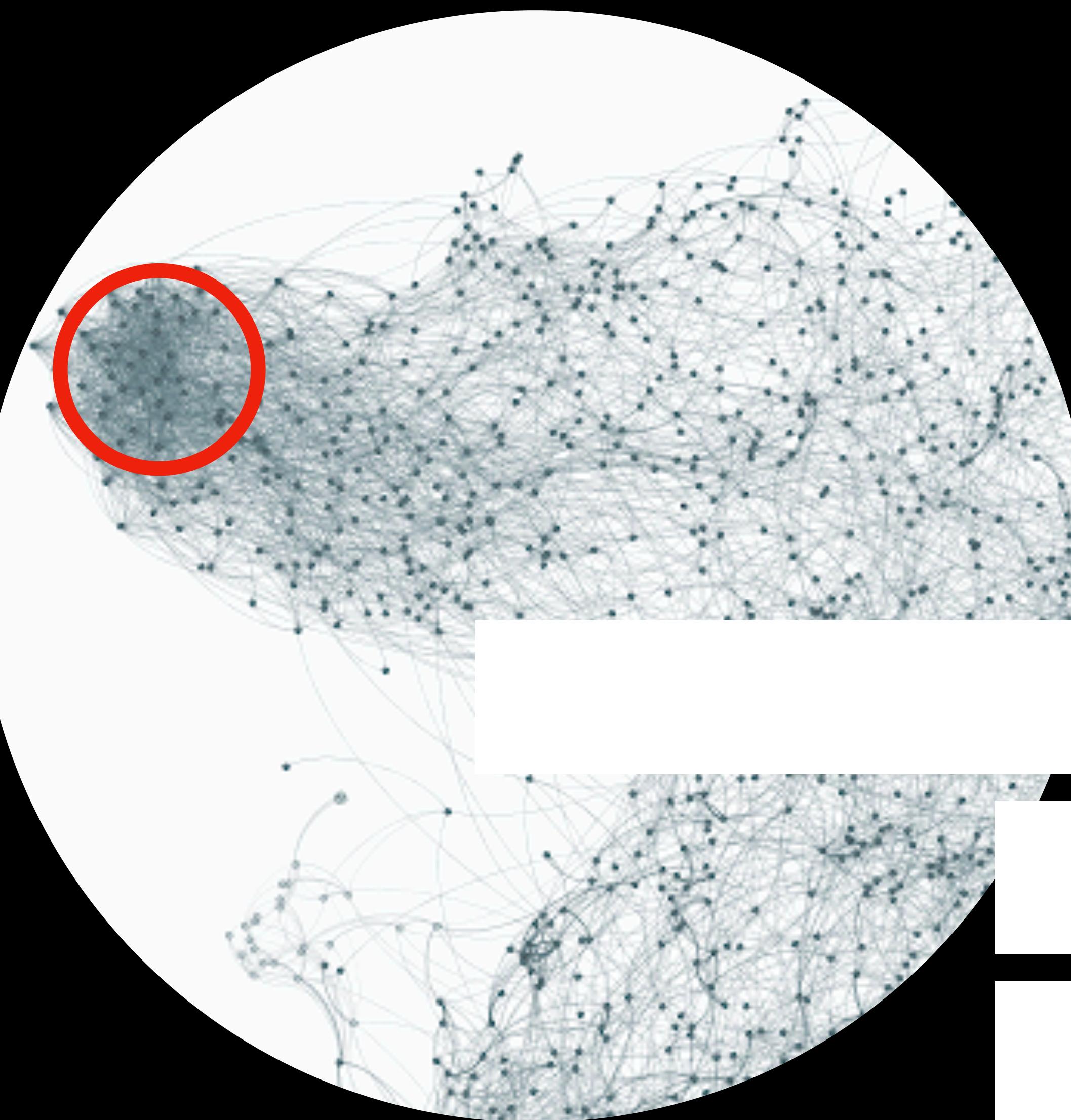
Pages with few links tend to be of
lower quality/less trustworthy



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



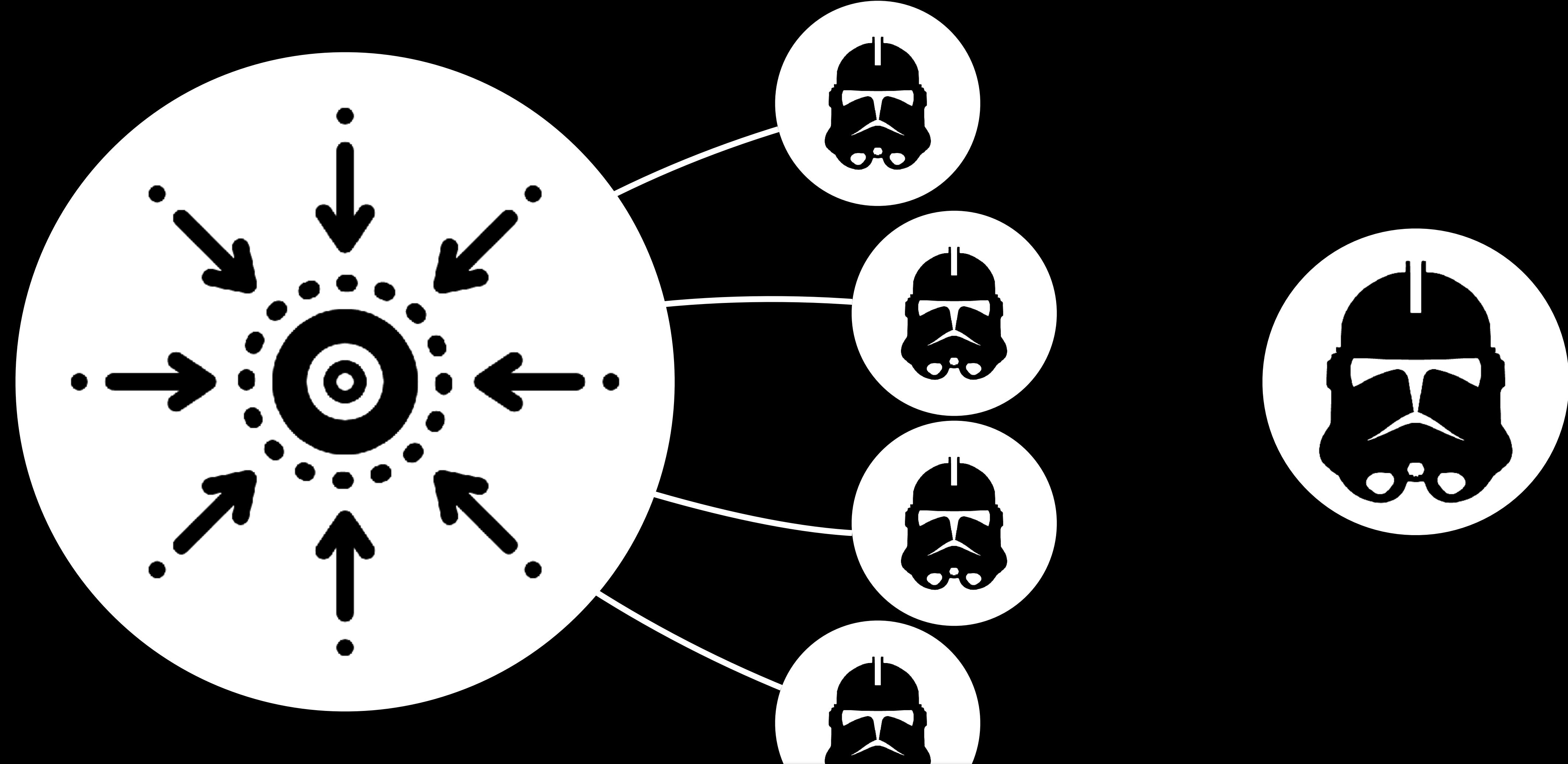
Groups of pages tend to be relevant to the same topic



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)

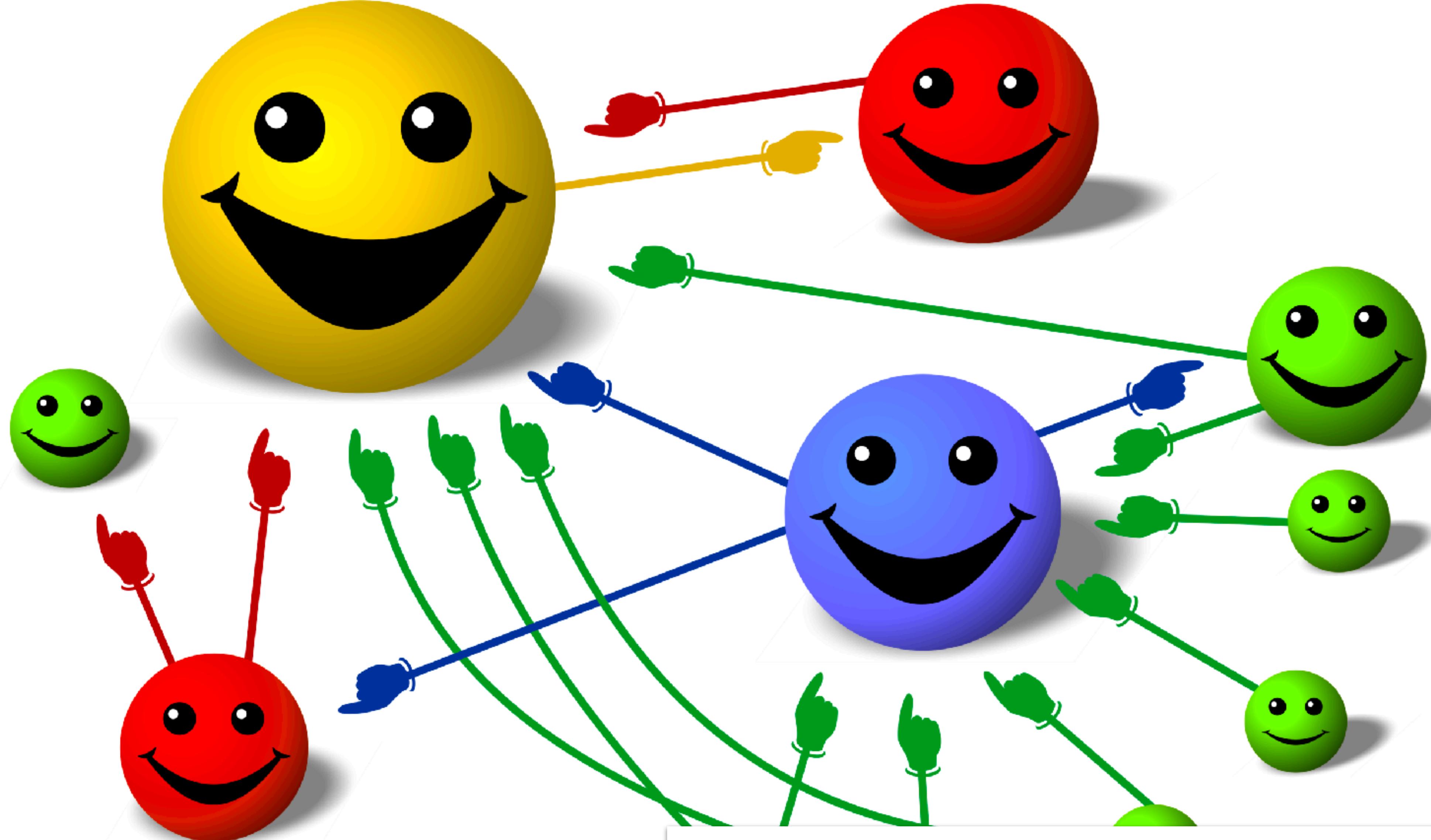
How do we find important, trustworthy pages?

