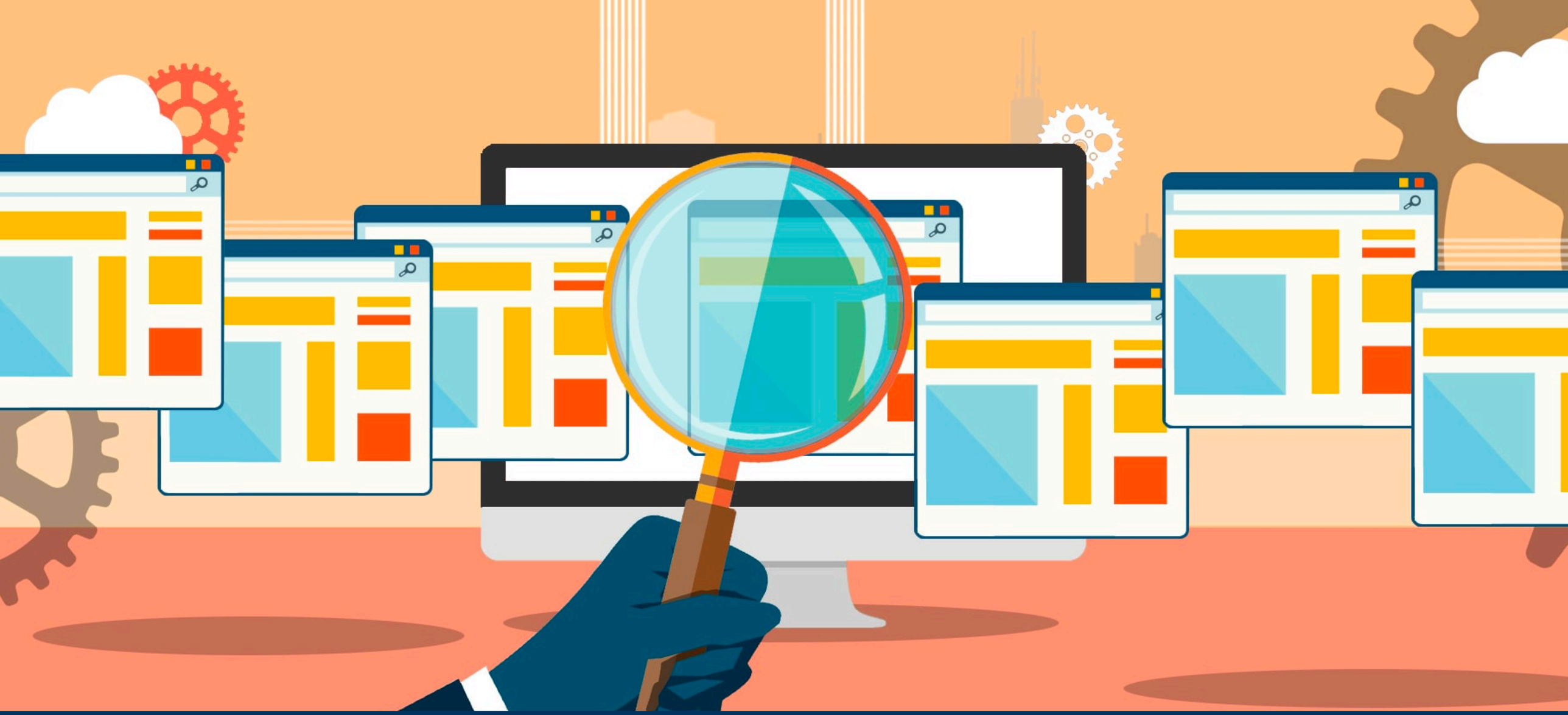


# Centrality in Graphs

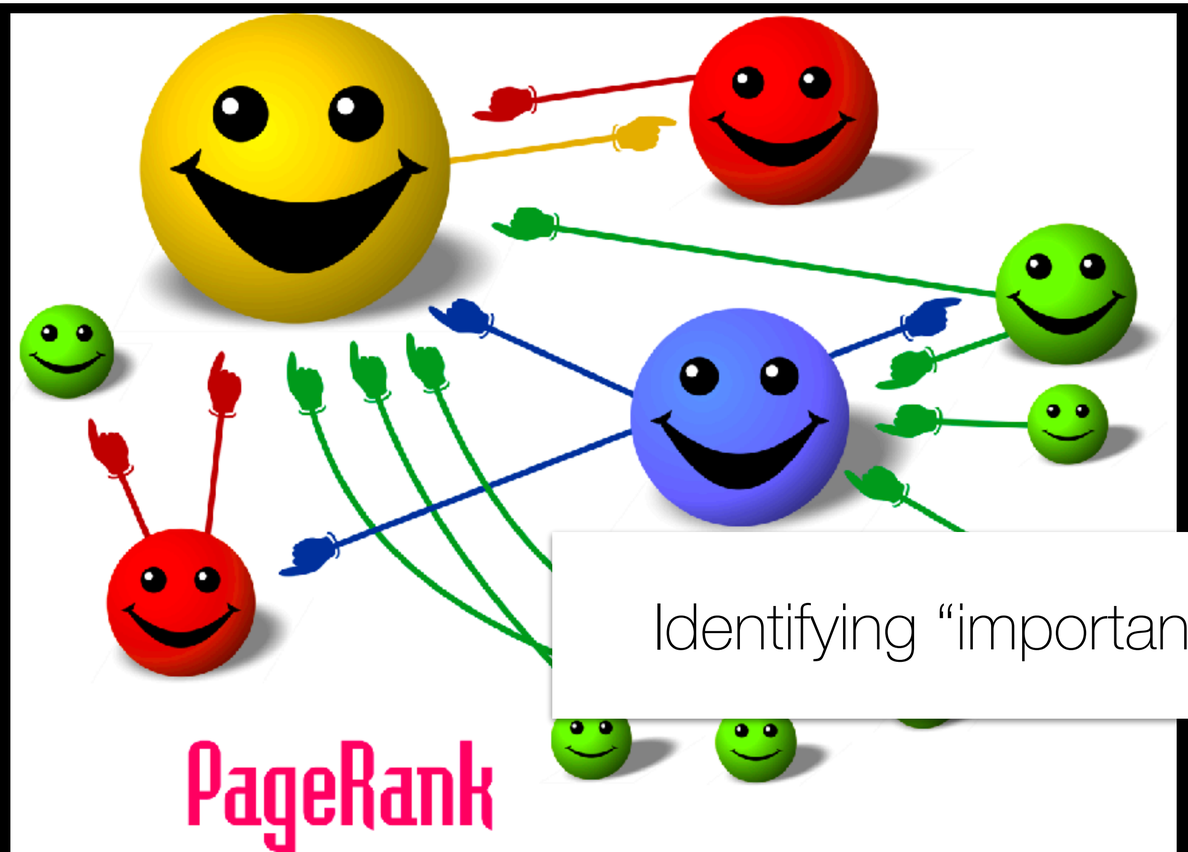
INST414 - Data Mining Techniques

# 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?

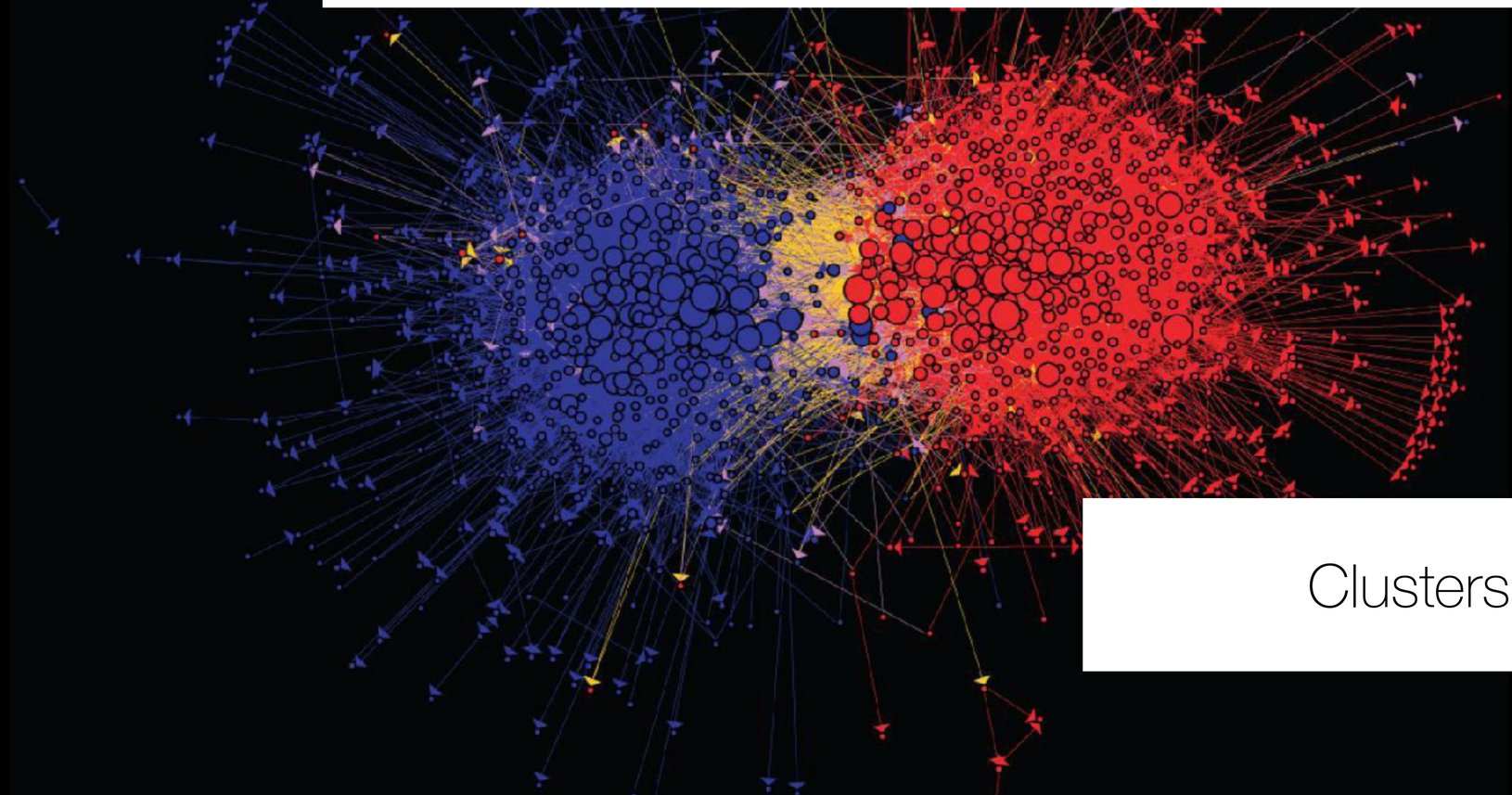
2. Articulate the math behind supervised and unsupervised techniques