Robust to keyword manipulation

Robust to link manipulation

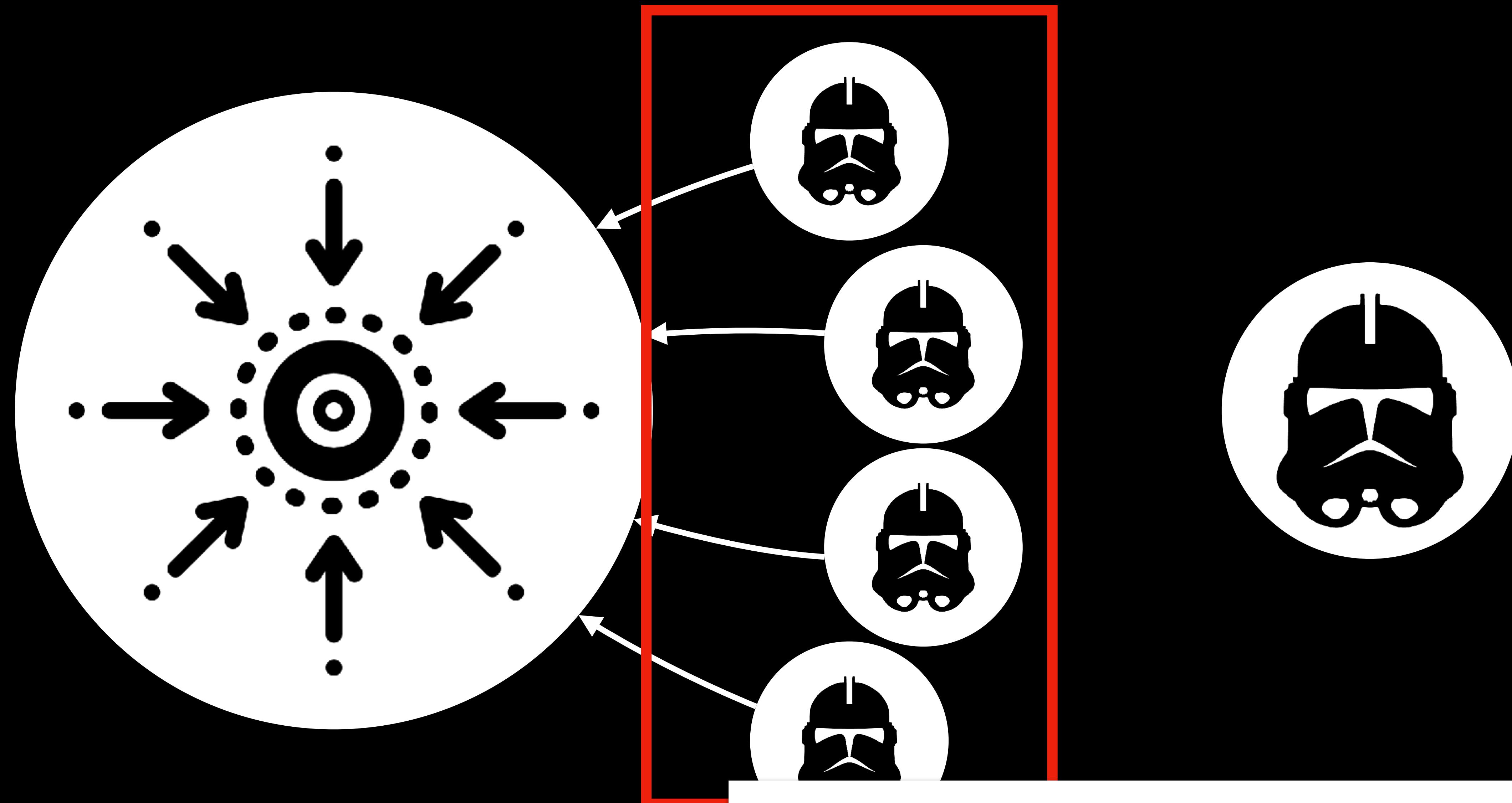


PageRank, by Larry Page and Sergey Brin



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

PageRank

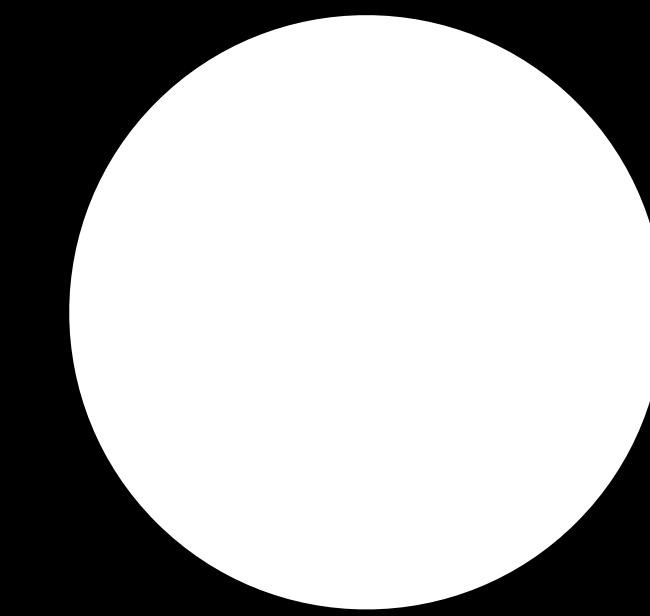
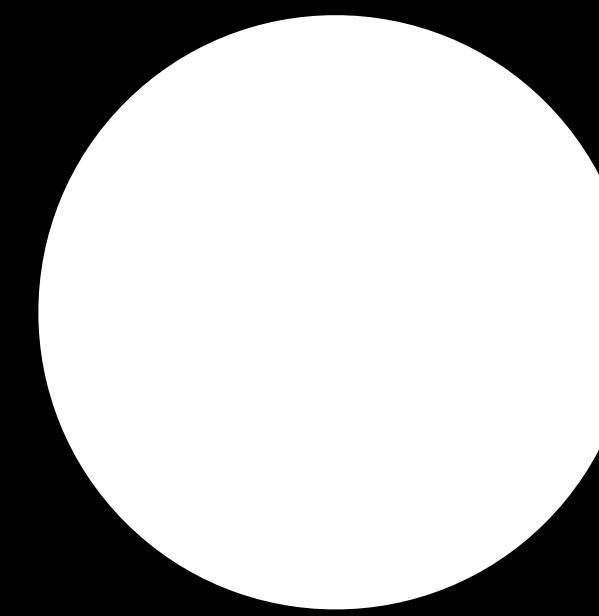


These “fake” nodes will have few incoming links

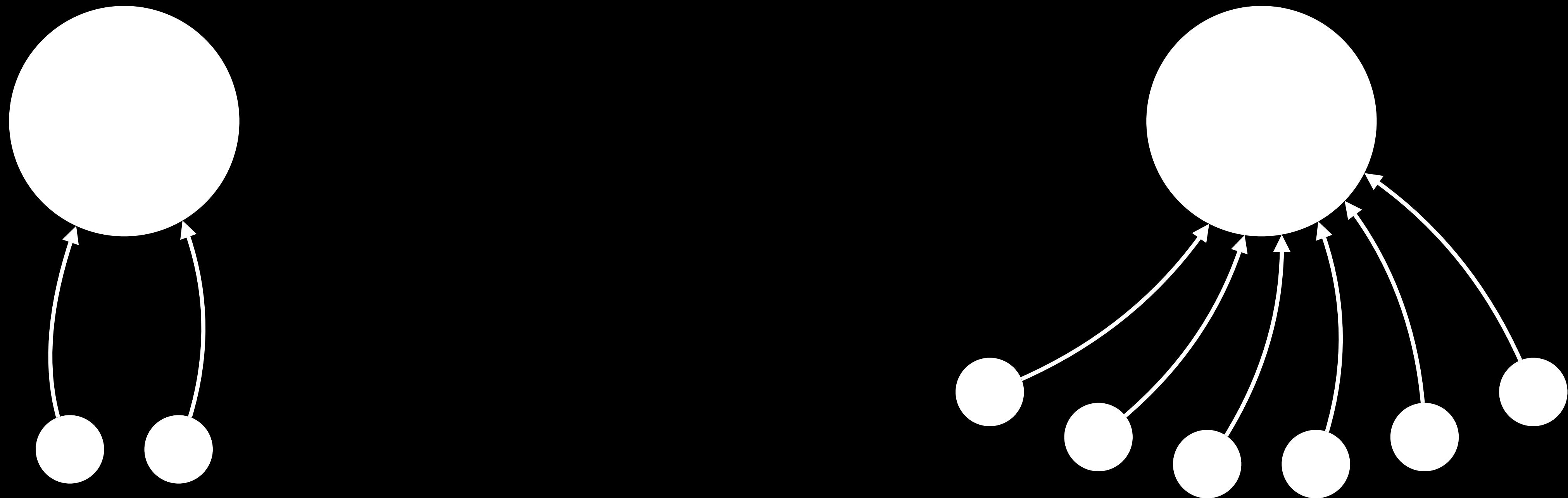


Think of the Web as a graph, with directed links
between web pages

Intuition: Pages vote on each other via linking

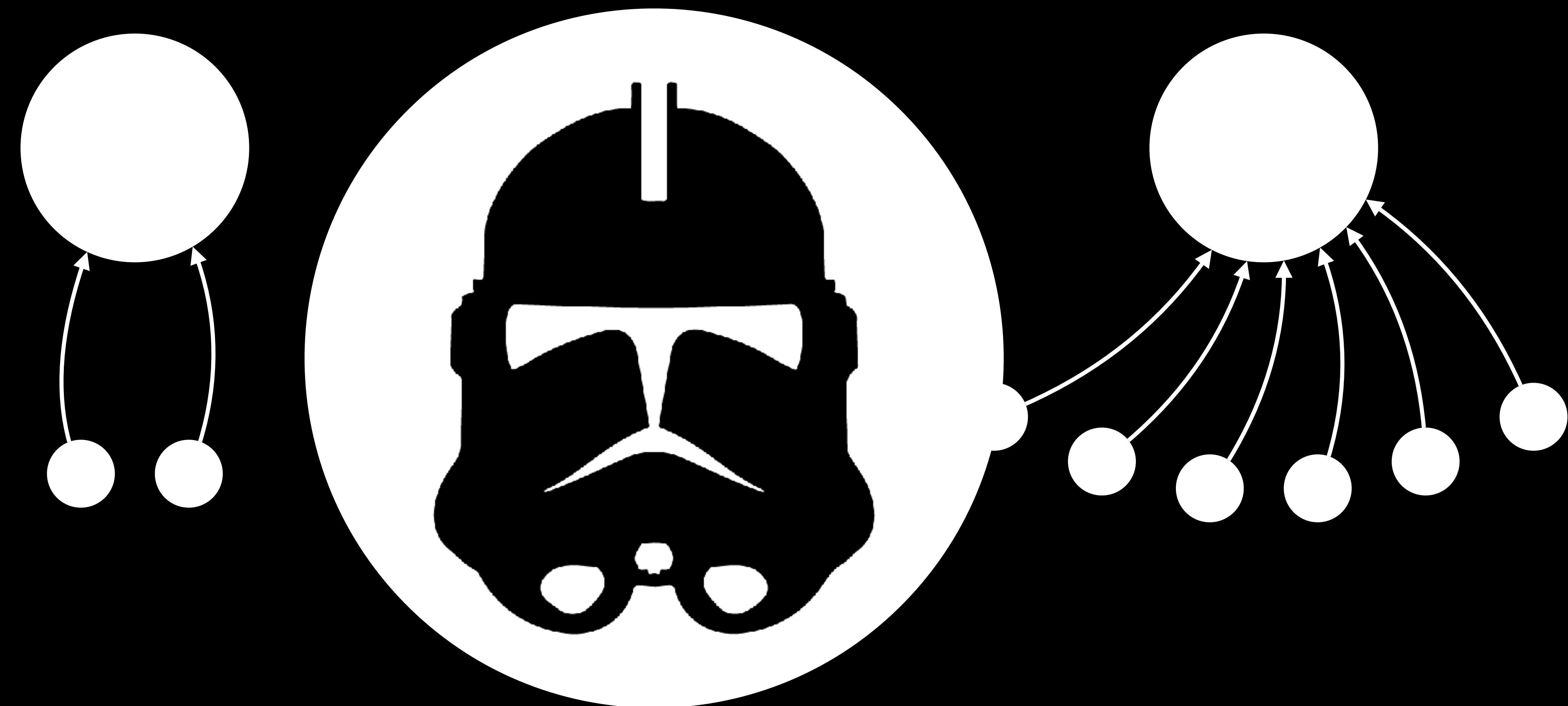


Intuition: Pages vote on each other via linking

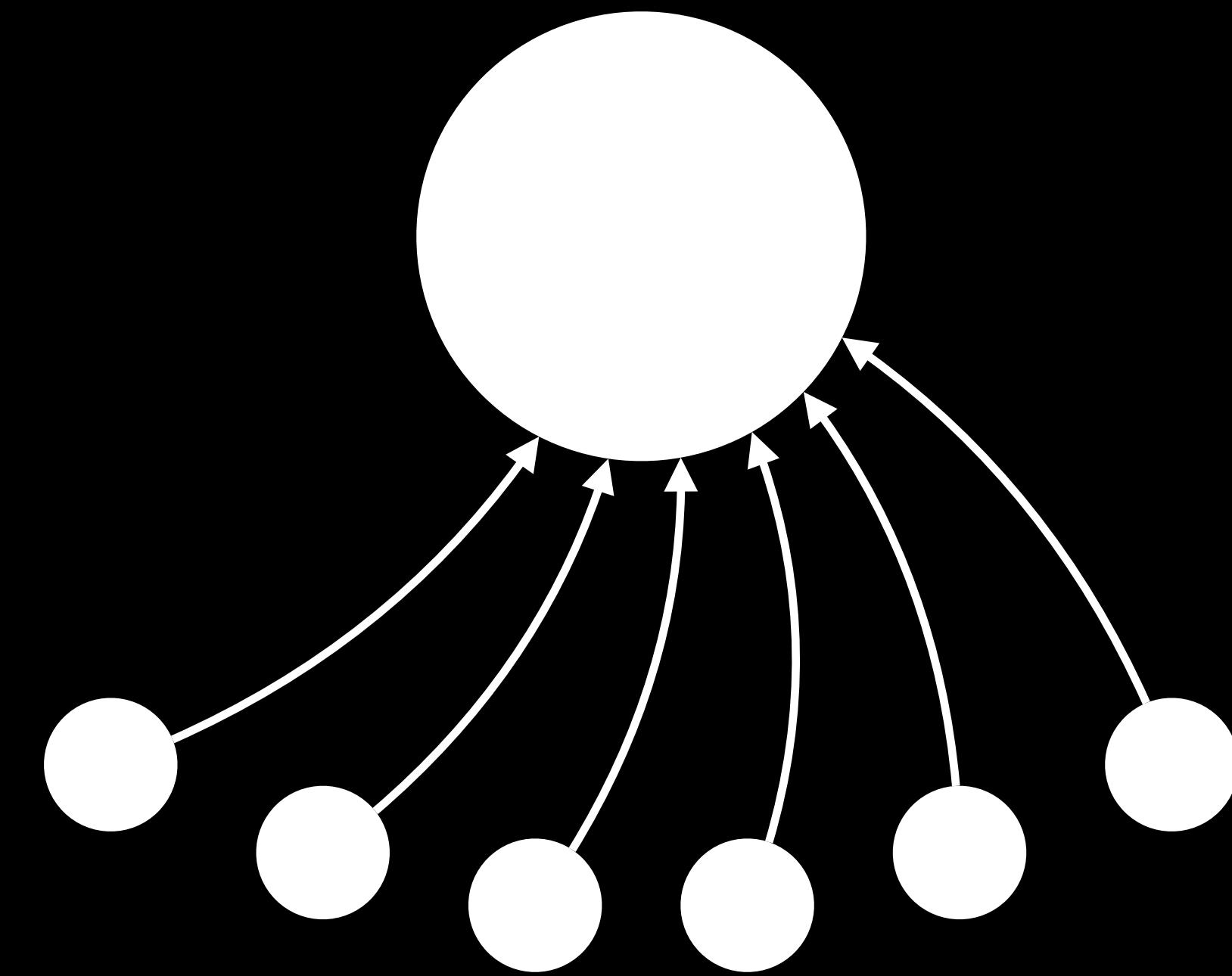
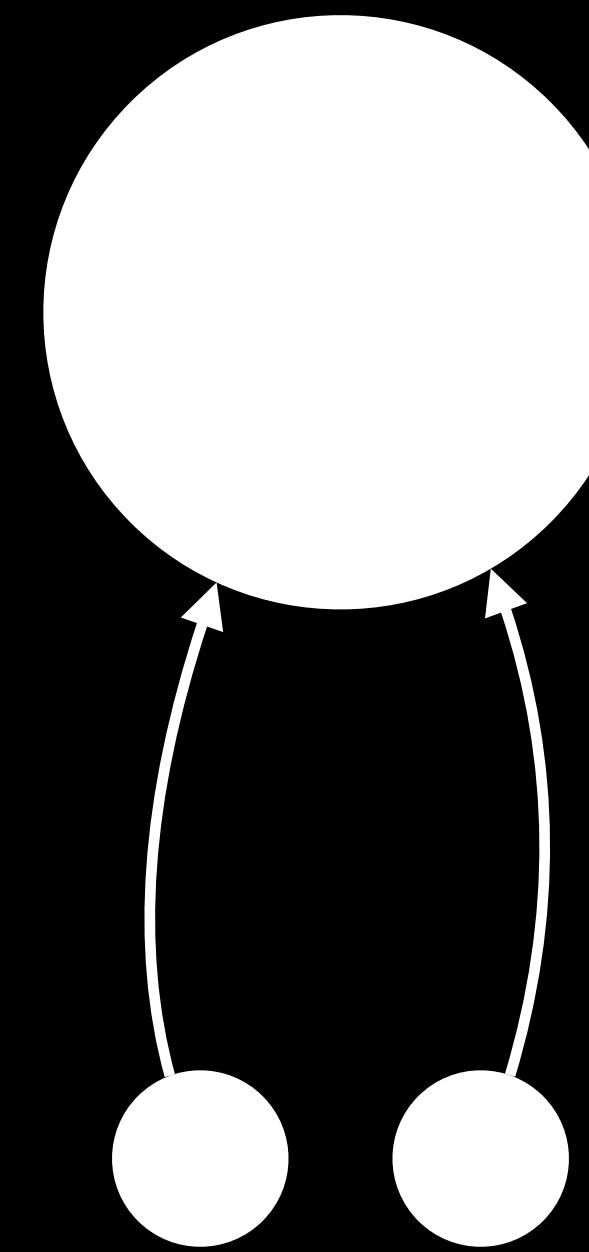