Identifying “important” elements in a graph

# Where are we?

3. Execute supervised and unsupervised machine learning techniques



Clusters, or “communities”, in graphs

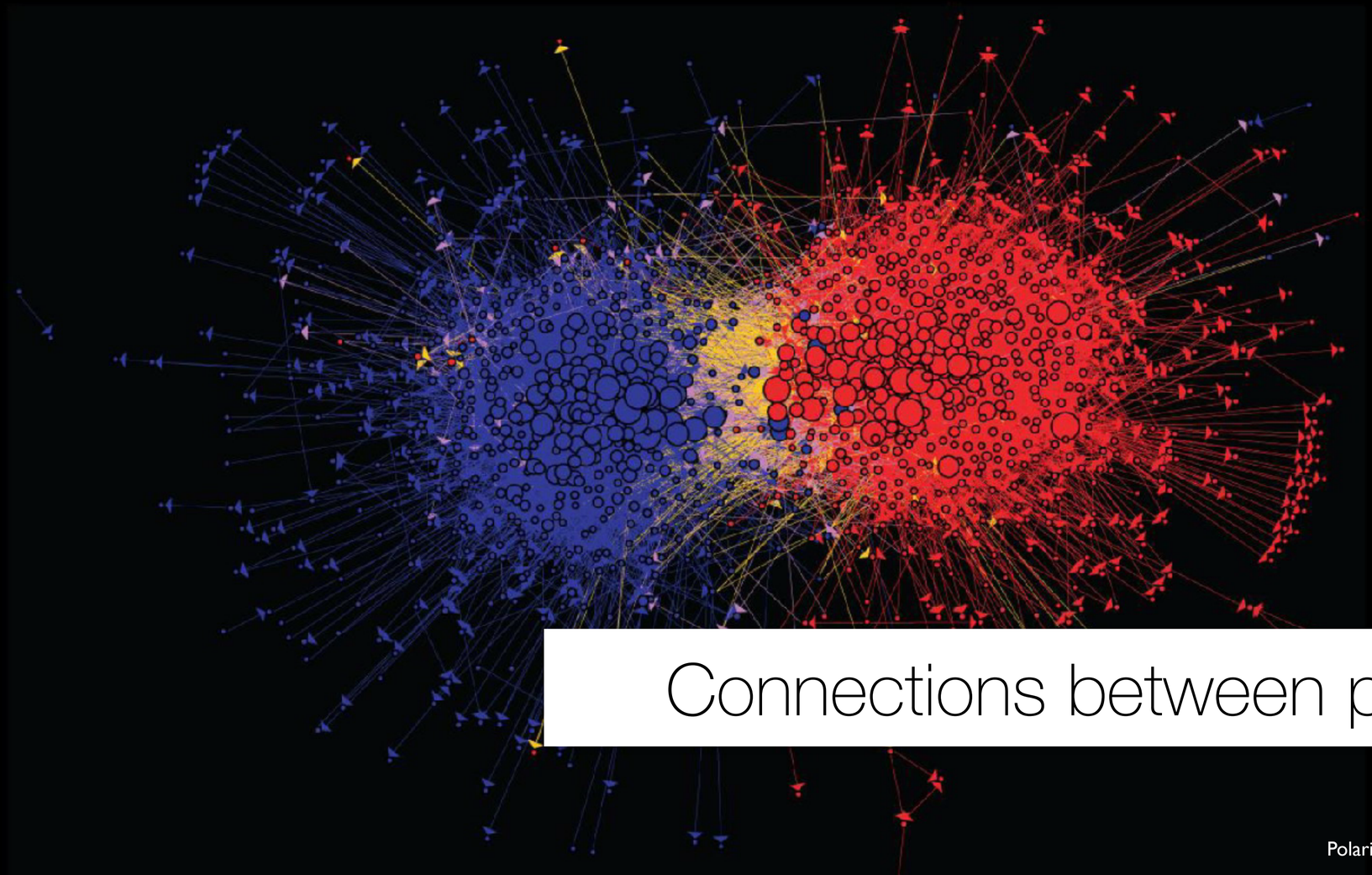
# This Lecture's Learning Objectives

Describe at least three network centrality metrics

Use the Girvan-Newman method to identify communities in graphs

The web is full of networks

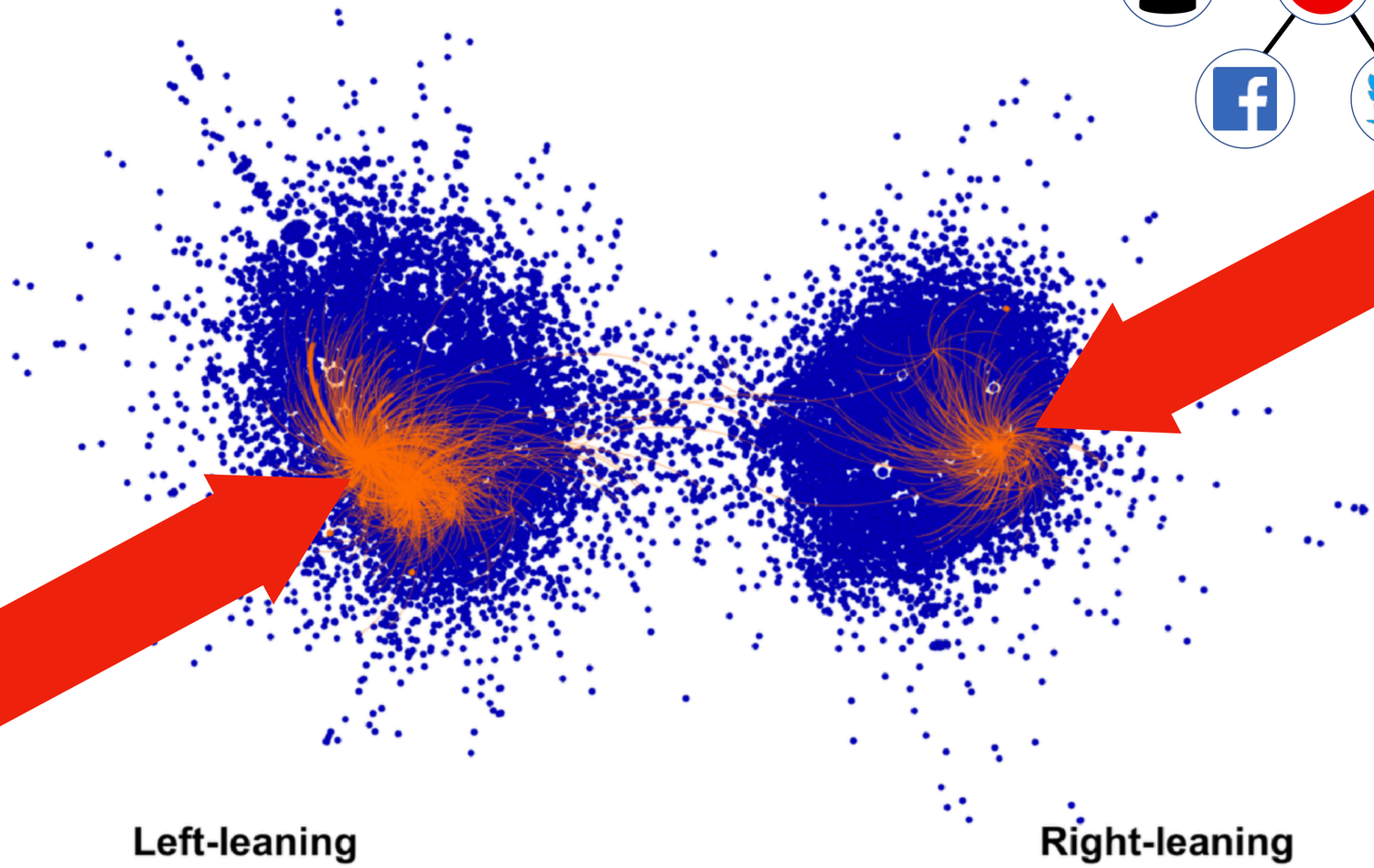
# Graphs in Media Networks



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

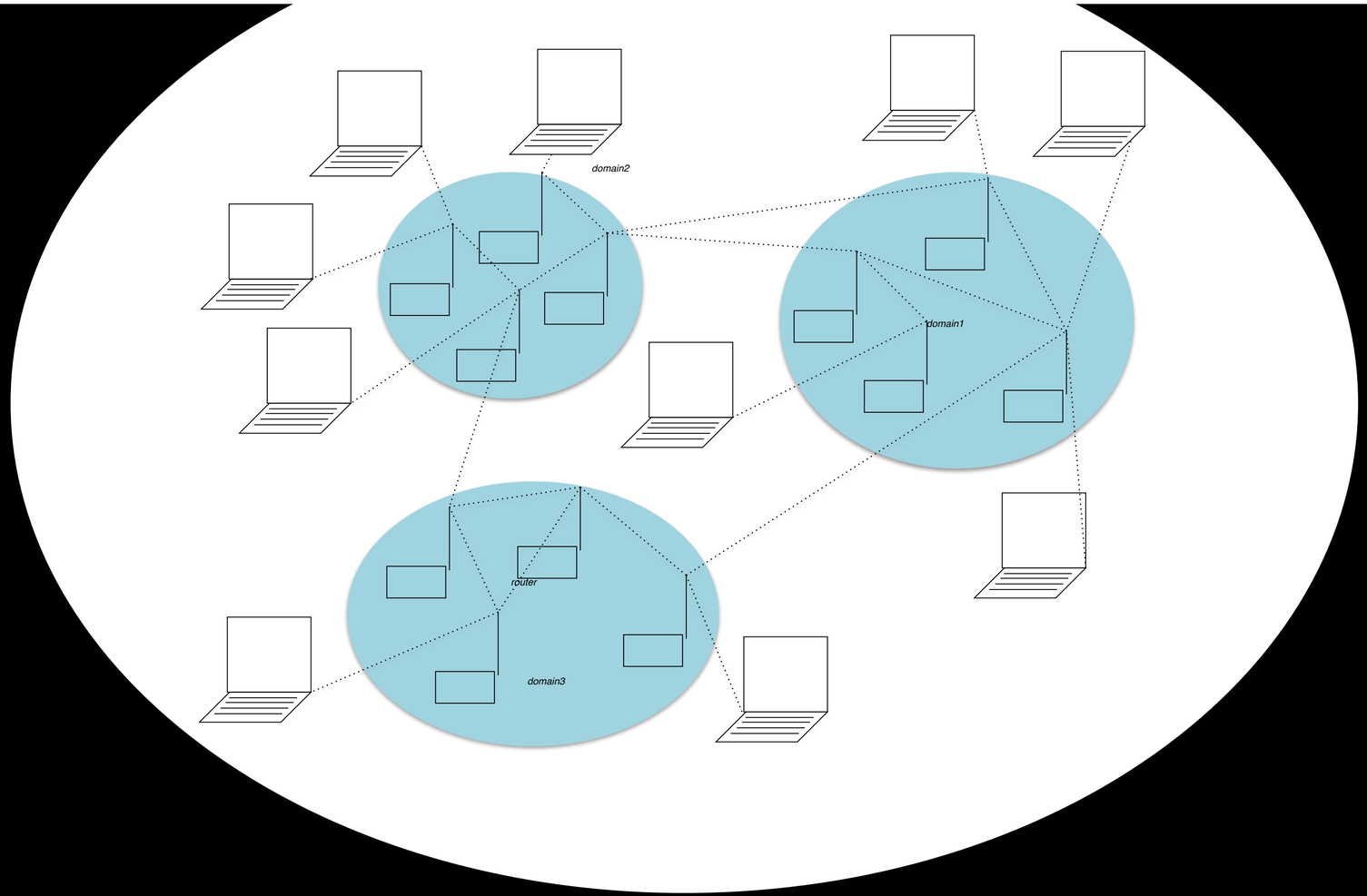
with the IRA accounts in orange and other accounts in blue.



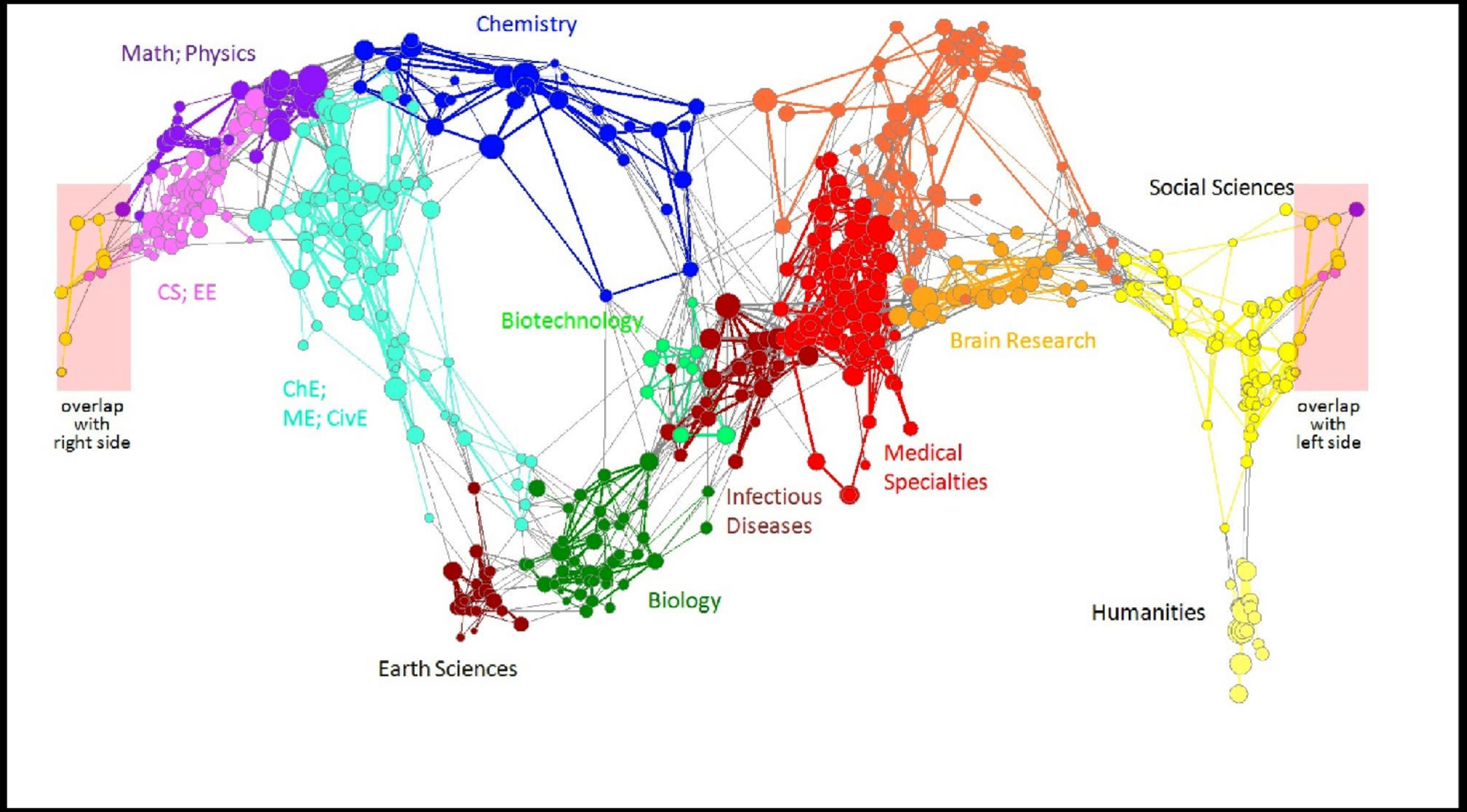

Retweet Network Graph plus IRA Troll Accounts. Courtesy of Kate Starbird/University of Washington

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

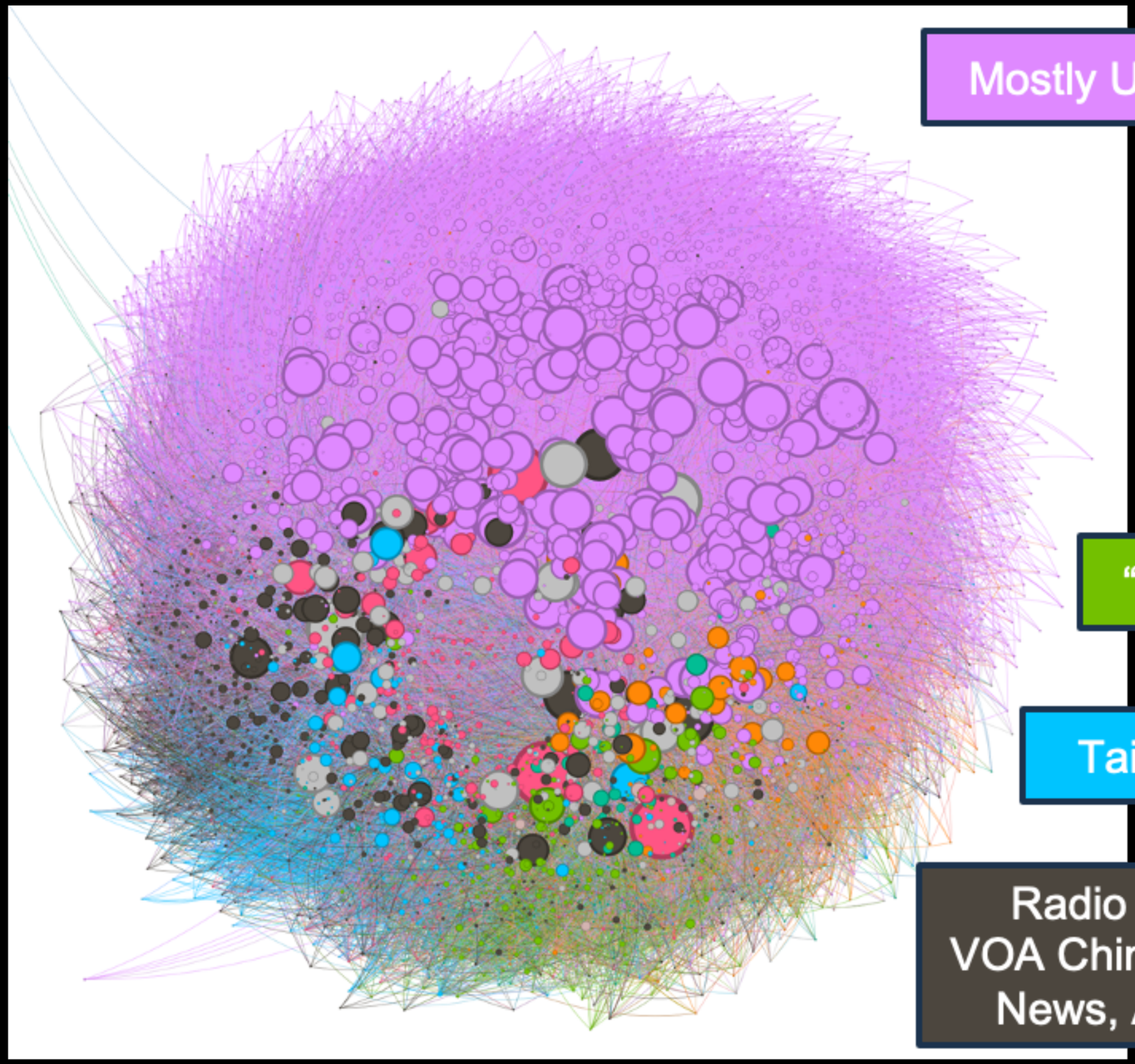
# Graphs in the Internet



# Graphs in Citations Networks




[Börner et al., 2012]



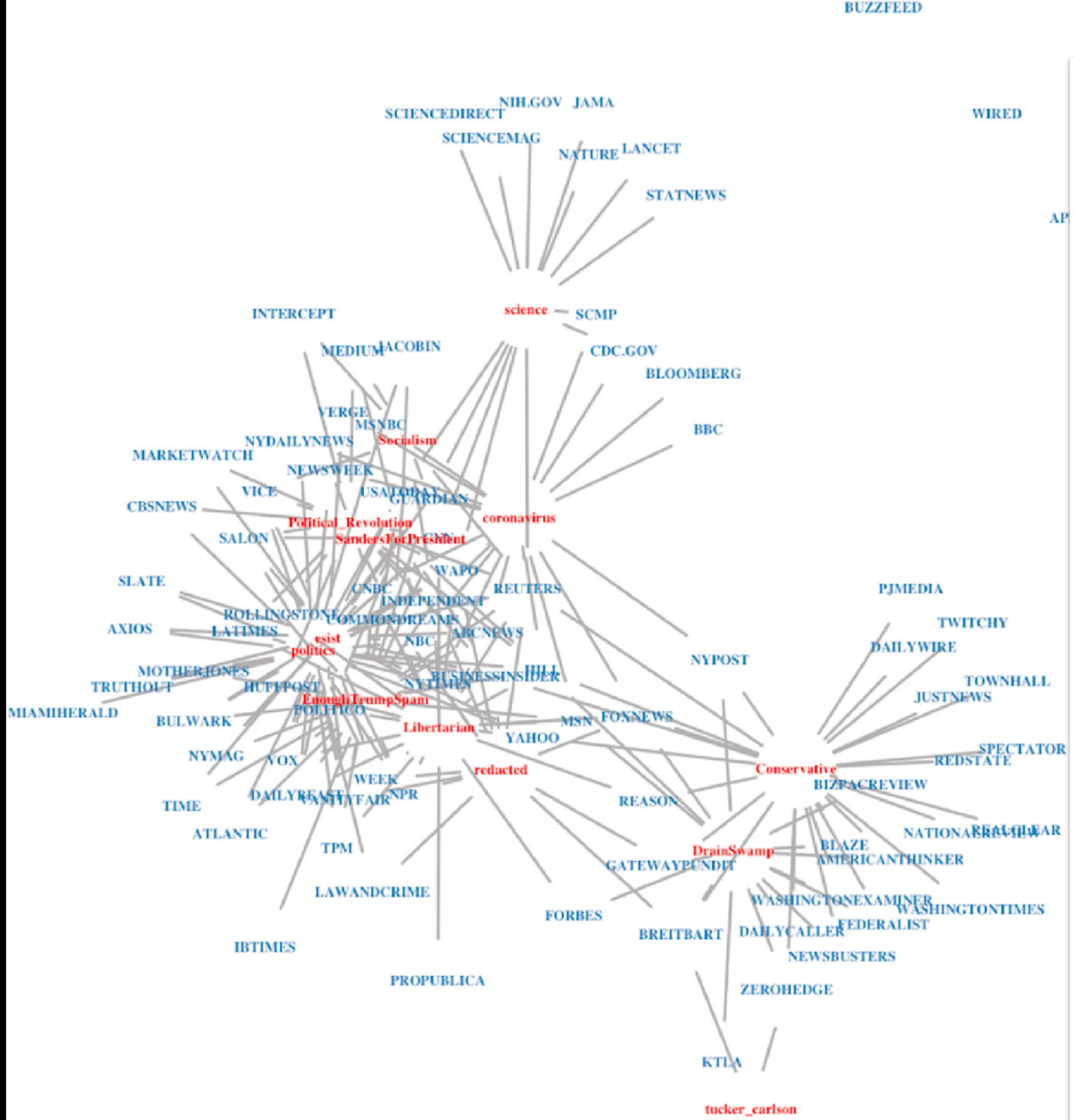
Mostly US news

Network of information sources in Hong Kong during a major protest

“Alternative” News

Taiwanese News

Radio Free Asia, VOA China, The Stand News, Apple Daily




International Journal of Information Management 61 (2021) 102397

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Research Article

# The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: An affiliation network approach

Wallace Chipidza

*Center for Information Systems and Technology, Claremont Graduate University, Claremont, CA 91711, USA*

ARTICLE INFO

*Keywords:*
Social media
Affiliation networks
Political polarization
Exponential random graph modeling
Toxicity
News sharing

ABSTRACT

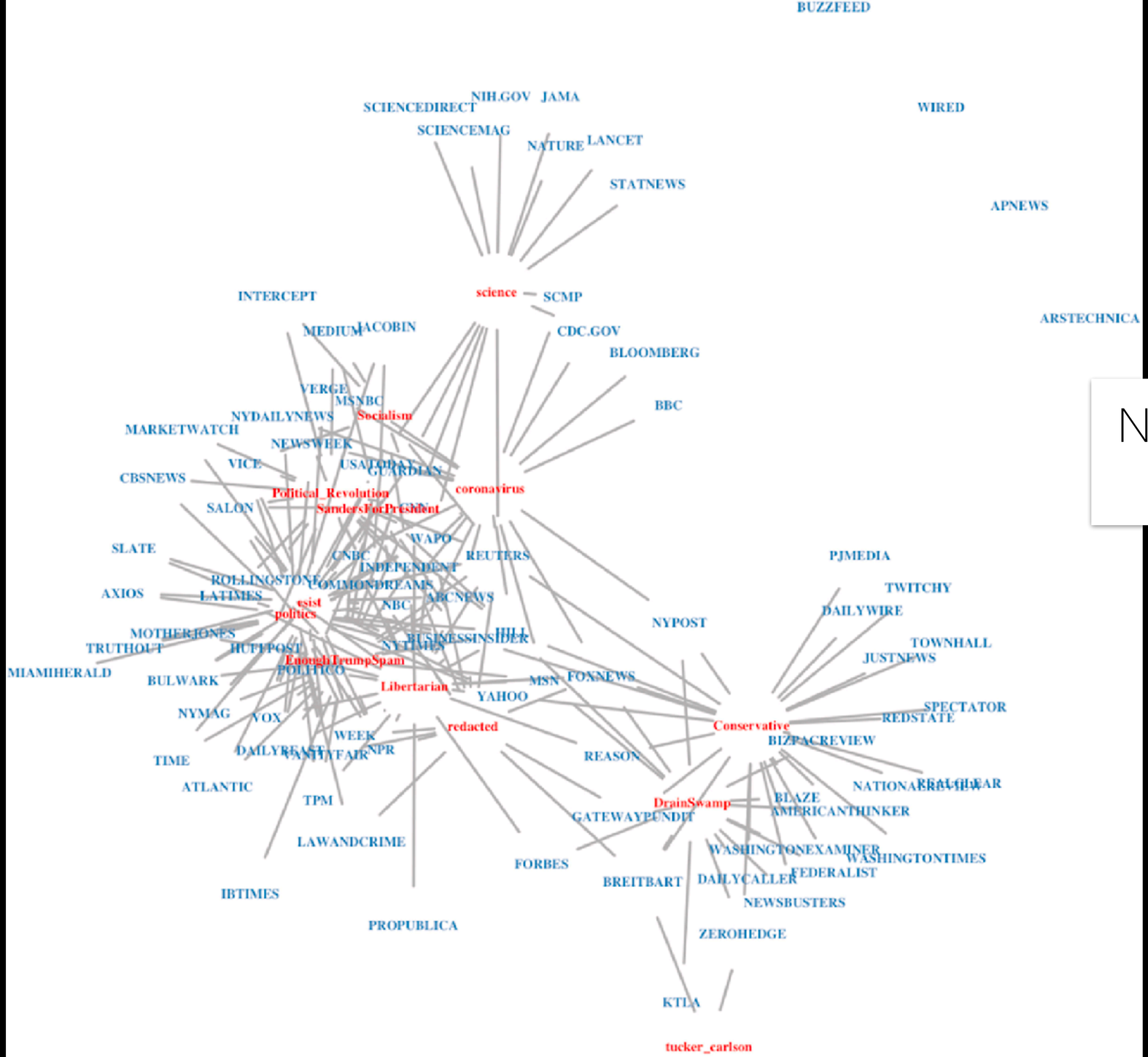
Political polarization remains perhaps the "greatest barrier" to effective COVID-19 pandemic mitigation measures in the United States. Social media has been implicated in fueling this polarization. In this paper, we uncover the network of COVID-19 related news sources shared to 30 politically biased and 2 neutral subcommunities on Reddit. We find, using exponential random graph modeling, that news sources associated with highly toxic – "rude, disrespectful" – content are more likely to be shared across political subreddits. We also find homophily according to toxicity levels in the network of online news sources. Our findings suggest that news sources associated with high toxicity are rewarded with prominent positions in the resultant network. The toxicity in COVID-19 discussions may fuel political polarization by denigrating ideological opponents and politicizing responses to the COVID-19 pandemic, all to the detriment of mitigation measures. Public health practitioners should monitor toxicity in public online discussions to familiarize themselves with emerging political arguments that threaten adherence to public health crises management. We also recommend, based on our findings, that social media platforms algorithmically promote neutral and scientific news sources to reduce toxic discussion in subcommunities and encourage compliance with public health recommendations in the fight against COVID-19.

## 1. Introduction

The COVID-19 pandemic has wreaked much havoc on human health and well-being through loss of lives, enforced isolation, and devastated economies (Baker, Bloom, Davis, & Terry, 2020). Humans have no natural immunity to SARS-CoV-2, the coronavirus that causes COVID-19, and as yet, vaccines are in limited supply both in the United States and elsewhere (McClung et al., 2020). There are also no known effective pharmaceutical therapies for the disease (Grossman, Kim, Rexer, & Thirumurthy, 2020), meaning that preventative measures are paramount in curtailing loss of human life and other negative effects accompanying the pandemic. Although populations across the world have complied with public health recommendations to slow the disease's spread, there has also been significant resistance, especially in the US (Barrios & Hochberg, 2020). Recently, evidence is accumulating that intention to receive COVID-19 vaccines is politically polarized, with Democrats much more willing to receive it than Republicans (Fridman, Gershon, & Gneezy, 2021). This resistance has been fueled by political polarization and is consistent with past responses to the H1N1 pandemic (Lee & Basnyat, 2013).

The traditional media has been implicated in fueling political polarization, and social media intensifies it further (Halberstam & Knight, 2014). Increasingly, Americans report getting more of their news from social media rather than traditional media (Pew Research Center, 2019). Social media has surpassed search engines as the main disseminator of news on the internet (Boxer, 2015). The characteristics of news content influence its sharing; content that invokes strong emotions is more likely to be shared than content that invokes weaker emotions (Berger & Milkman, 2012). It is feasible that news sources which create emotionally provoking content are likely to be shared more often on social media platforms and thus occupy influential positions of the resulting network of news sources.

Understanding news sharing on social media is important, because users on the same social media platform may choose different information sources and are thus exposed to and ultimately consume different content (Halberstam & Knight, 2014). COVID-19 is a serious threat to societal well-being and warrants shared understanding of the nature of the threat and concerted action to prevent its spread. Ideally, qualified public health professionals should be conveying COVID-19 related information to the public (Regidor et al., 2007); but currently a significant proportion of society obtains health news from social media (Allington, Duffy, Wessely, Dhavan, & Rubin, 2020). This study




Node/edge diagram of subreddits and news sources around COVID

What's the most important news sources in this network?

How might we define “importance” in a graph?