Pages with more in-links are more important

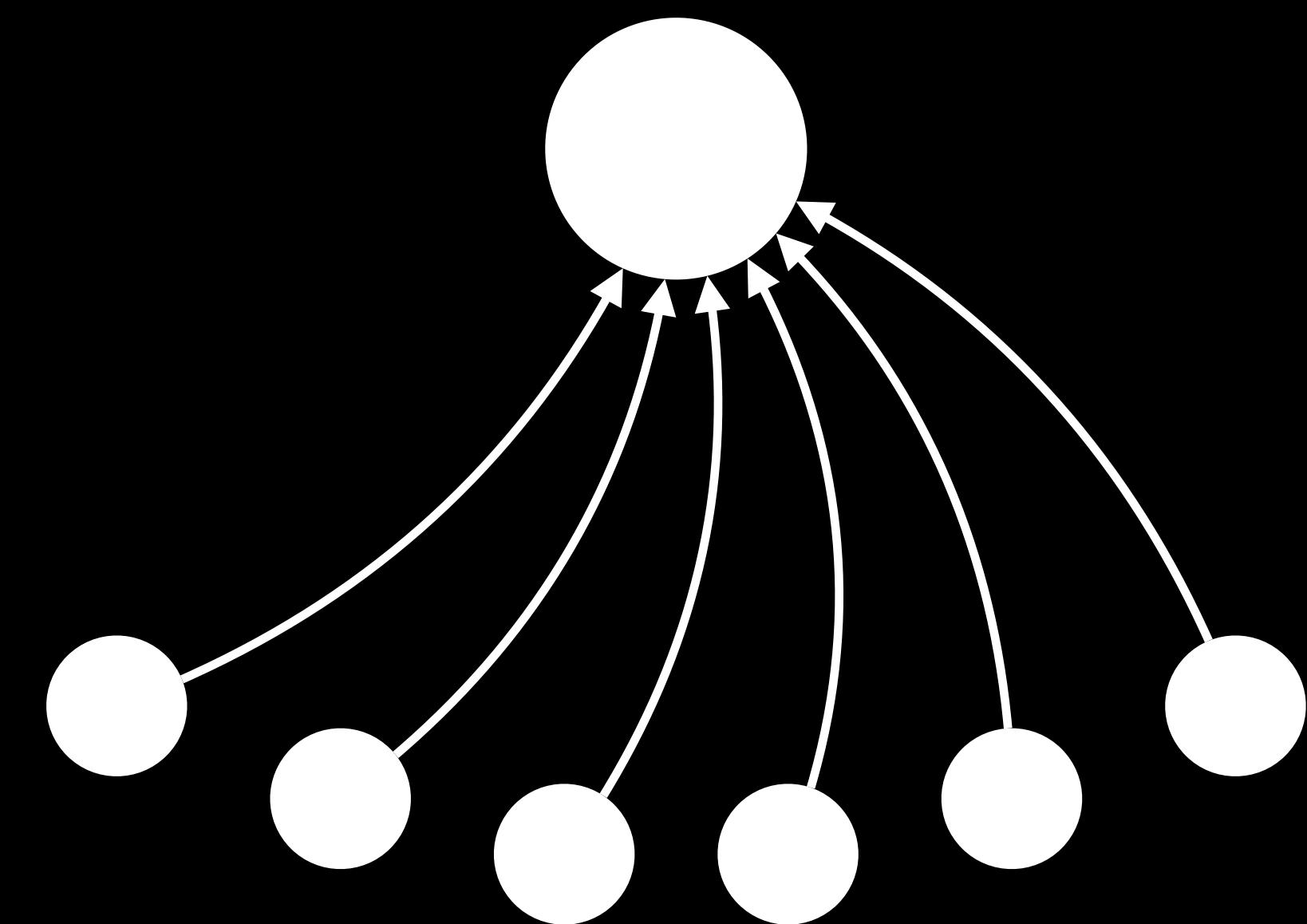
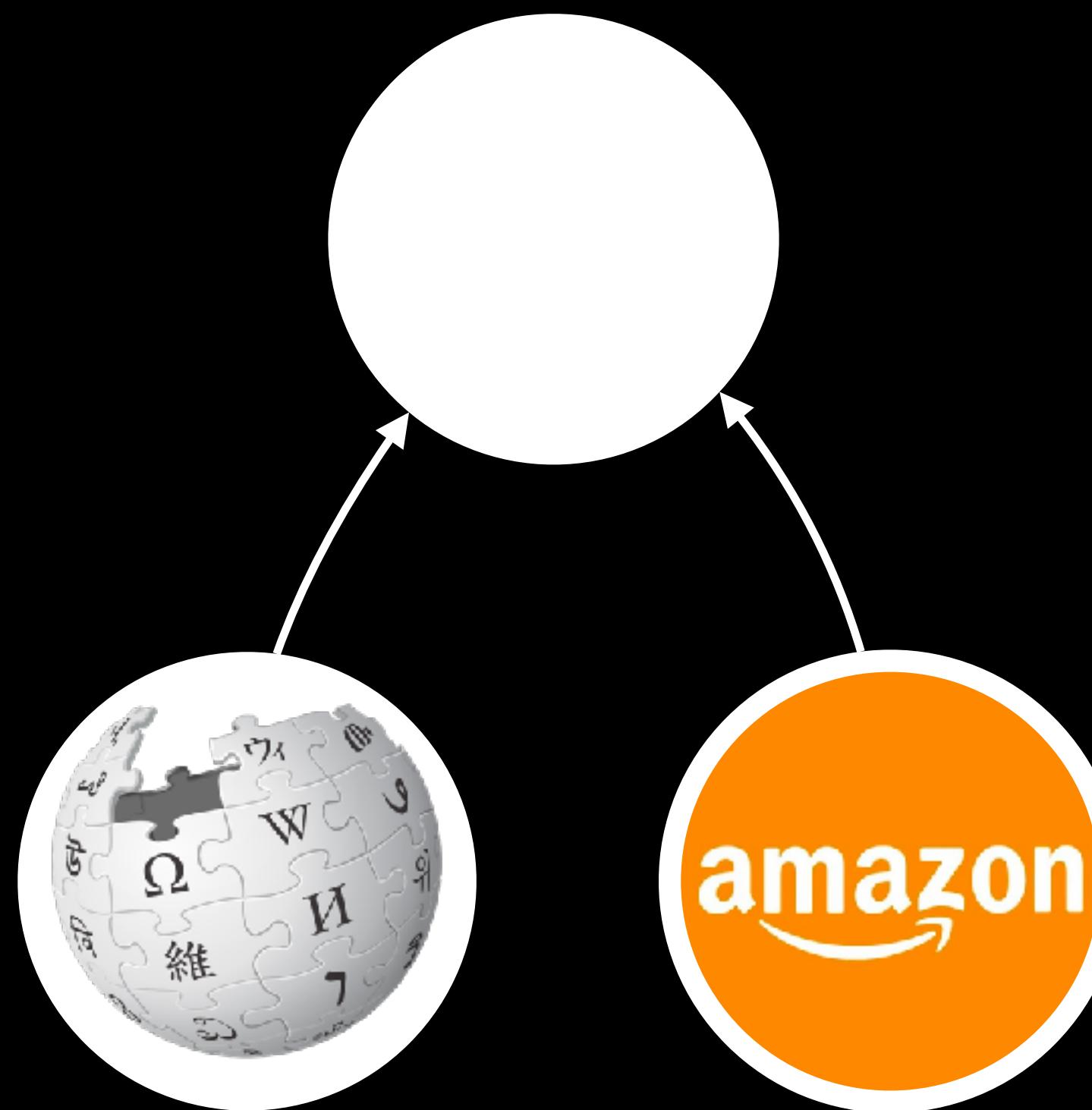
Intuition: Pages vote on each other via linking



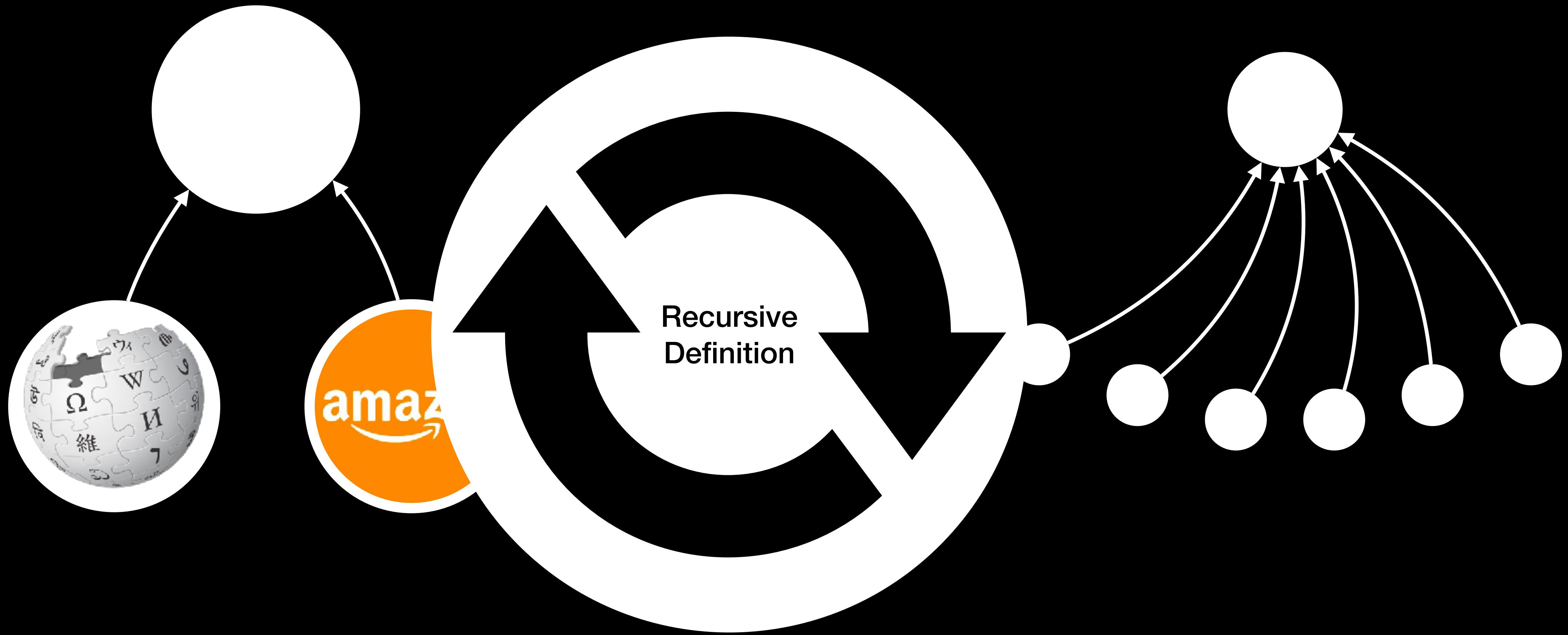
Addendum: Votes from important pages count more