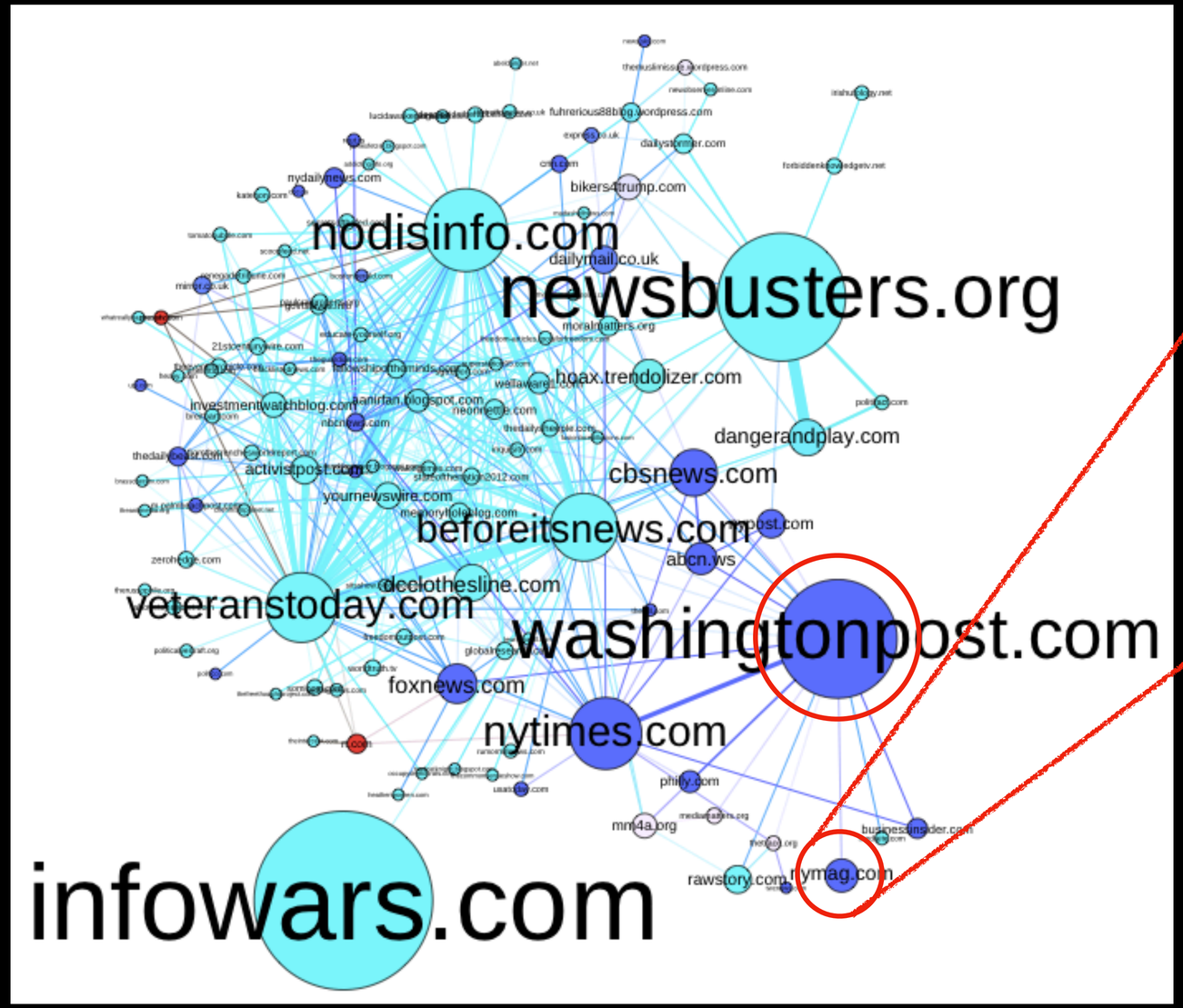
Via “centrality” metrics

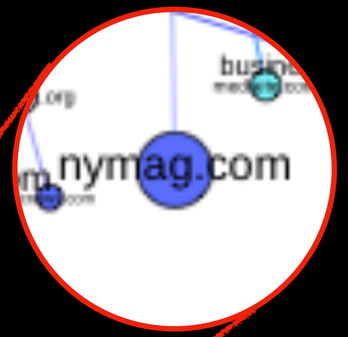
# This Lecture's Learning Objectives

Describe at least three network centrality metrics

Use the Girvan-Newman method to identify communities in graphs

# HOW DO TWO NODES DIFFER?

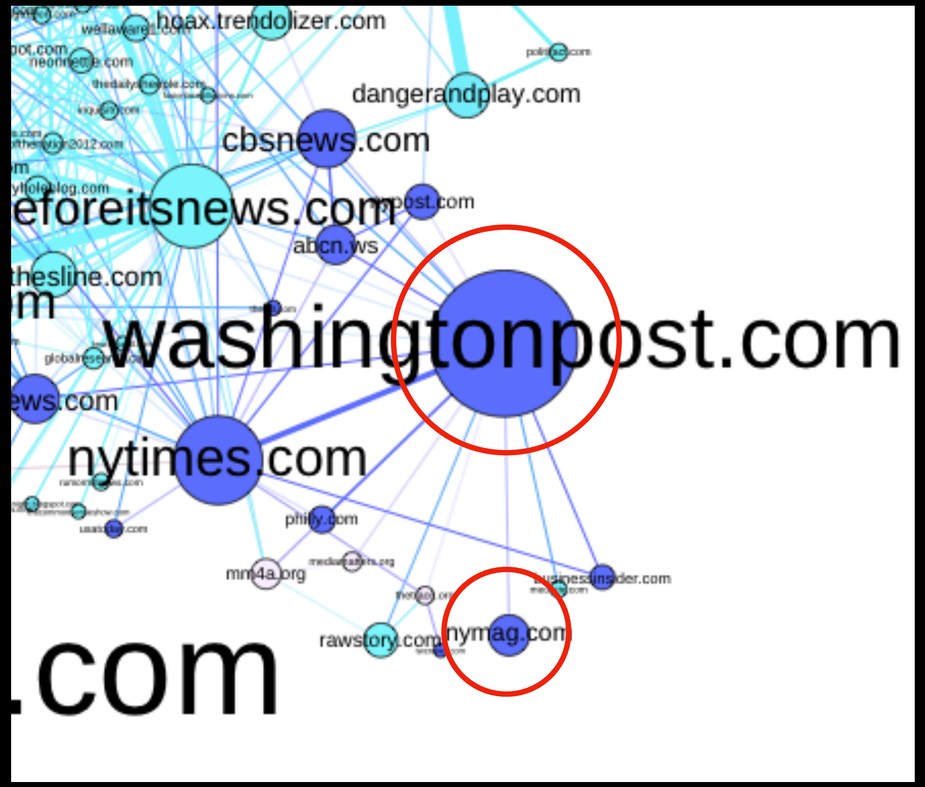



WashingtonPost.com has many more neighbors than nymag.com

# What is “Degree”?

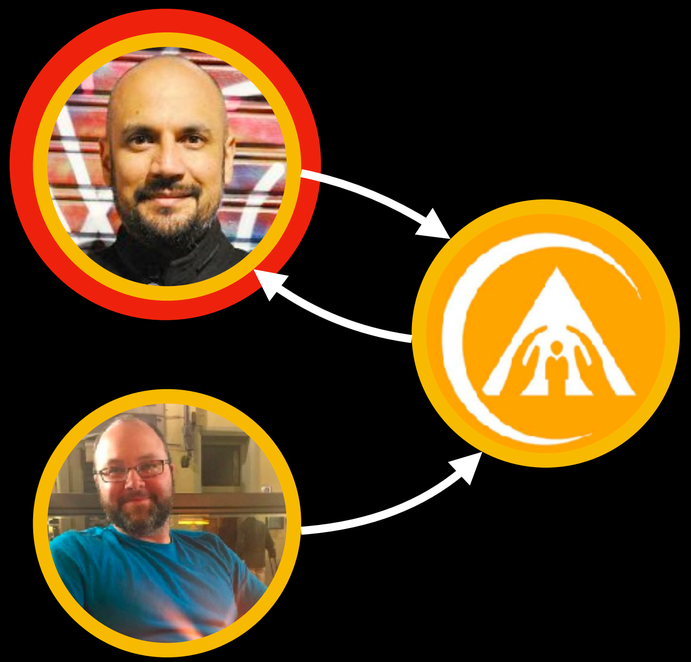
- “Degree” of a node
  - “Number of neighbors” a node has
- Degree(WaPo.st) >> Degree(NyMag.com)



# What is “Degree”?

- Undirected graphs
  - Degree = number neighbors
  - Also # of *incident* edges
- Directed graphs
  - in-degree - edges coming “into”
  - out-degree - edges going “out of”

# What is “Degree”?



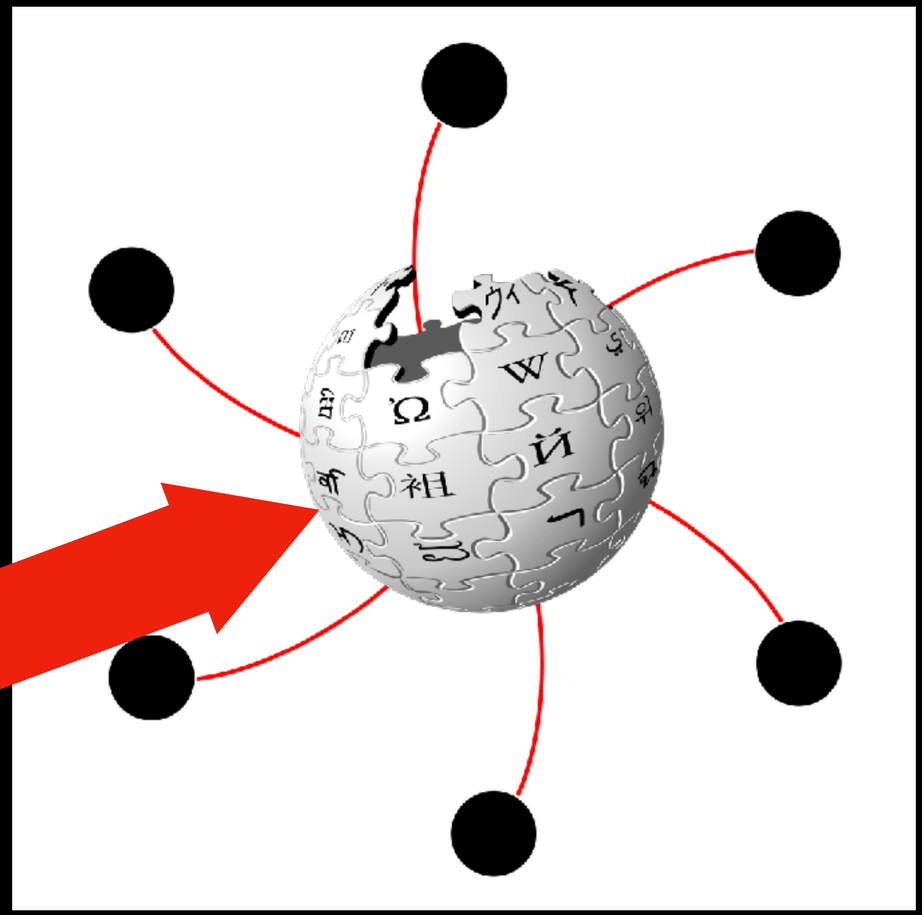
- In-Degree = 1
- Out-Degree = 1
- Degree = In-Degree + Out-Degree = 2

Does higher degree mean more influence?

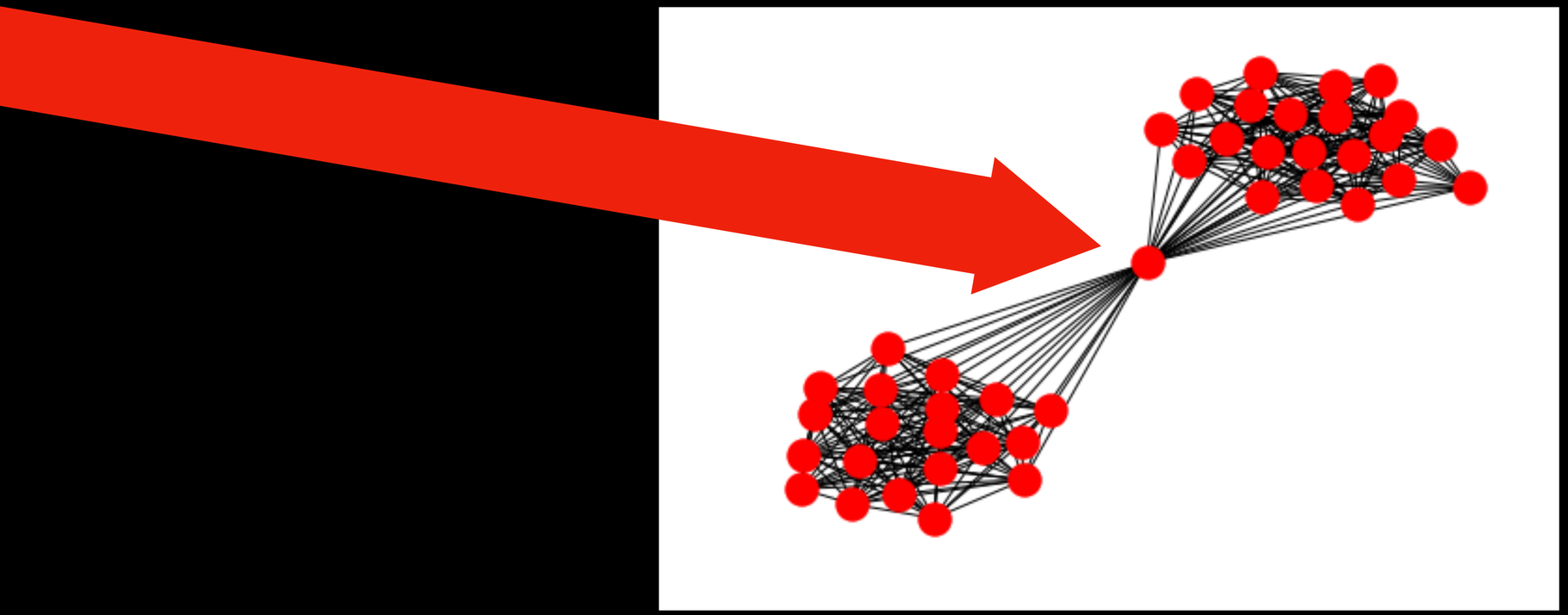
What about in-degree vs out-degree?

In Twitter, does more followers means more influence than more friends?

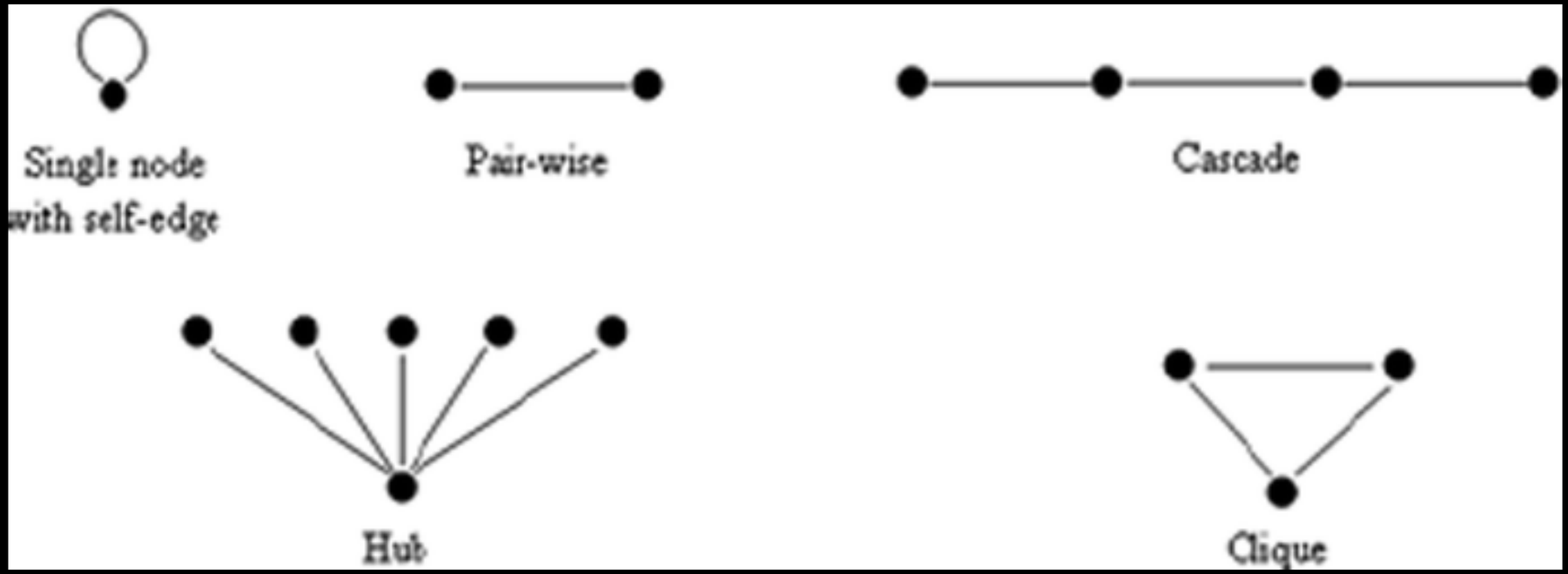
# Types of Nodes - Hubs and Bridges

Hubs



Bridges




Motifs

# What does it mean for a node to be “important”?

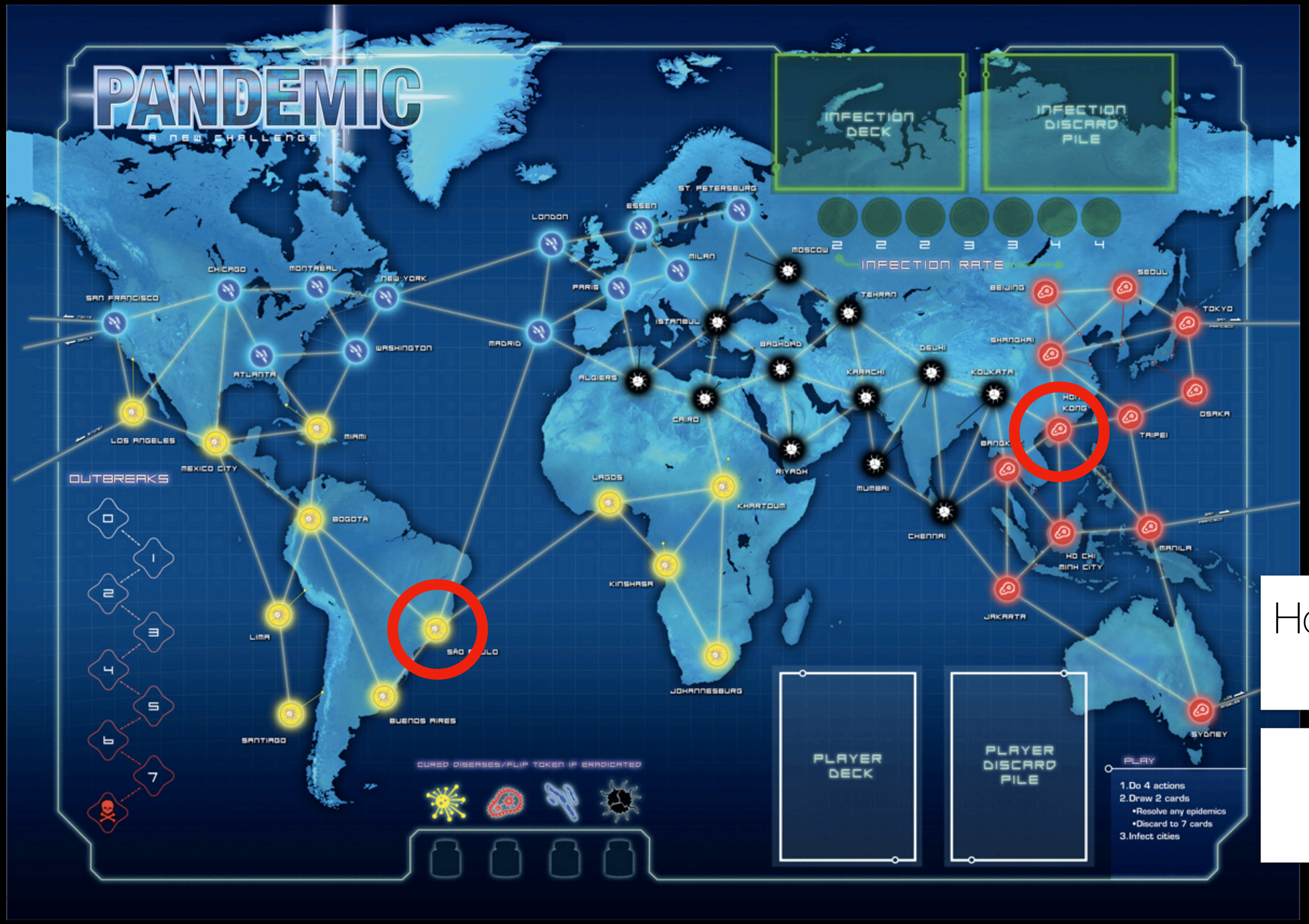
Highest degree?

Having a high PageRank?

Being an authority?

Being a hub *and* an authority?

Something else?




Hong Kong has highest degree

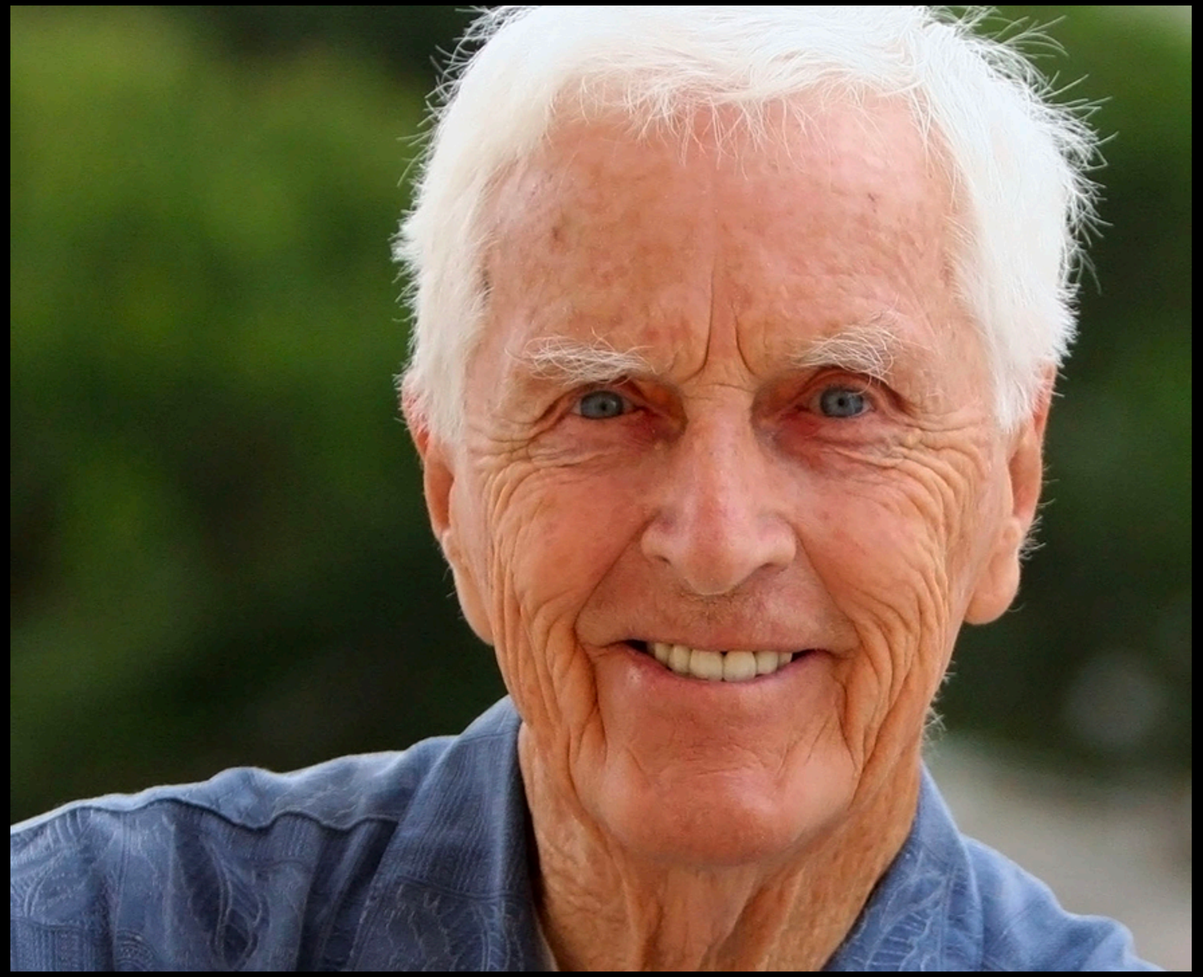
Sao Paolo bridges South America

# Defining “Centrality”

- Influence?
- Prestige?
- Power?
- Control?

"There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement.”

*– Linton Freeman, 1979*



# Five common types of centrality

Eigenvector Centrality

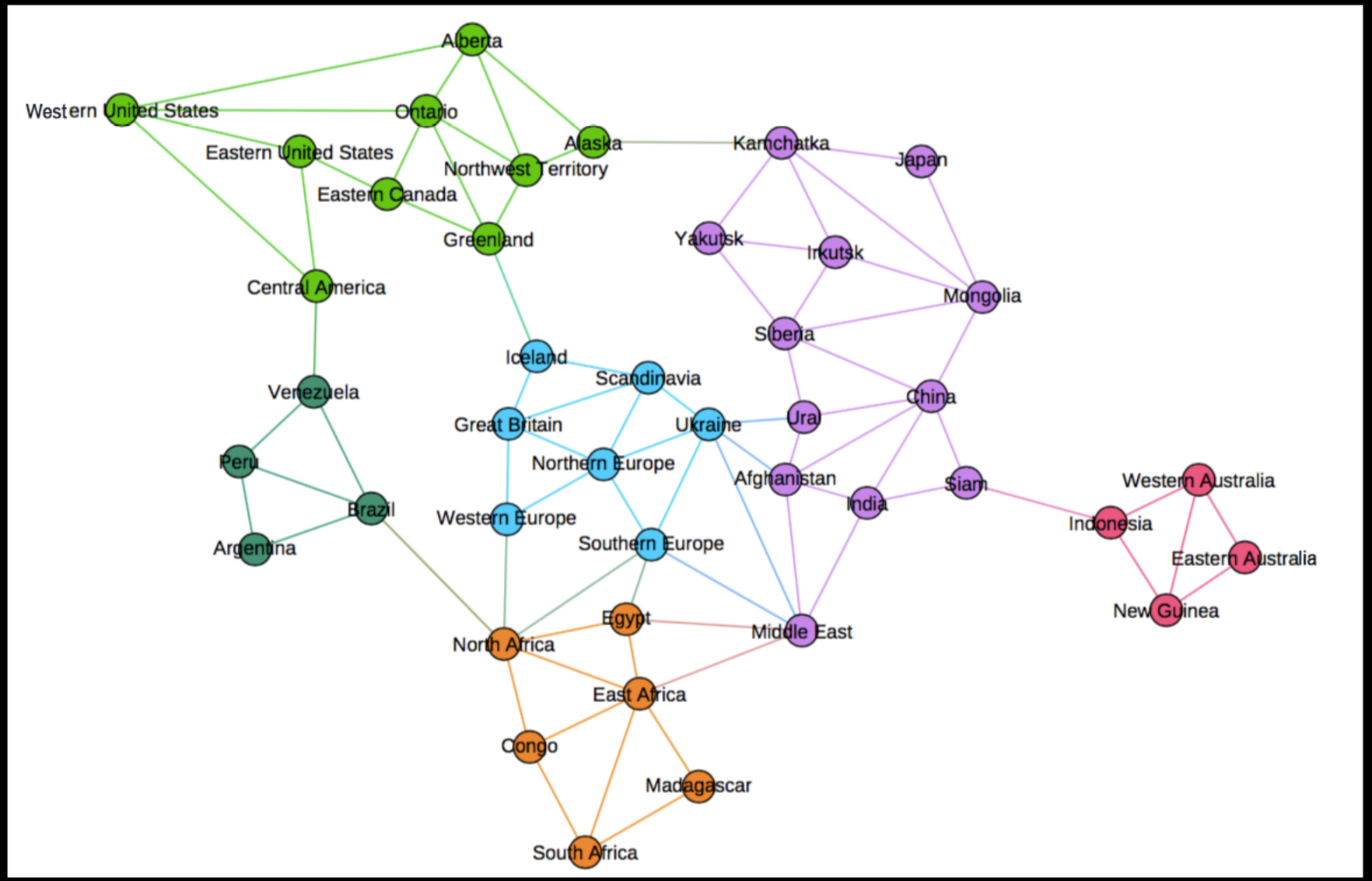
Degree Centrality

Closeness Centrality

Betweenness Centrality

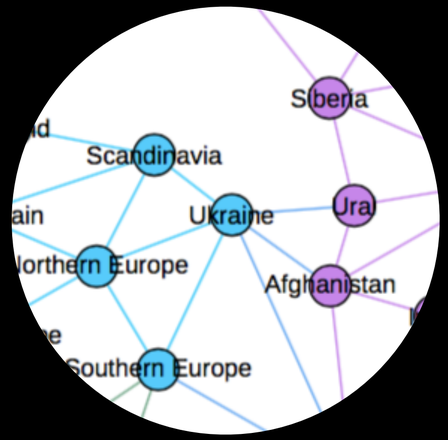
PageRank

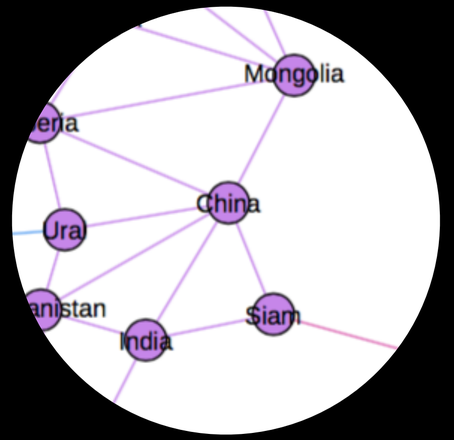

Greenland
Alaska
North West Territory
Alberta
Ontario
Quebec
Western United States
Eastern United States
Central America
Iceland
Scandinavia
Great Britain
Northern Europe
Western Europe
Ukraine
Ukraine
Siberia
Yakutsk
Kamchatka
Ural
Irkutsk
Afgharistan
Mongolia
Middle East
China
India
Siam
Japan
Egypt
North Africa
East Africa
Congo
South Africa
Madagascar
Venezuela
Brazil
Peru
Argentina
Indonesia
New Guinea
Western Australia
Eastern Australia
Risk

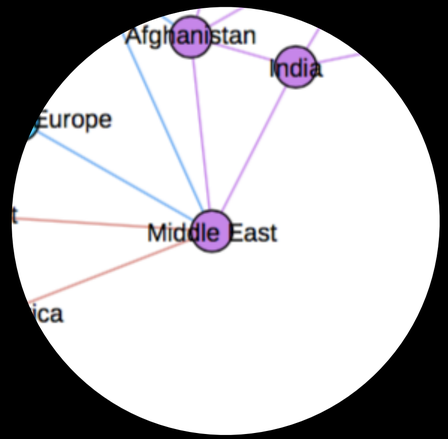

Alberta
Western United States
Ontario
Eastern United States
Alaska
Northwest Territory
Eastern Canada
Greenland
Central America
Kamchatka
Japan
Yakutsk
Irkutsk
Mongolia
Siberia
Iceland
Scandinavia
China
Venezuela
Great Britain
Ukraine
Ural
Peru
Northern Europe
Afghanistan
Siam
Western Australia
Brazil
Western Europe
India
Indonesia
Argentina
Southern Europe
Eastern Australia
Egypt
New Guinea
Middle East
North Africa
East Africa
Congo
Madagascar
South Africa

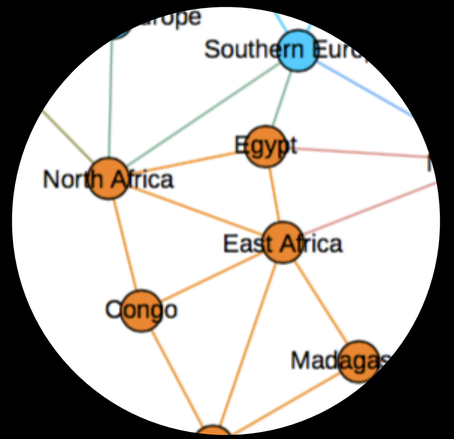
# Degree Centrality

- Nodes with highest degree have highest *degree* centrality
- Can immediately reach the most neigbors



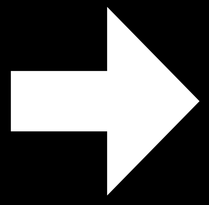



# Degree Centrality

**Binary Adj Matrix A =**

$$\begin{bmatrix} 1 & 0 & 1 & 0 & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & & \vdots & & & \vdots & \\ a_{n1} & \dots & a_{nj} & \dots & \dots & \dots & 1 & \end{bmatrix}$$