Addendum: Votes from important pages count more



Addendum: Votes from important pages count more



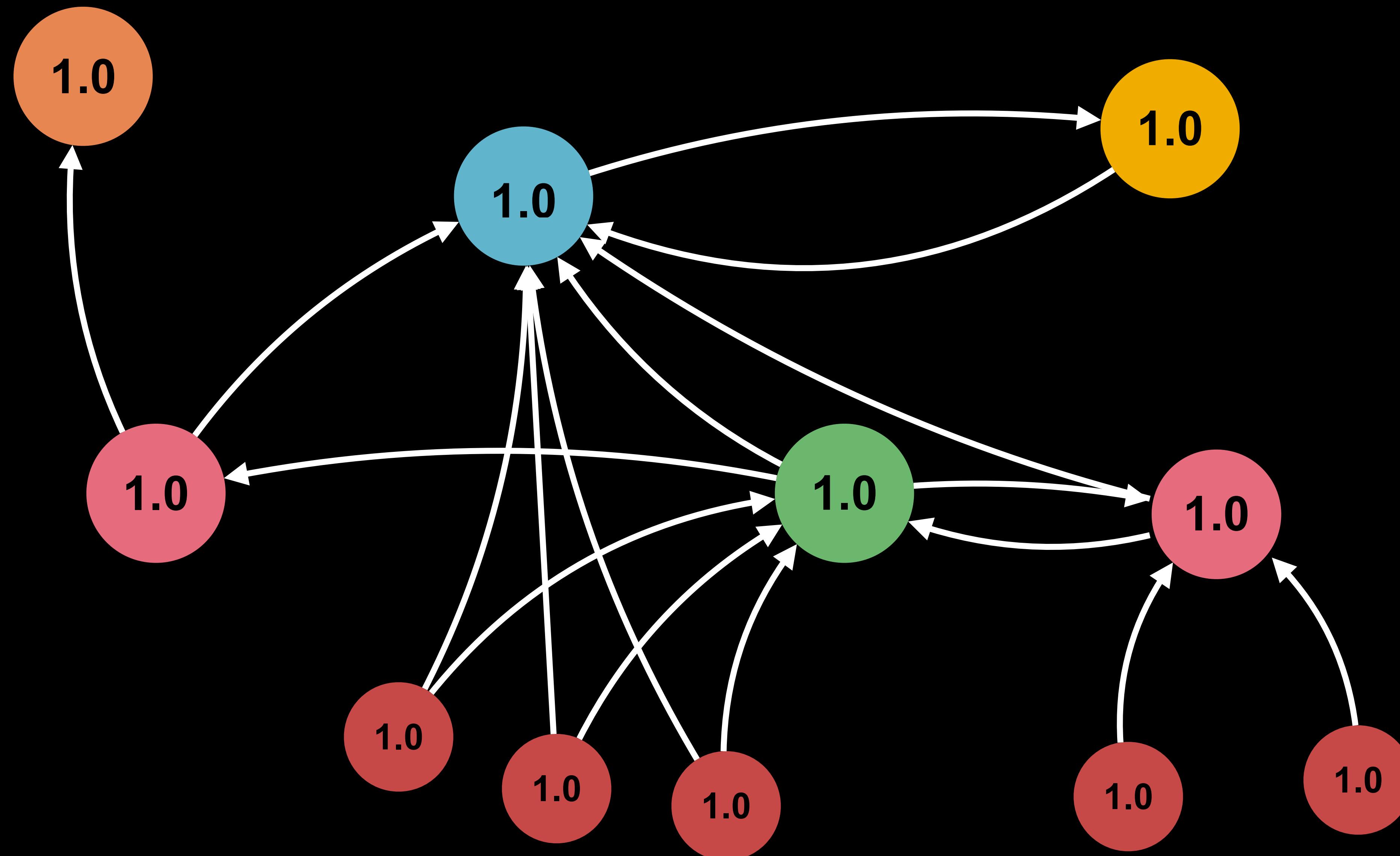
PageRank's Recursive Definition

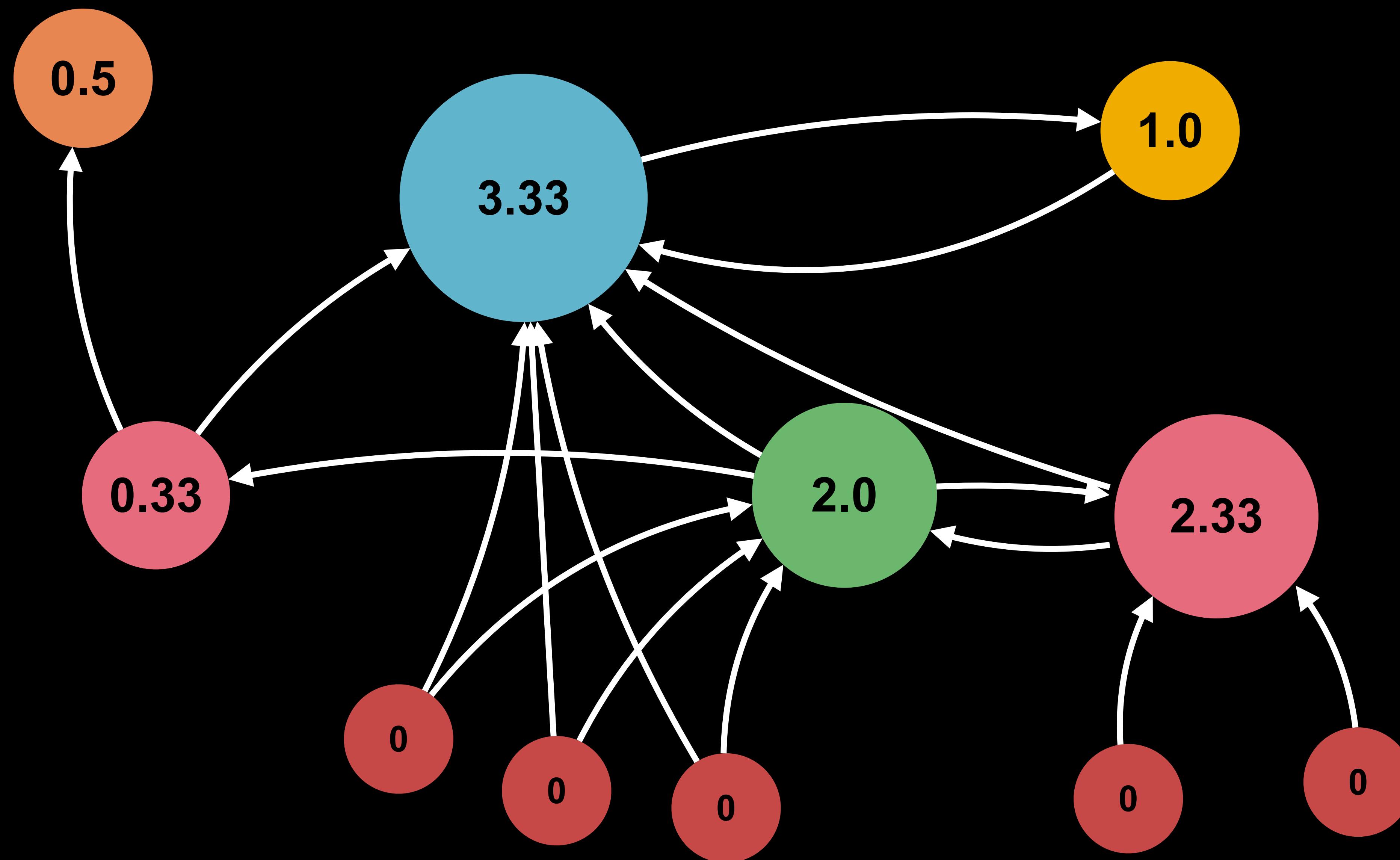
$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

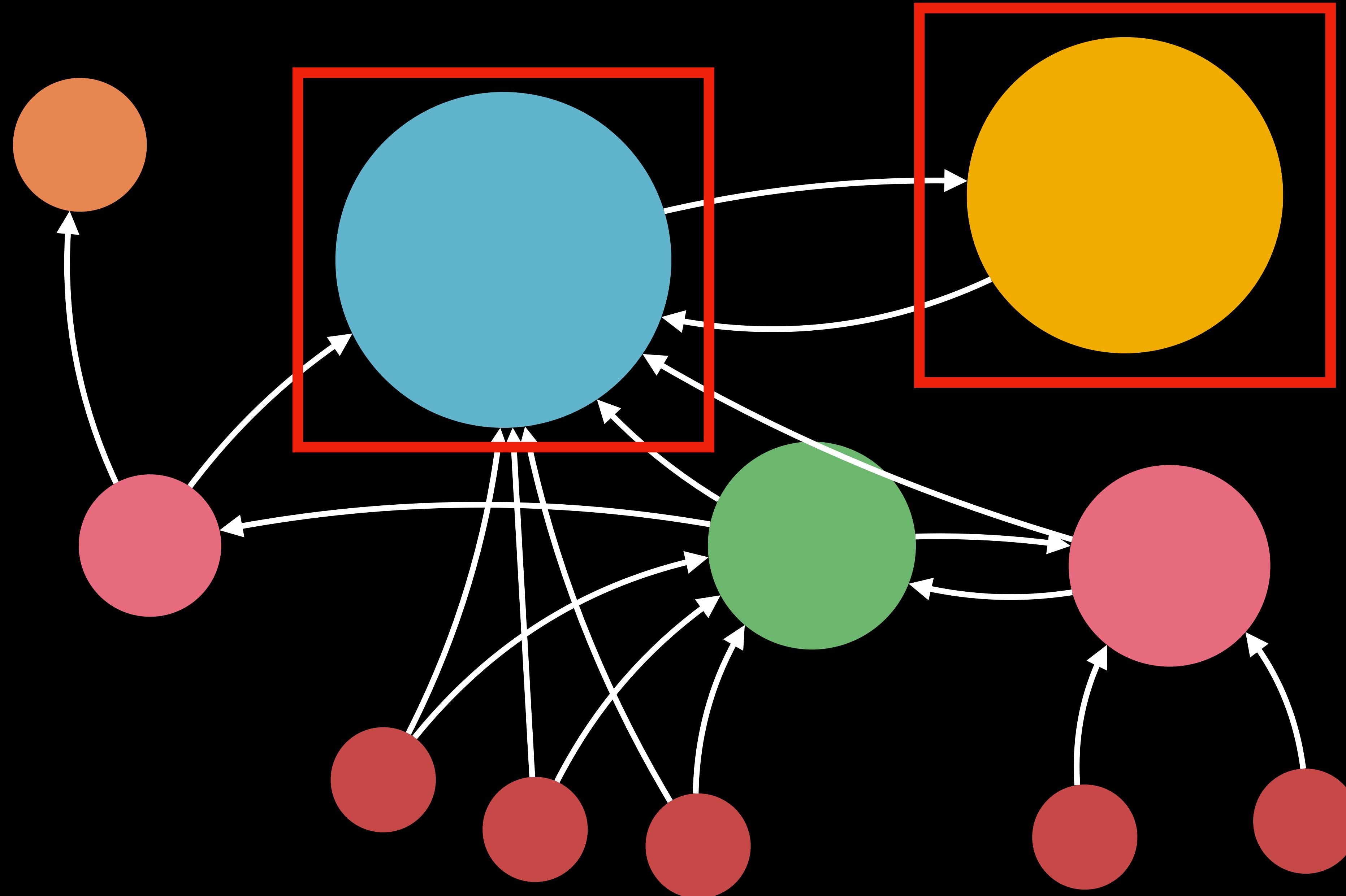
PageRank score r_j for page j

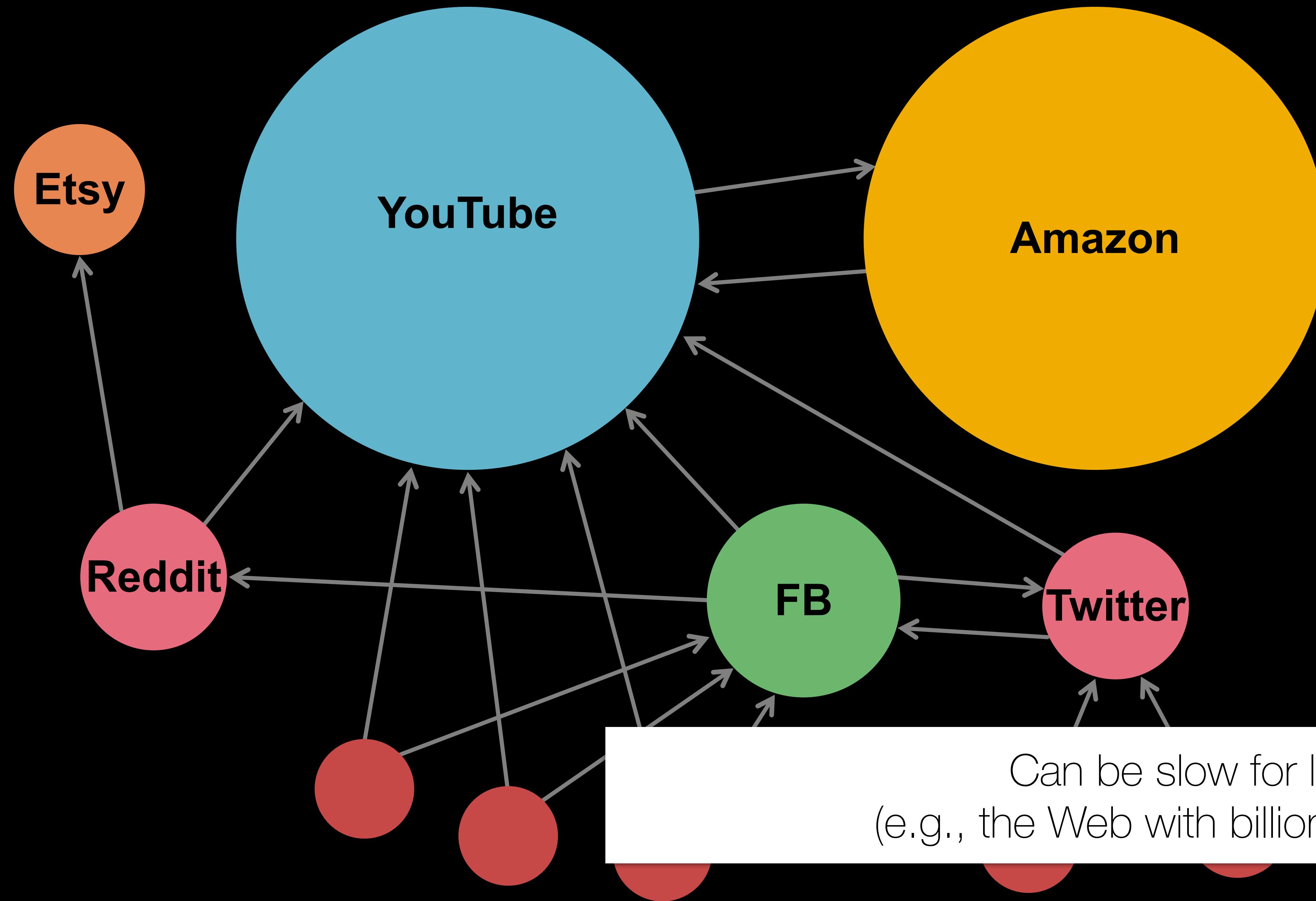
A node's rank can be distributed across its d_j outgoing links

Can be calculated in a distributed fashion via message-passing









What insights might this algorithm provide in other network contexts?