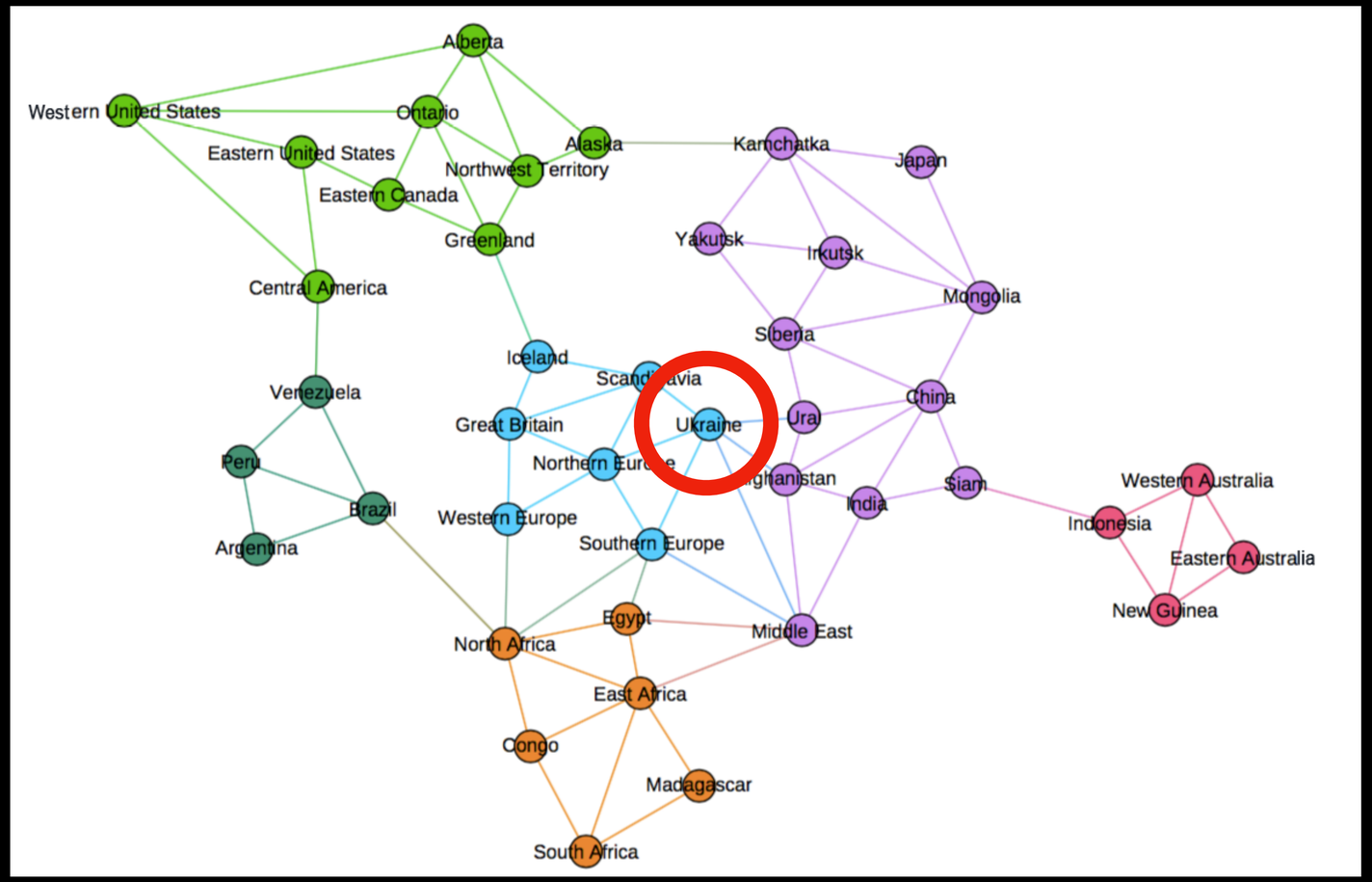
$$\mathbf{c}_i^d = \sum_{i=1}^{n} a_i$$



$$\bar{c}_i^d = \frac{c_i^d}{n-1}$$

# Closeness Centrality

- Nodes with the most shortest paths have highest *closeness* centrality
  - Inverse of “farness”
- Can reach the most (even distant) nodes quickly



# Closeness Centrality

Farness =>

$$\text{farness}(v_i) = \sum_{j \neq i} \text{shortest} \ldots \text{ance}(v_i, v_j)$$

Only defined for connected graphs

Closeness =>

$$\frac{1}{\text{farness}(v_i)}$$

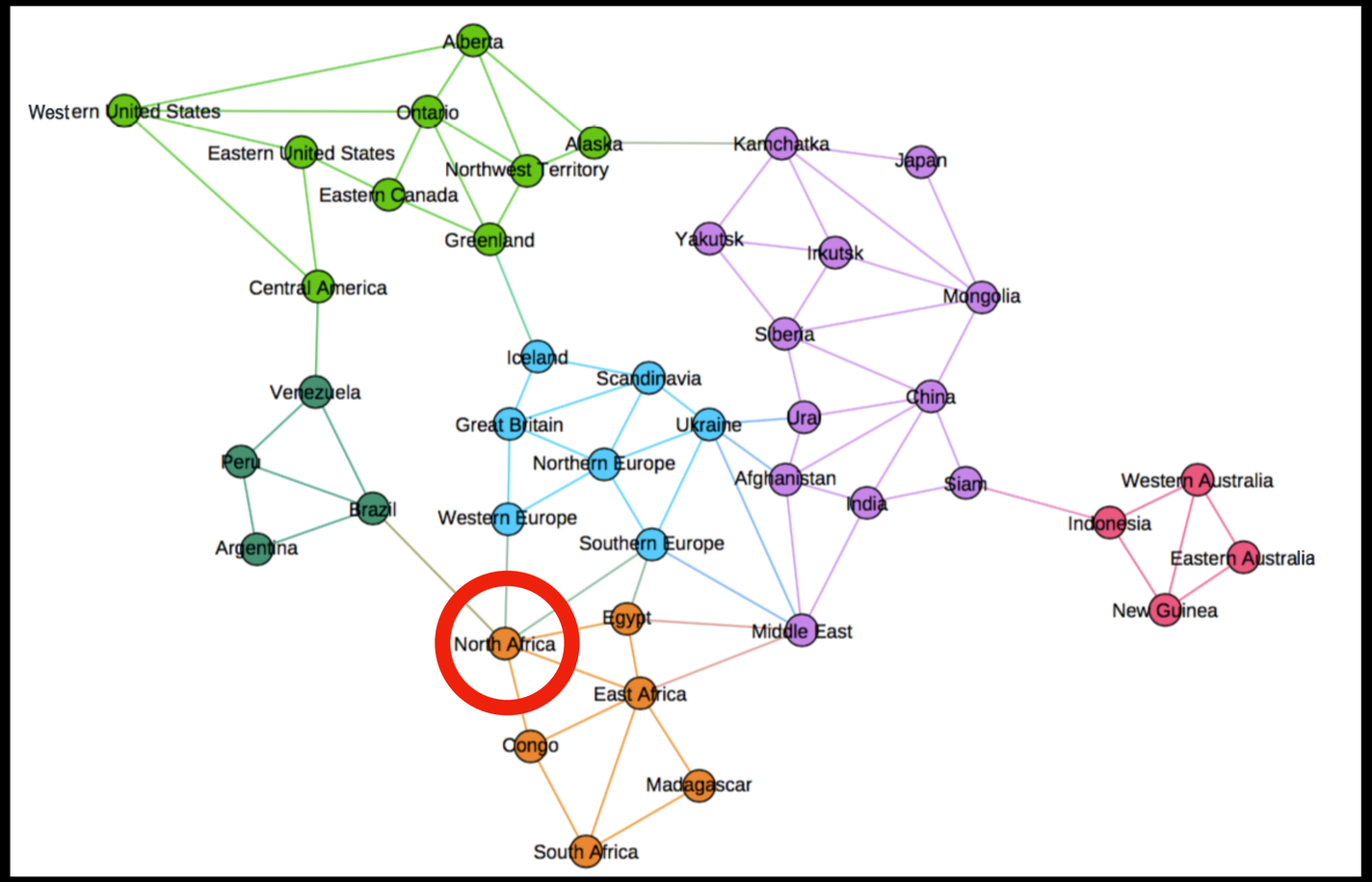
Maximum closeness

Closeness =>

$$\frac{n-1}{\text{farness}(v_i)}$$

# Betweenness Centrality

- Nodes that appear most often in shortest paths have highest *betweenness* centrality
- Many paths must flow throw nodes with high betweenness
  - i.e., bridges



# Betweenness Centrality

How many shortest paths between s and t go through v?

How many shortest paths exist between s and t?

$$Betwenness(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

**Undirected**

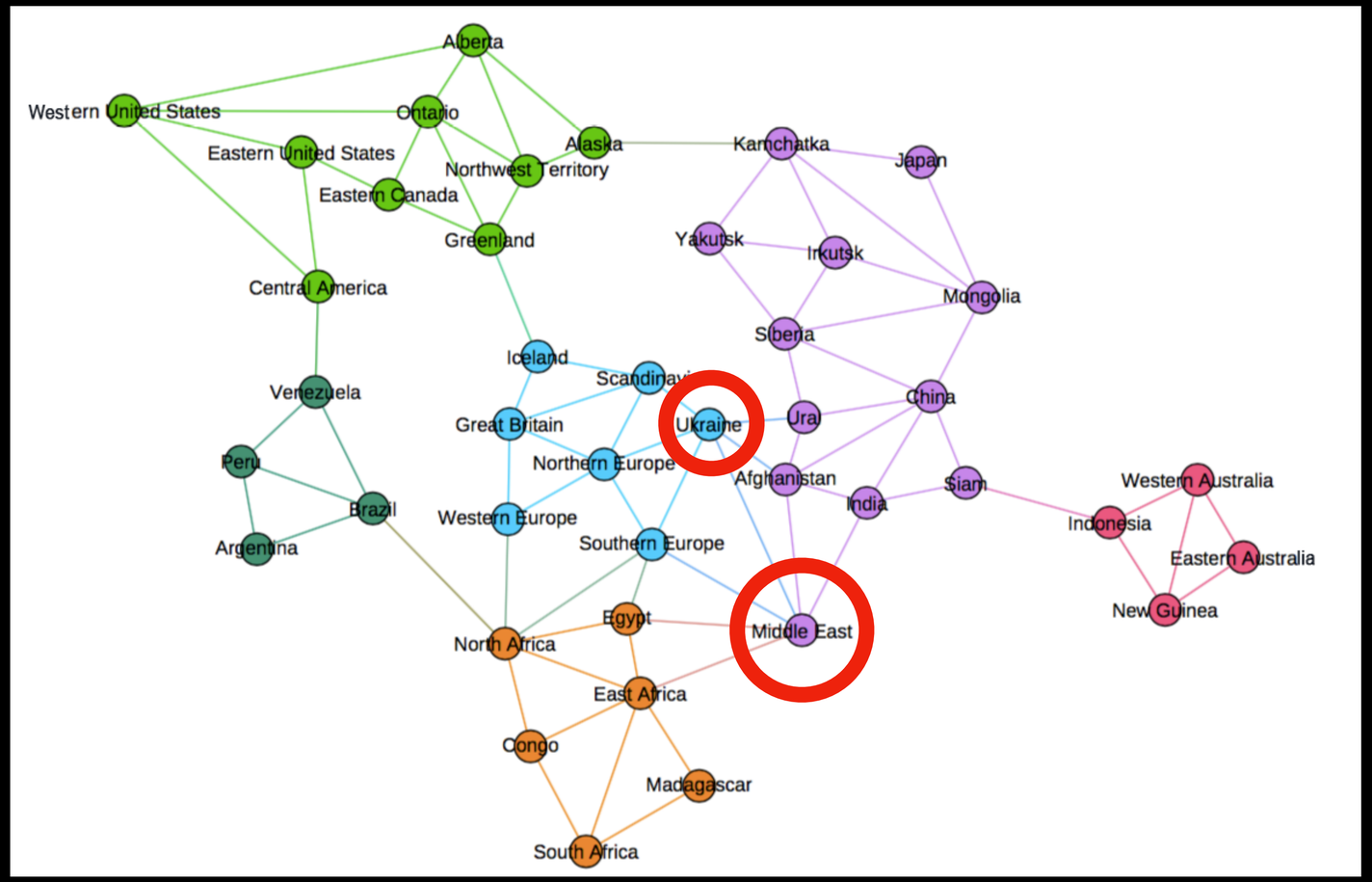
$$\overline{Betweenness}(v) = \frac{2 \times Betweenness(v)}{(n-1)(n-2)}$$

**Directed**

$$\overline{Betweenness}(v) = \frac{Betweenness(v)}{(n-1)(n-2)}$$

# Eigenvector Centrality

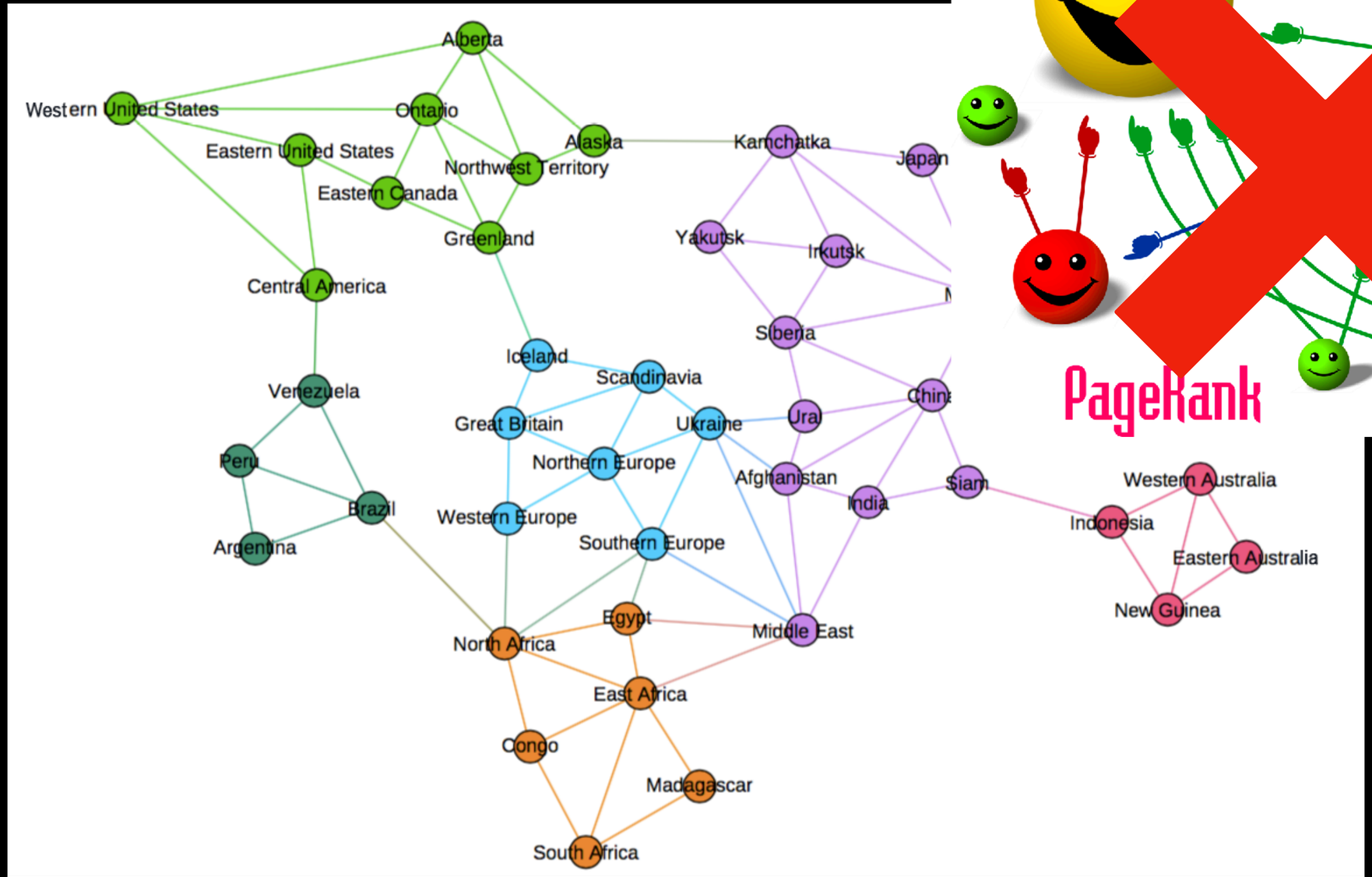
- Nodes that are close to other central nodes have high *eigenvector* centrality
- A node's eigenvector centrality relies on its neighbors'