This Lecture's Learning Objectives

Construct graphs from web structures

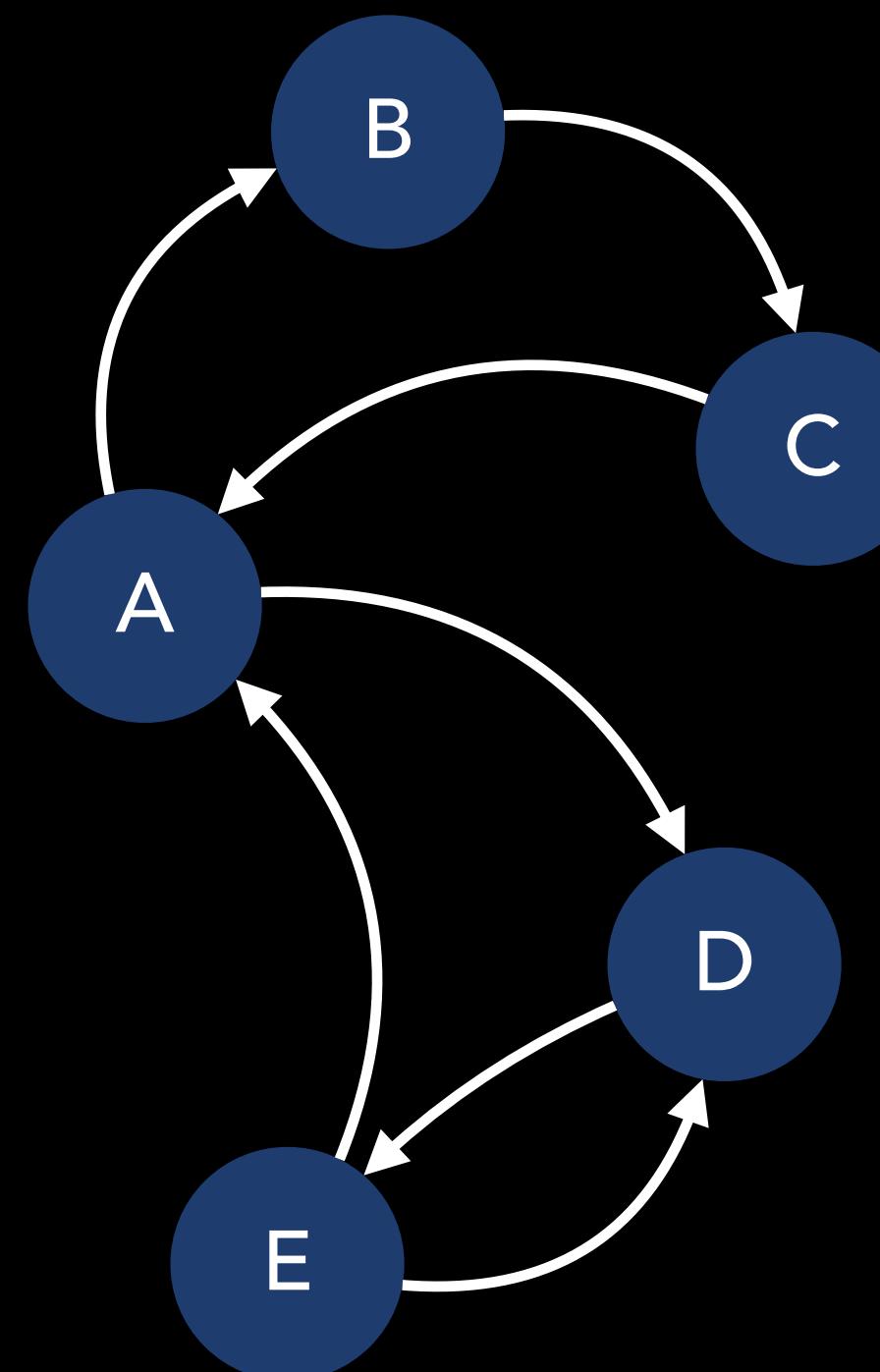
Describe at least two ways how one may try to exploit search engines

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

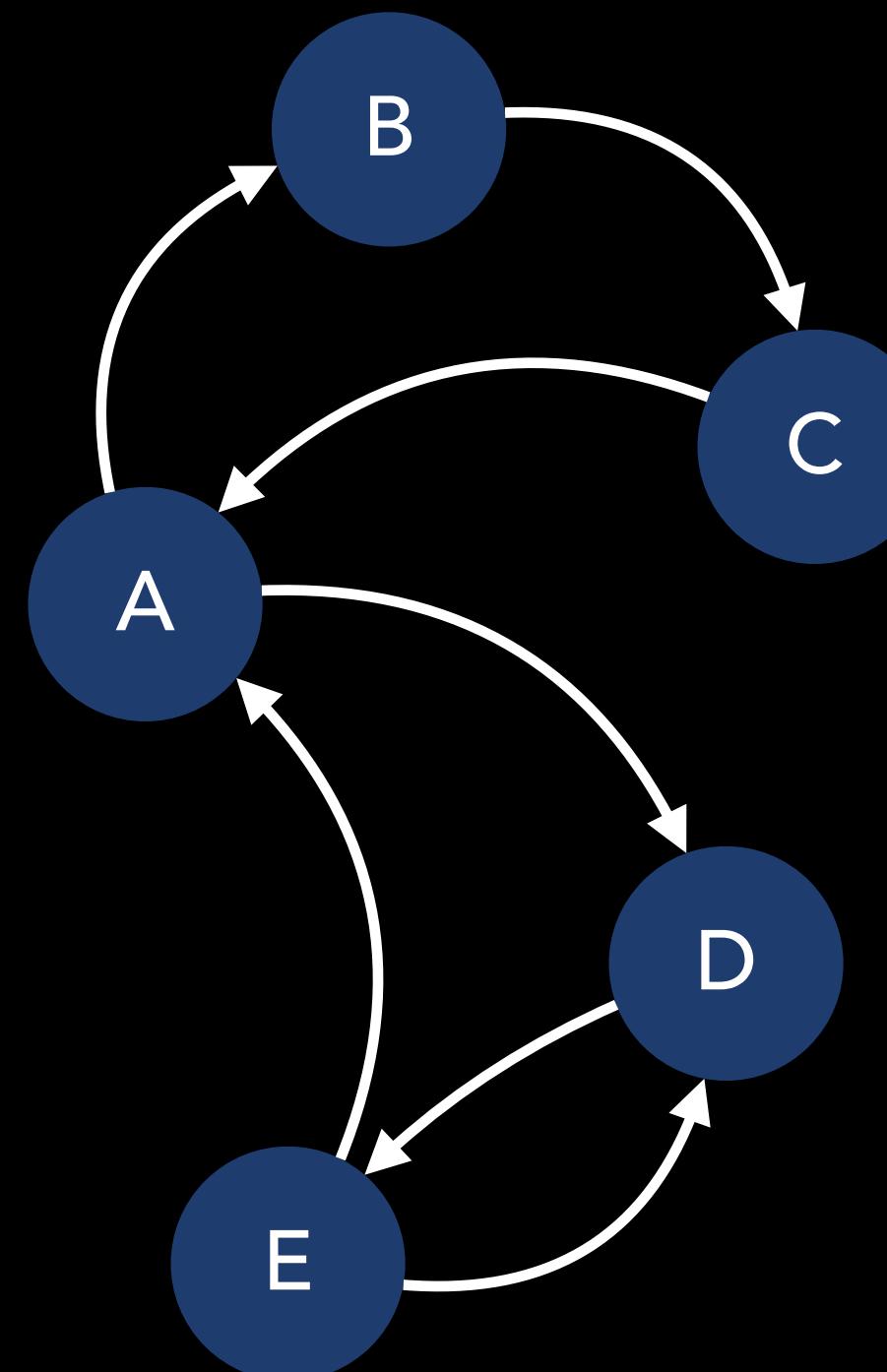


Adjacency Matrix $A =$

	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



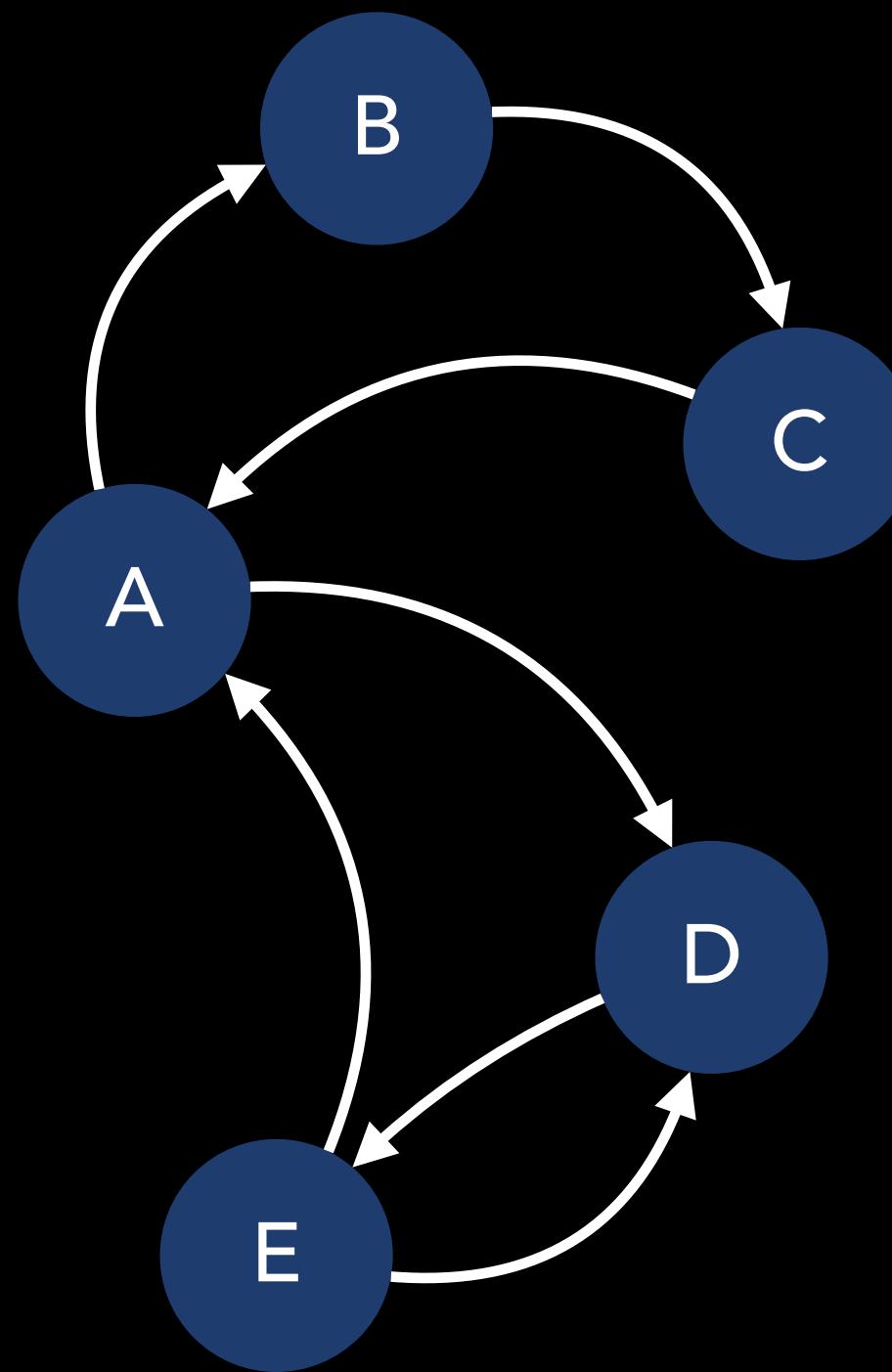
Transition Matrix $M =$

	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



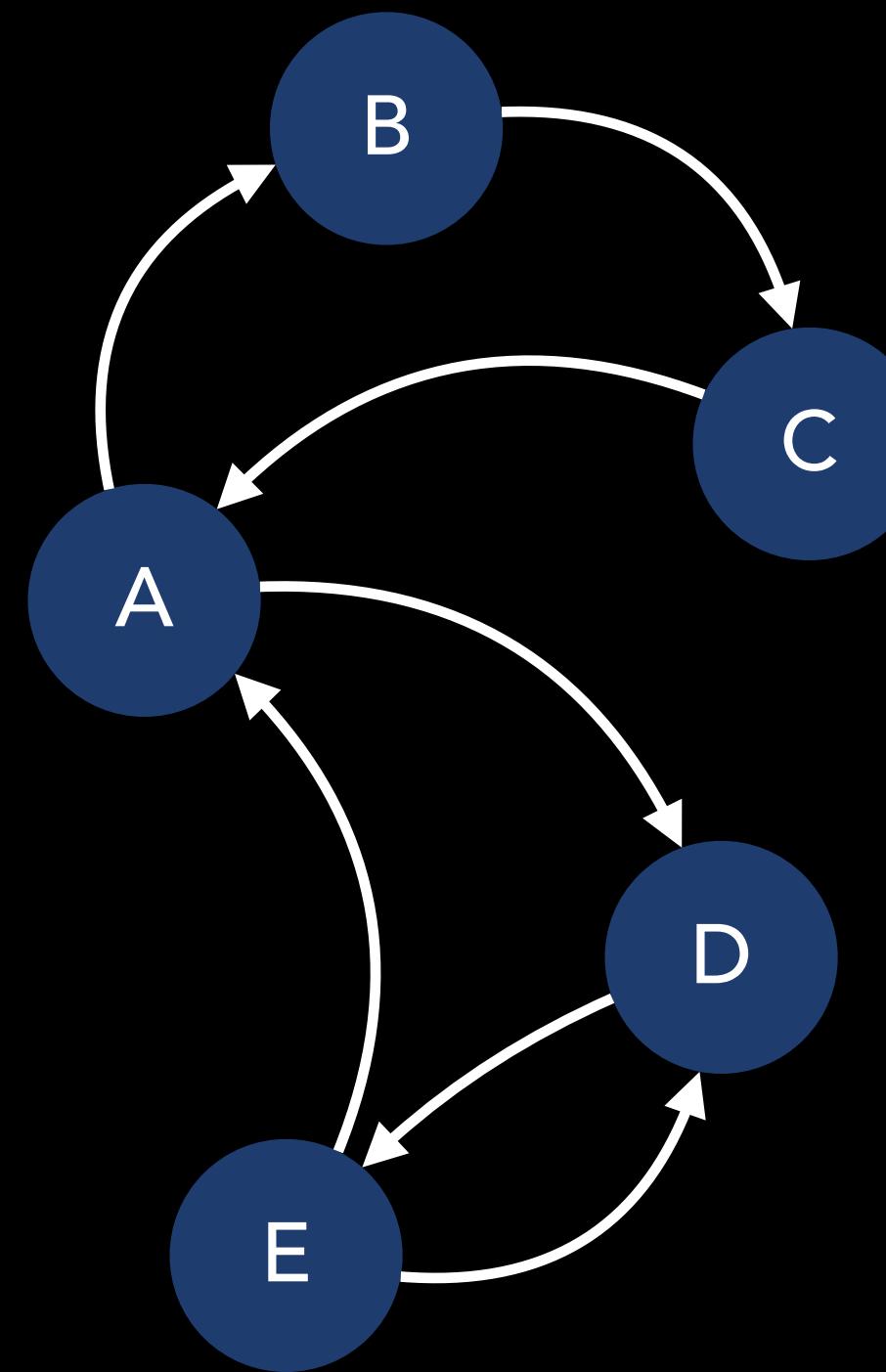
Transition Matrix $M =$

	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

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

An Alternative PageRank Definition



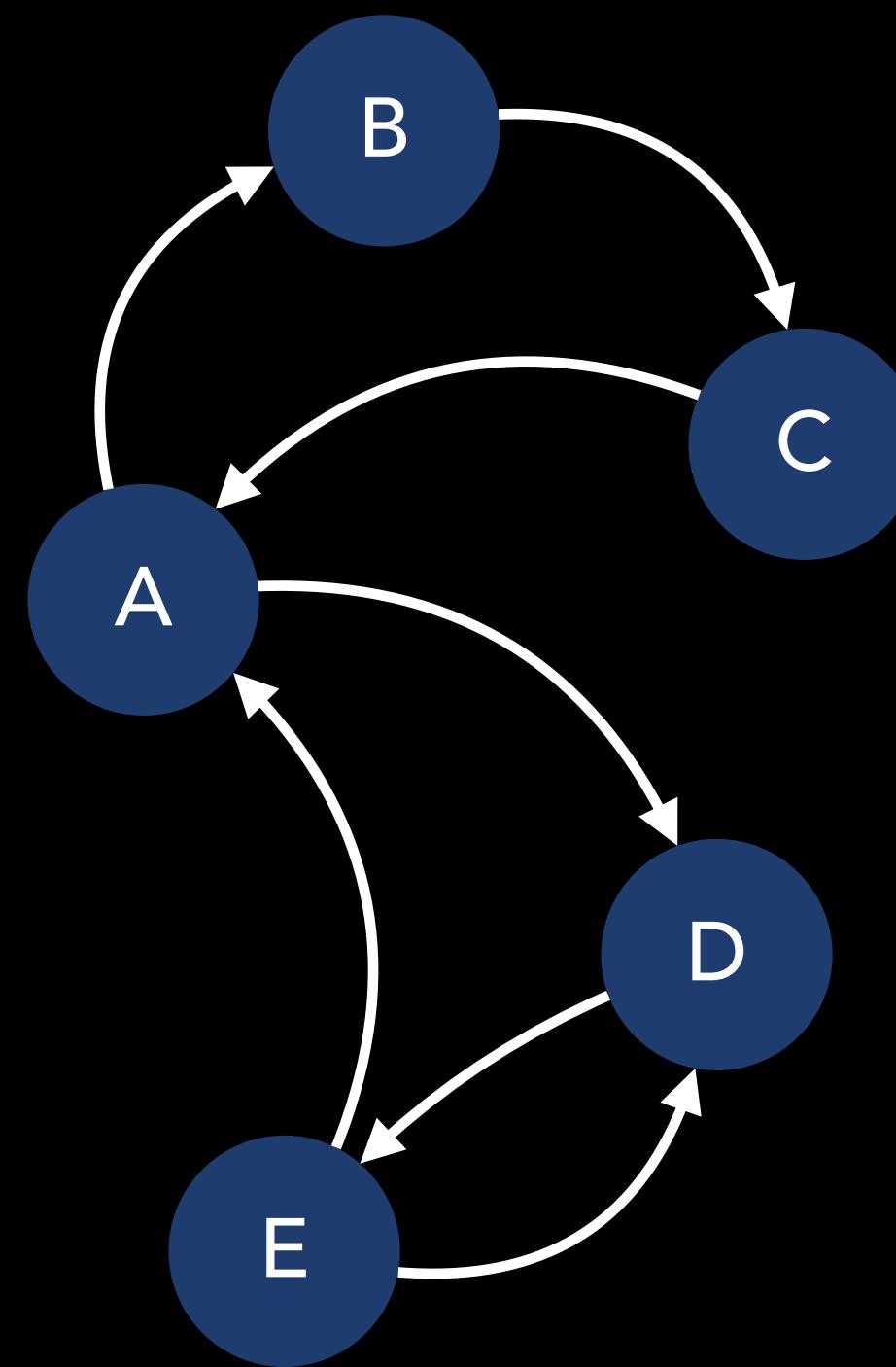
Transition Matrix $M =$

	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

$$r(\text{ank}) = M r$$

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

$$r(\text{rank}) = M r$$

$$\lambda r = M r$$

An Alternative PageRank Definition

PageRank scores then
are given in the largest
eigenvector of M

Can be found via
Power Iteration

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

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

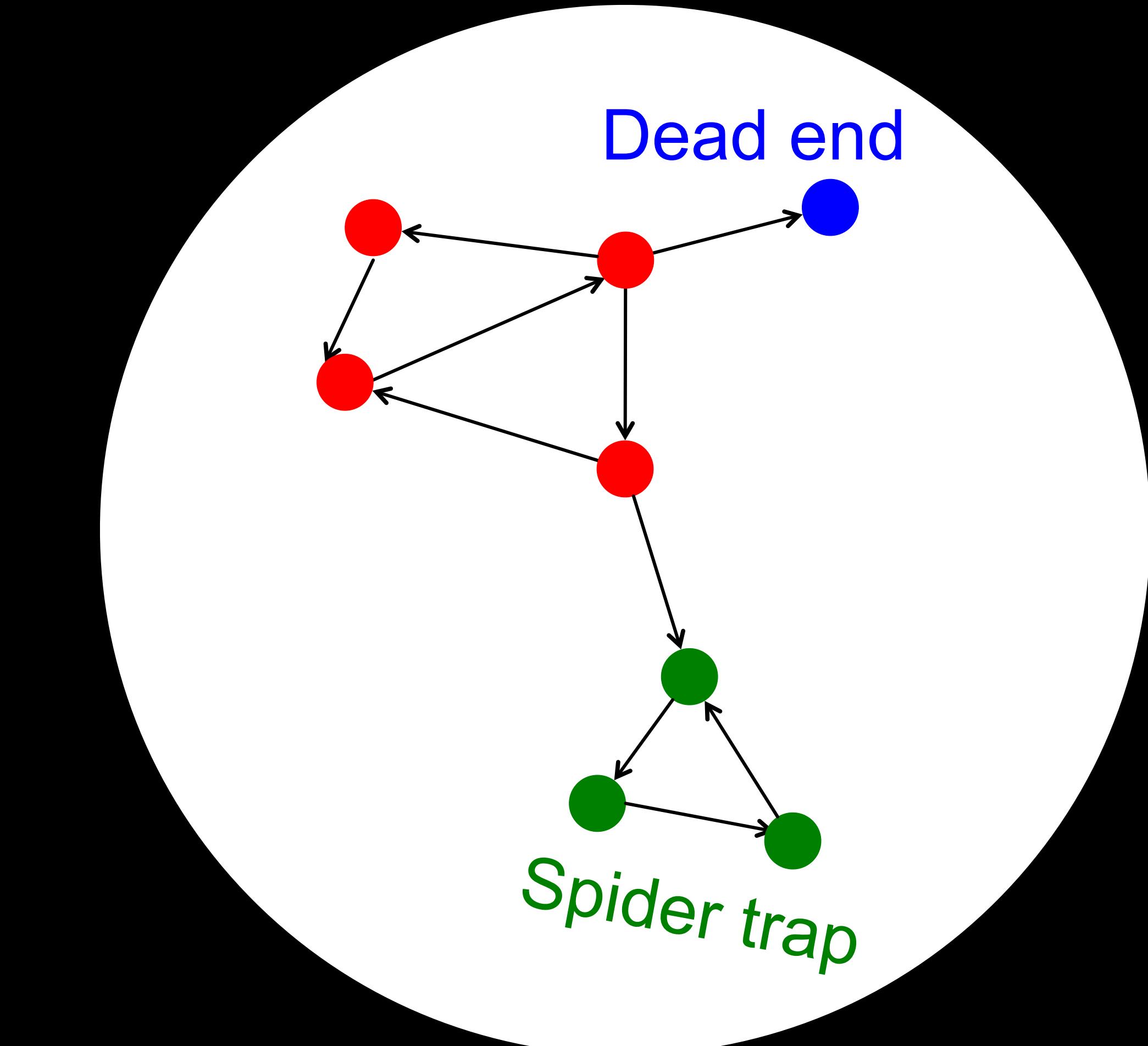
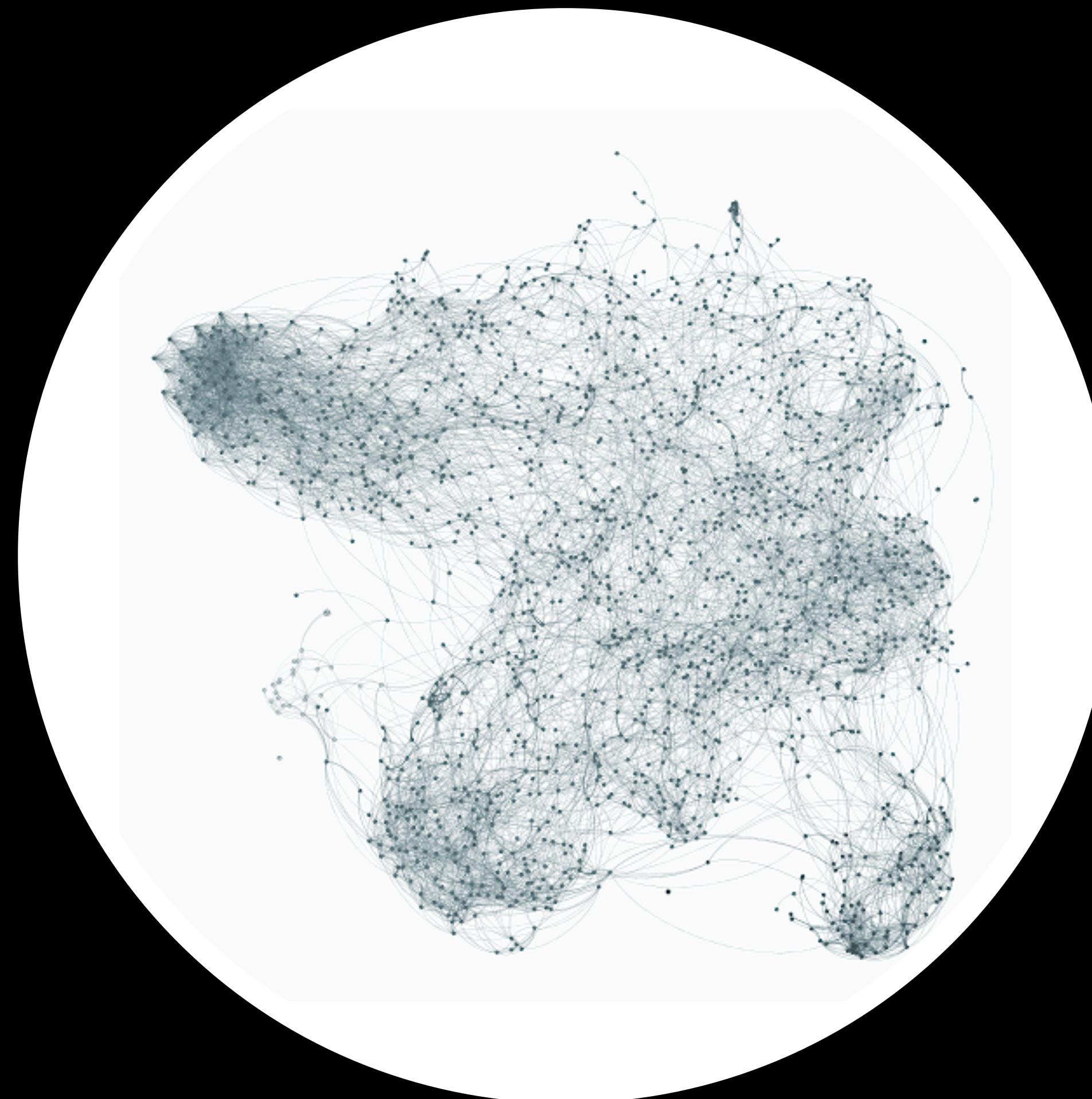
Transition Matrix M =