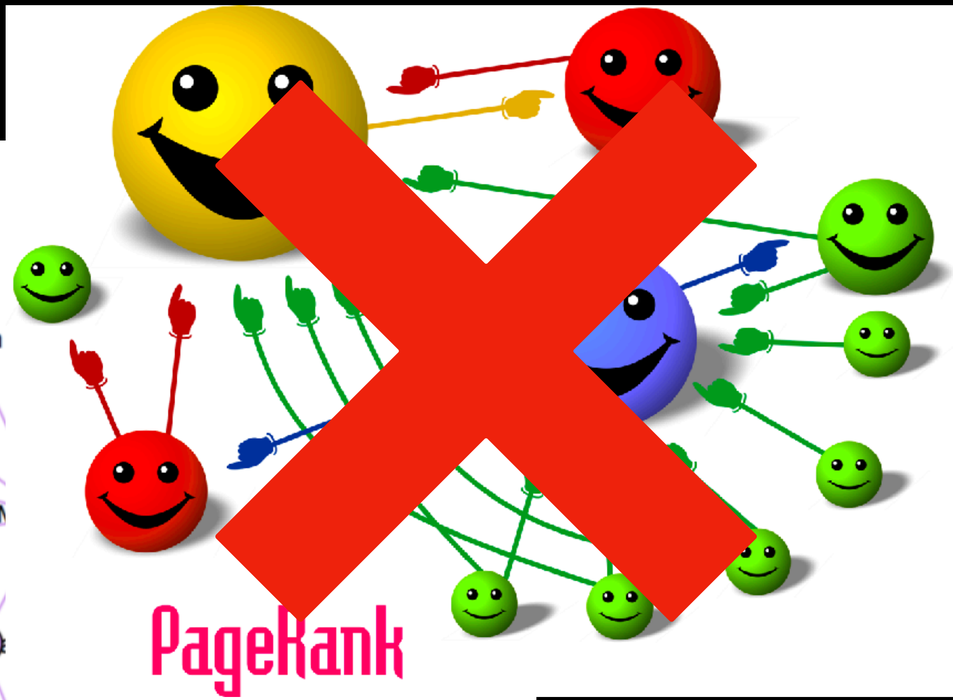


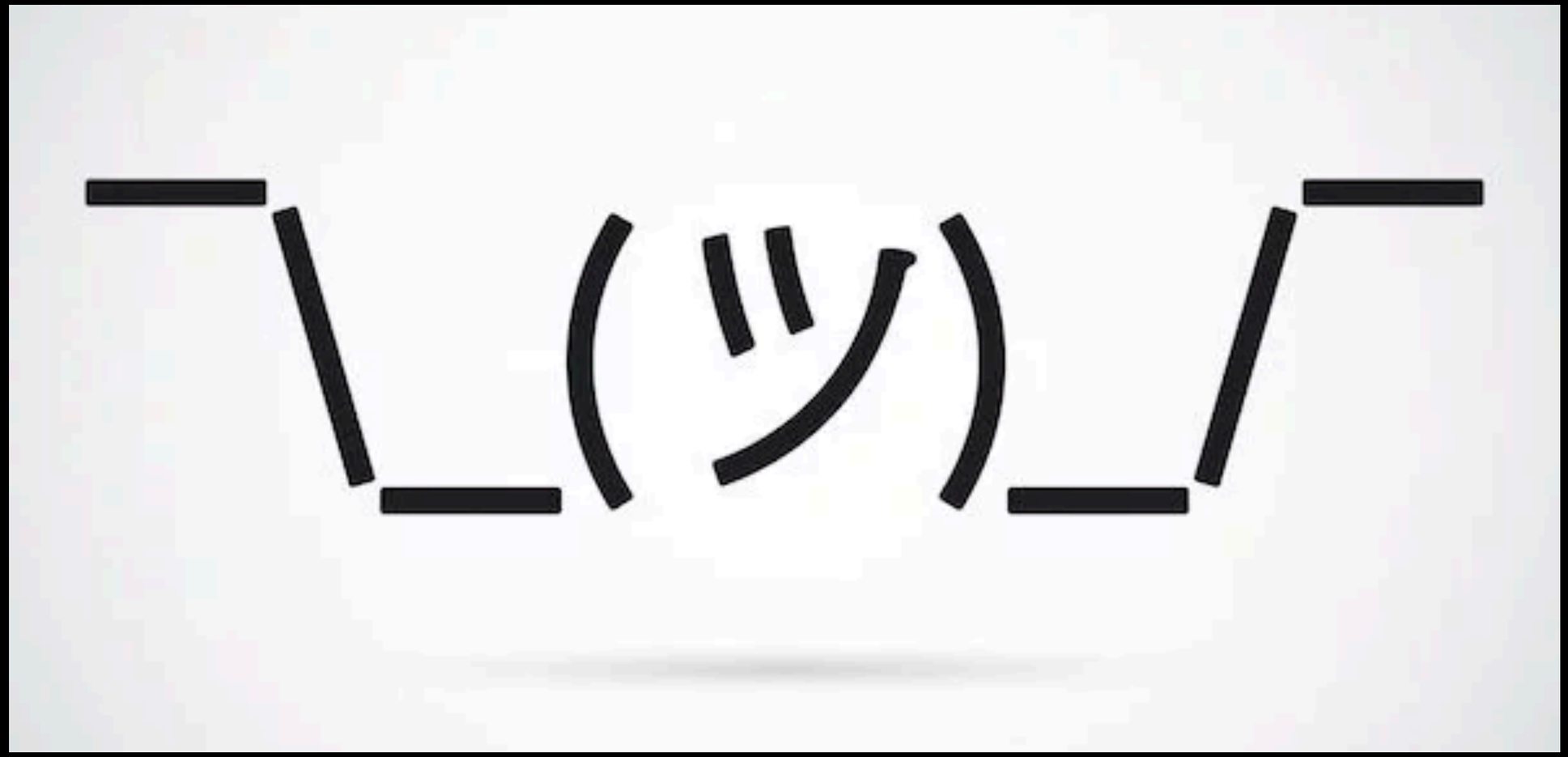
# Eigenvector Centrality

- PageRank but for undirected graphs
- Calculable through the Power Iteration

$$A\pi = \lambda\pi$$

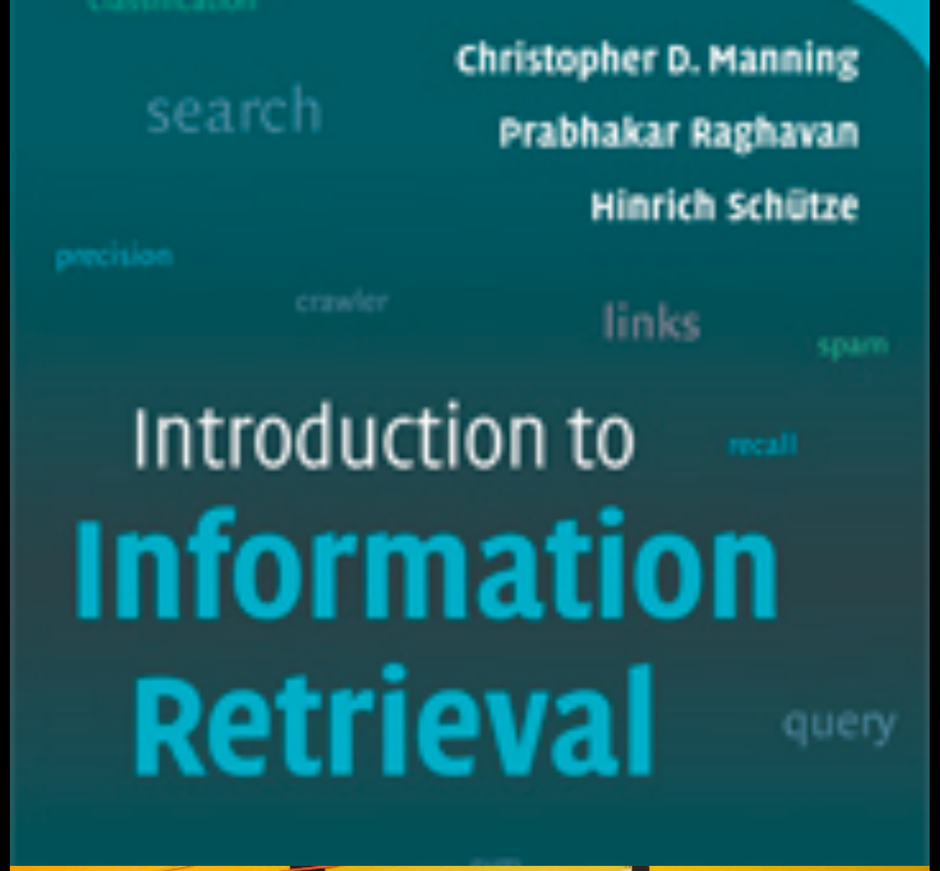

Alberta
Western United States
Ontario
Eastern United States
Alaska
Northwest Territory
Eastern Canada
Greenland
Central America
Kamchatka
Japan
Yakutsk
Irkutsk
Siberia
Iceland
Scandinavia
Venezuela
Great Britain
Ukraine
Ural
Peru
Northern Europe
Afghanistan
Siam
Brazil
Western Europe
India
Indonesia
Western Australia
Argentina
Southern Europe
Eastern Australia
Egypt
Middle East
New Guinea
North Africa
East Africa
Congo
Madagascar
South Africa


PageRank




Which centrality measure is the best?

How might we use these centrality metrics?





# Query: Amanda Gorman

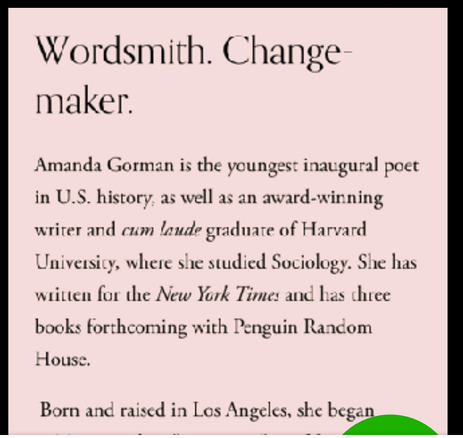
Largest *r*

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 b status, and earned her a professiona management contract.

Next *r*



Wordsmith. Change-maker.

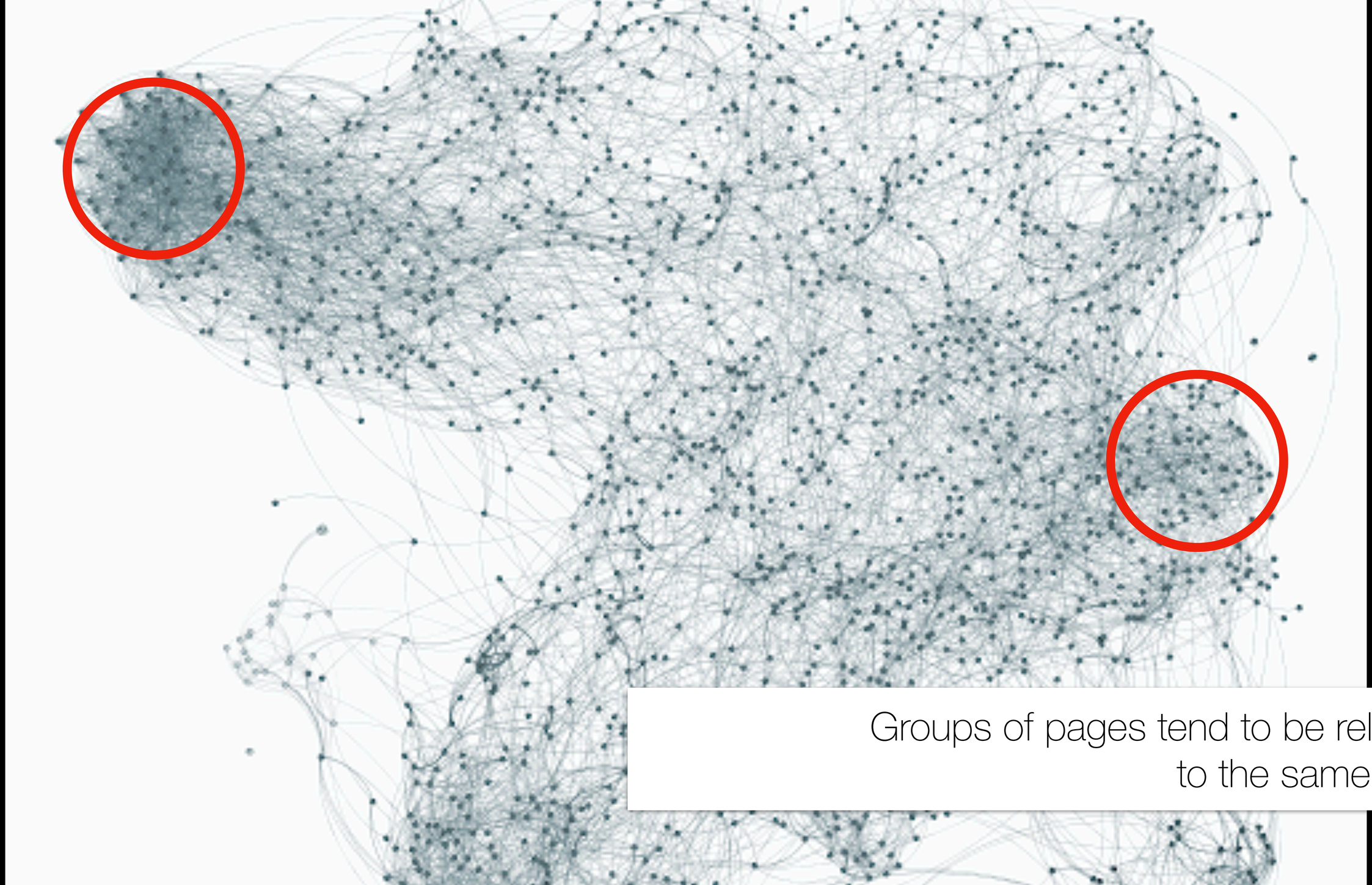
Amanda Gorman is the youngest inaugural poet in U.S. history, as well as an award-winning writer and *cum laude* graduate of Harvard University, where she studied Sociology. She has written for the *New York Times* and has three books forthcoming with Penguin Random House.

Born and raised in Los Angeles, she began

Google's PageRank Algorithm

Small *r*

Motivating Question: How do we organize the Web for search?