	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

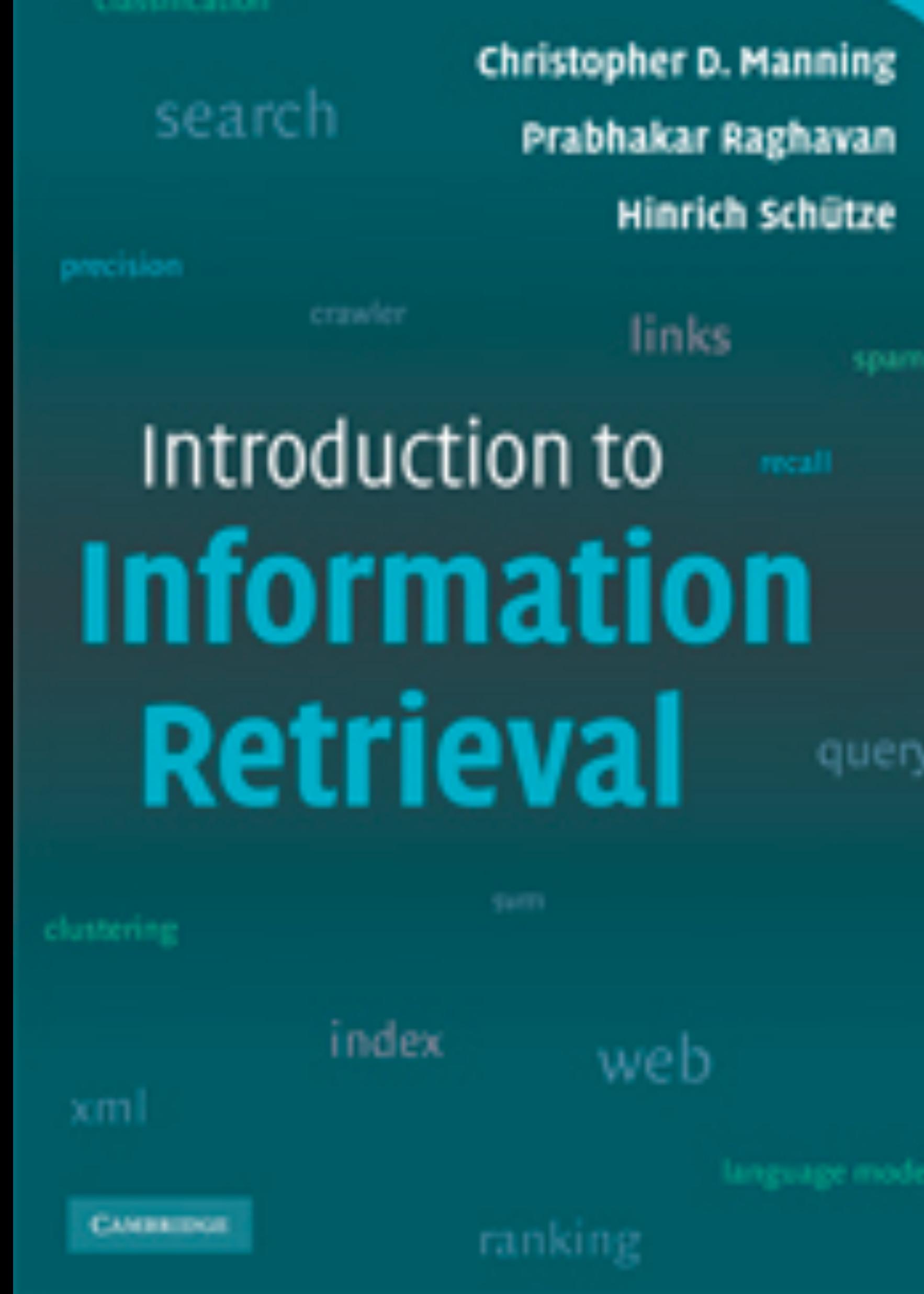
Vector r is an eigenvector
of the transition matrix M

$$r(\text{ank}) = M r$$

This formulation works well for
strongly connected graphs



Many extensions to PageRank to deal with
these irregular structures



Query: Amanda Gorman

Amanda Gorma

From Wikipedia, the free encyclopedia

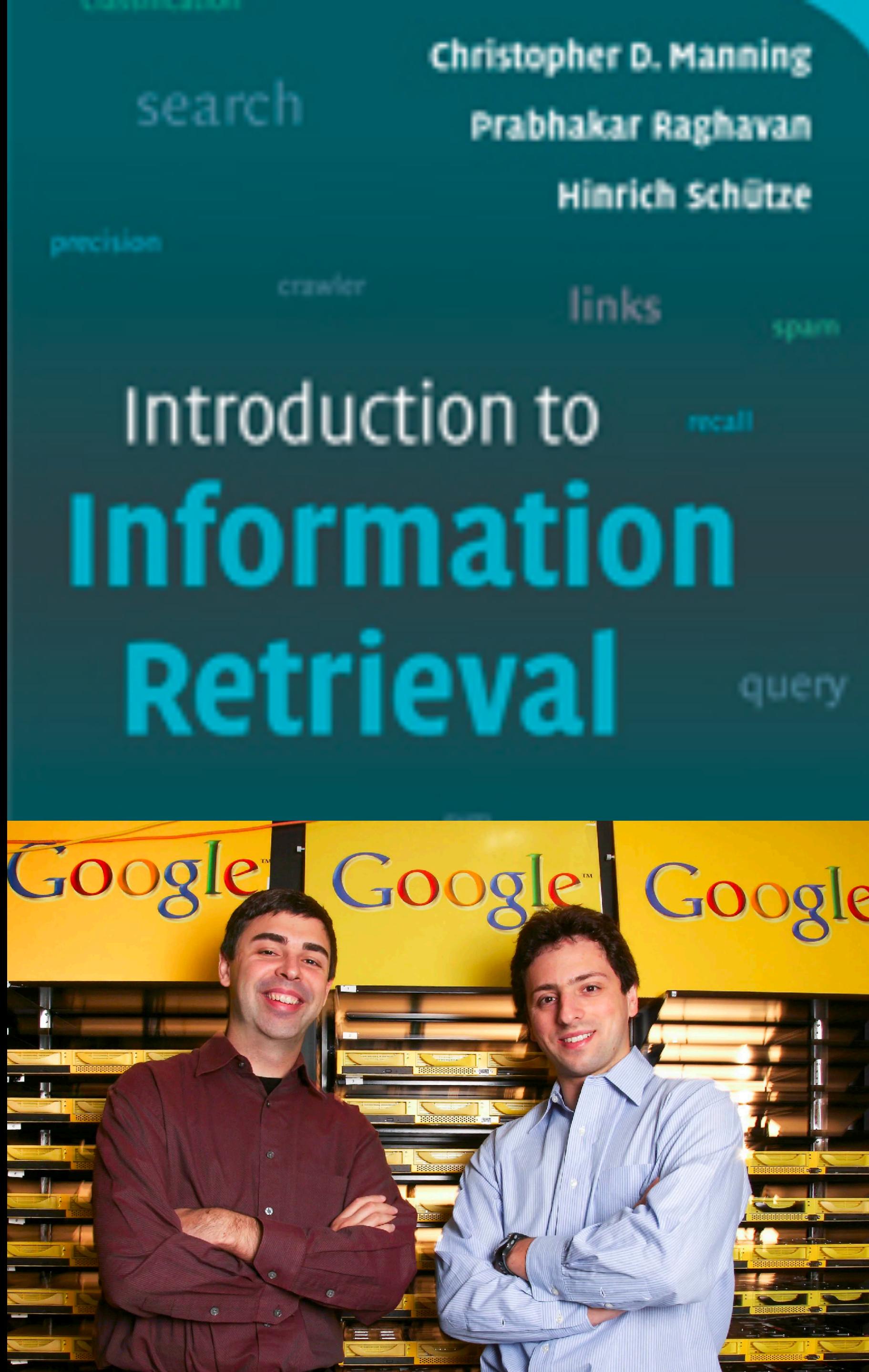
Amanda S. C. Gorman^[1] (born 1998) is an American poet and activist. Her work focuses issues of oppression, feminism, race, and marginalization, as well as the African diaspora. Gorman was the first person to be named National Youth Poet Laureate. She published her poetry book *The One for Whom Food 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 and stimulated her two books to reach best-seller status.

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 one could have won her invitations to the Obamas' White House.





Query: Amanda Gorman

Largest 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 bestseller status, and earned her a professional management contract.



Next r

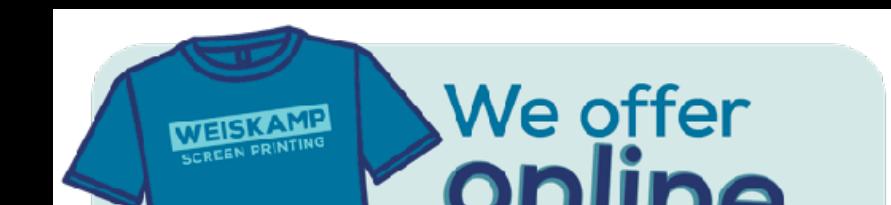
Wordsmith. Change-maker.

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, she has won her invitations to the White House and to perform for Lin-Manuel Miranda, Al Gore, Secretary Hillary Clinton, Malala Yousafzai,



Small r



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)

Open Problems with PageRank

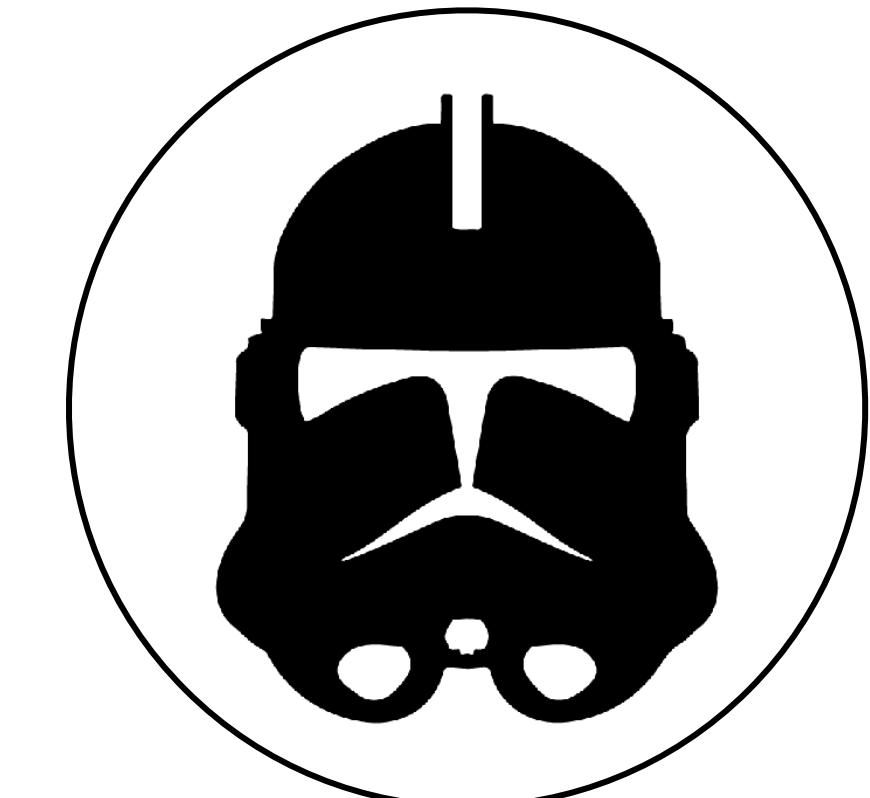
Measures generic popularity

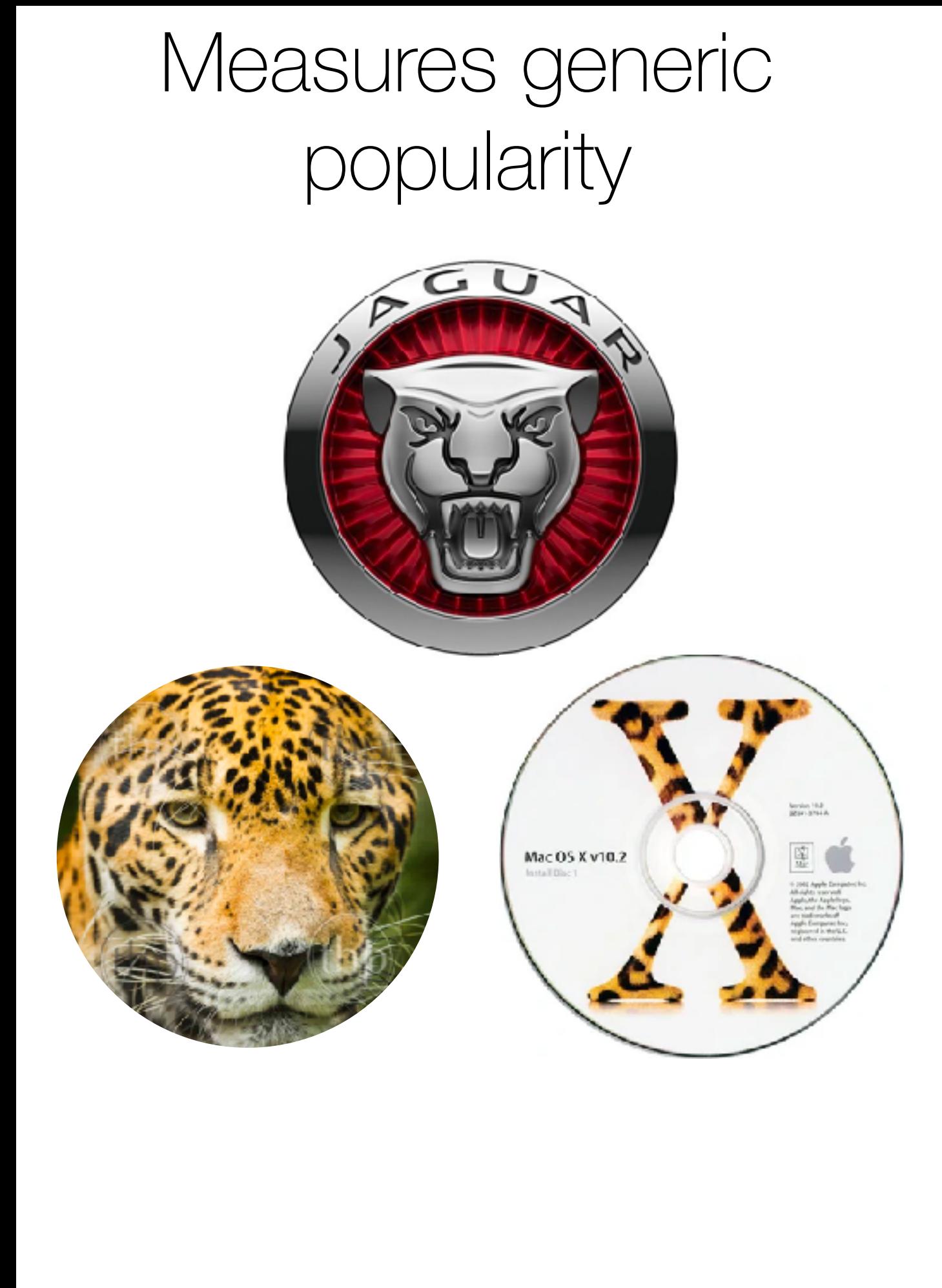


Only one measure of importance



Can be manipulated (though harder)





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

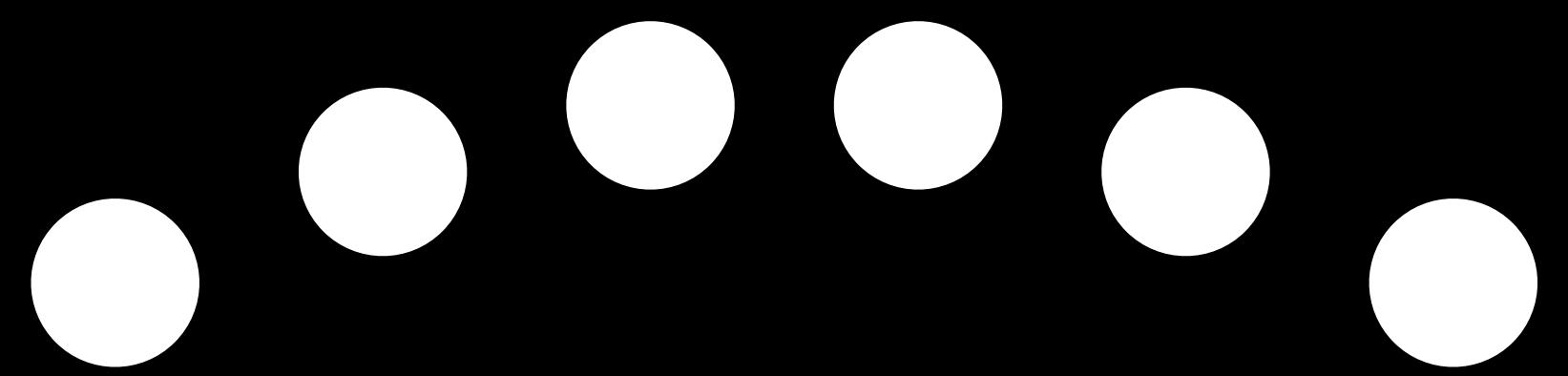
Construct graphs from web structures

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

Explain how PageRank captures “importance” in a graph structure

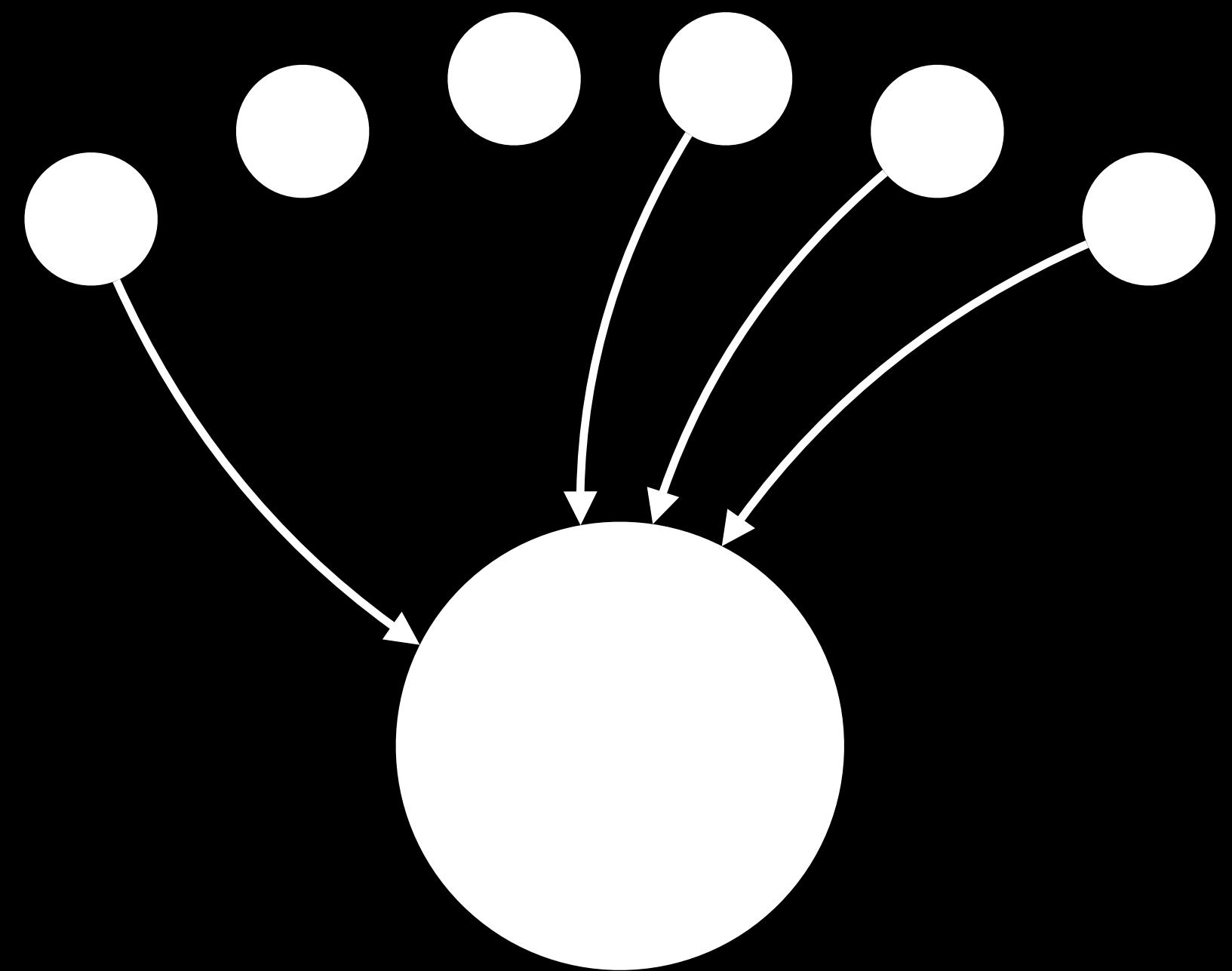
Construct a transition matrix to model a collection of web pages

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”

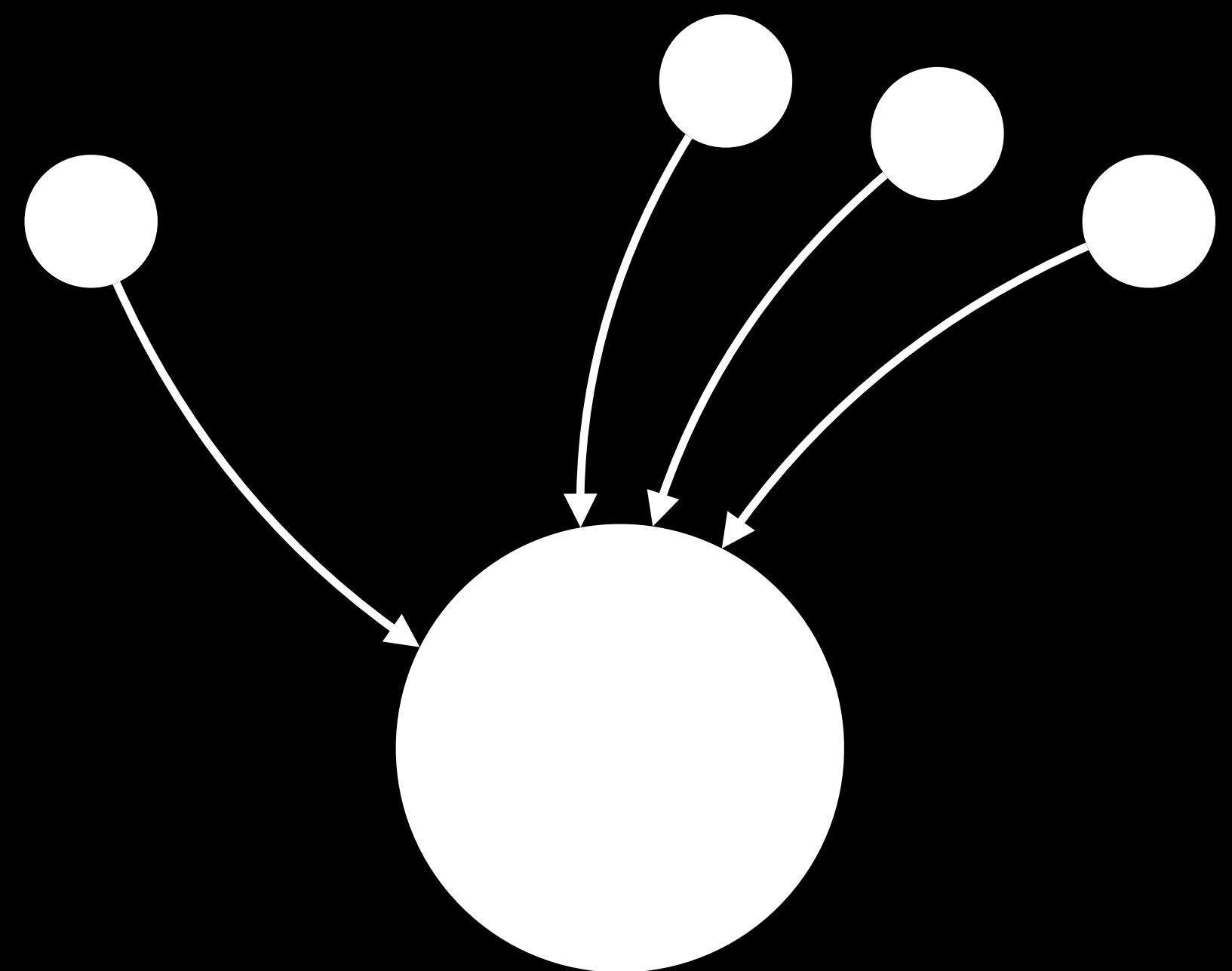
Pages that refer you to this common site are “hubs”



HITS vs. PageRank

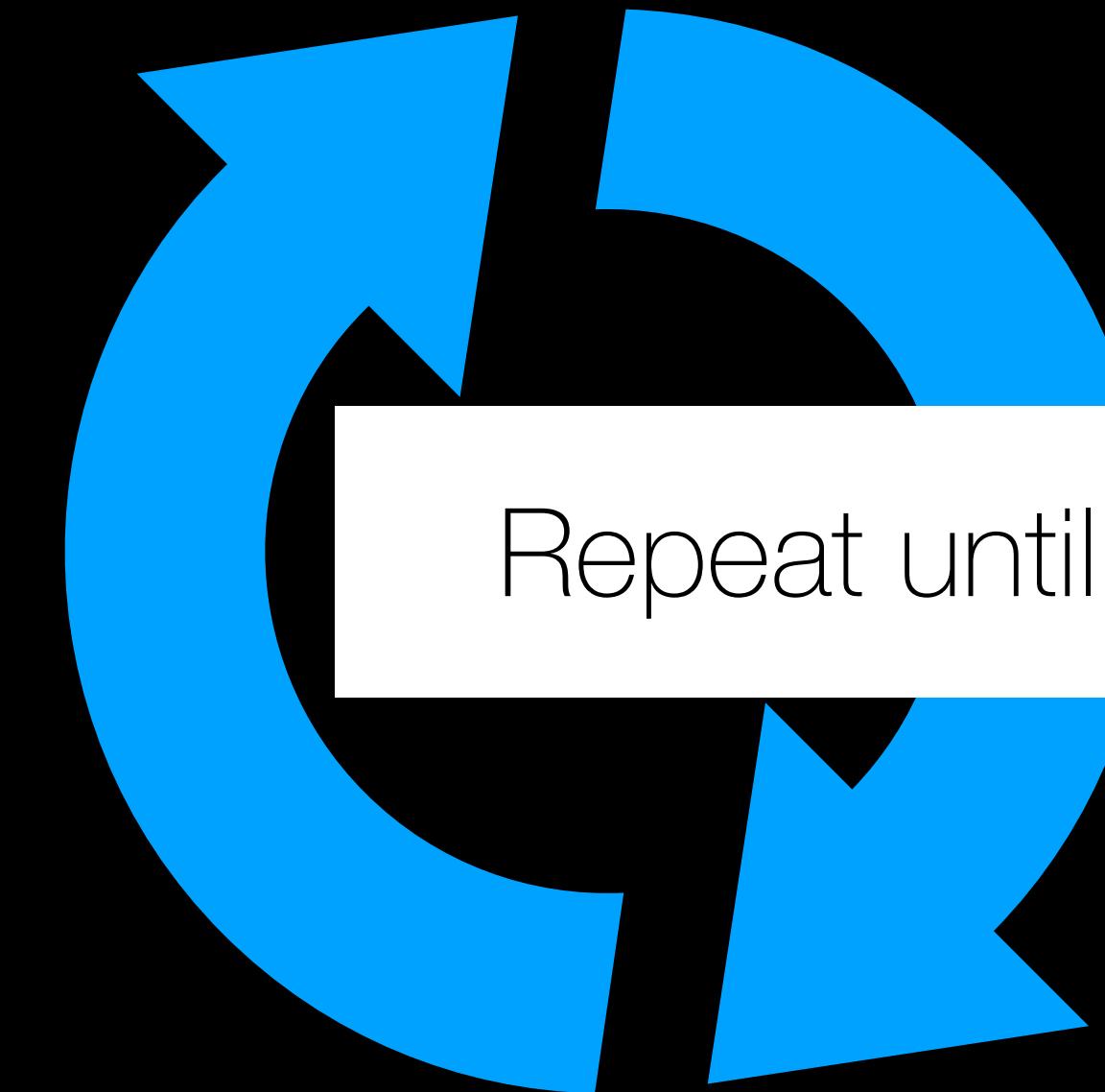


Each node receives separate hub- and authority-scores



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