Groups of pages tend to be relevant
to the same topic

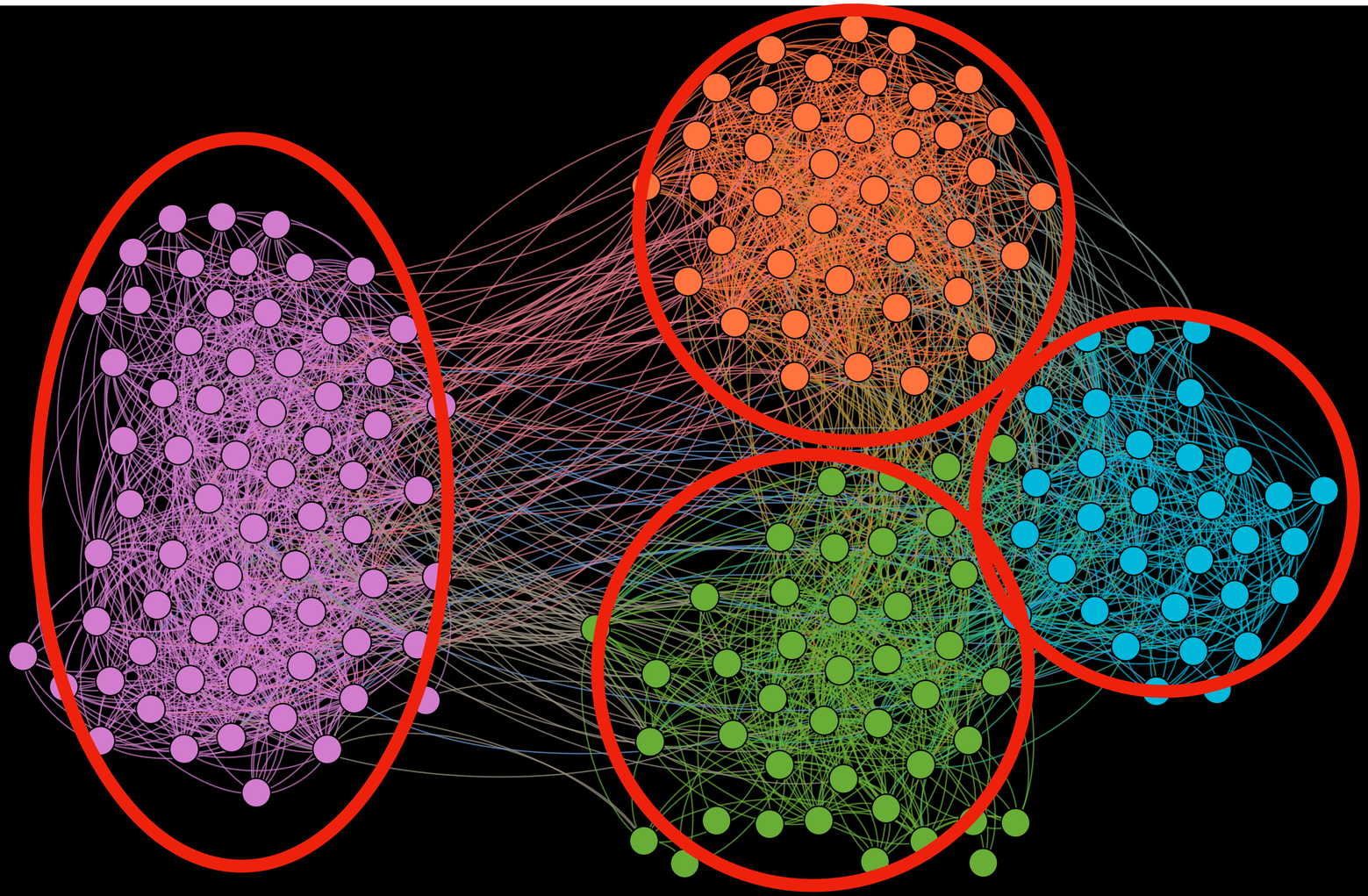
New Motivation: How do we find these *groups* of nodes?

# This Lecture's Learning Objectives

Describe at least three network centrality metrics

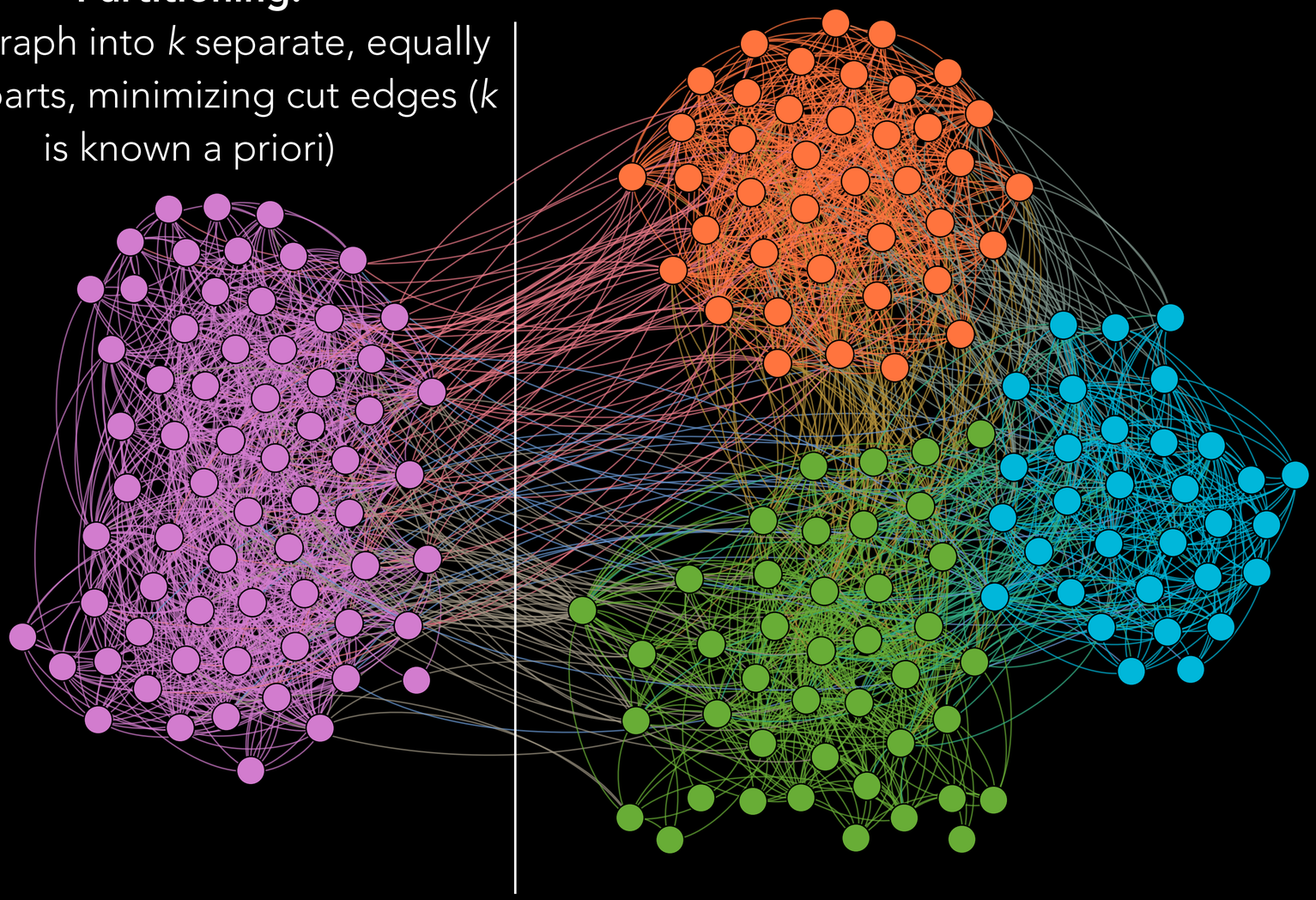
Use the Girvan-Newman method to identify communities in graphs

# Graph Clustering and Partitioning



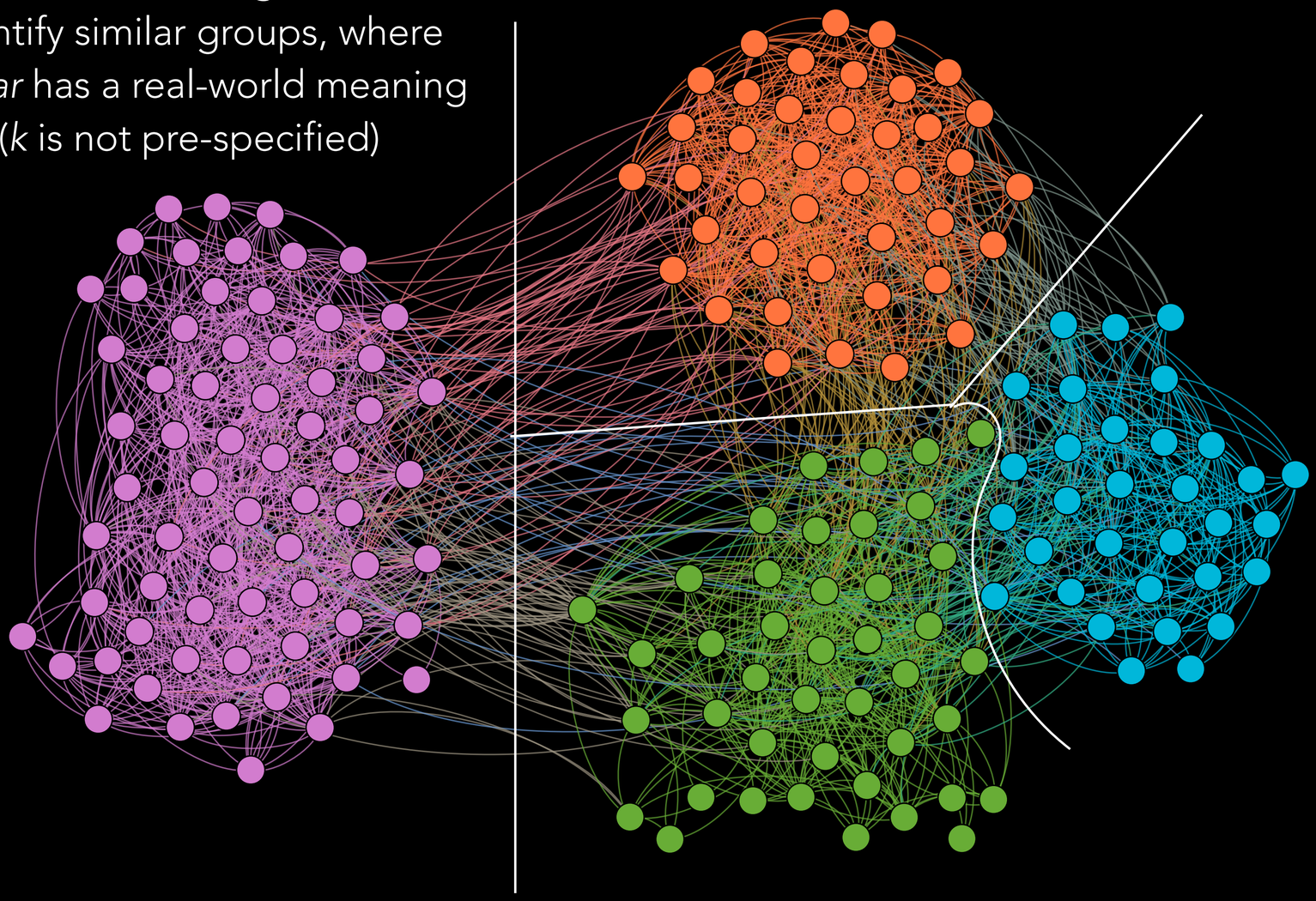
# PARTITIONING VS. CLUSTERING

**Partitioning:**
Split graph into *k* separate, equally sized parts, minimizing cut edges (*k* is known a priori)



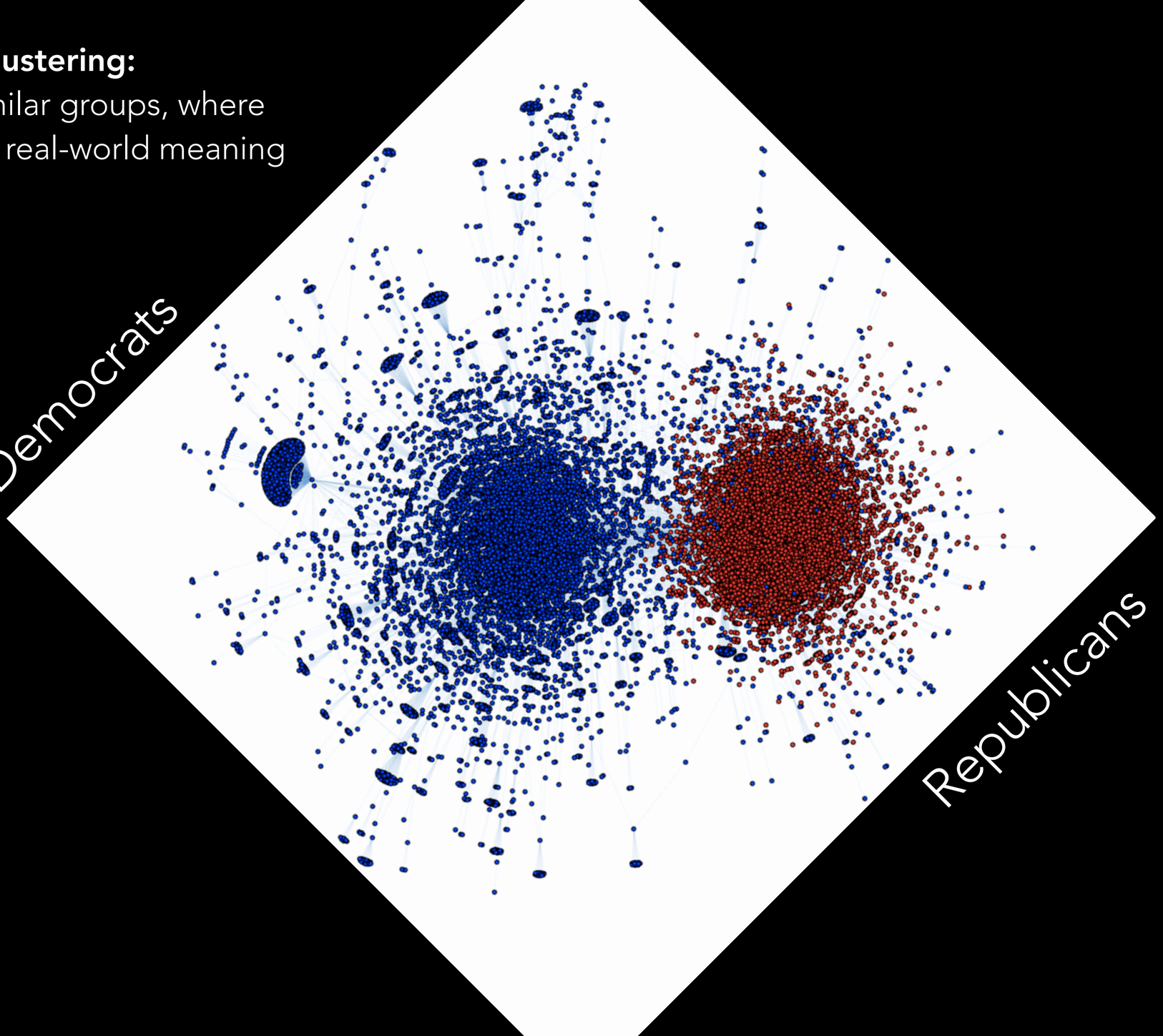
# PARTITIONING VS. CLUSTERING

**Clustering:**
Identify similar groups, where *similar* has a real-world meaning
(*k* is not pre-specified)



**Clustering:**

Identify similar groups, where *similar* has a real-world meaning



# Clustering Defined

- Cluster task:
  - **Maximize** the number of within-cluster edges
  - **Minimize** the number of between-cluster edges

How do we know a particular set of clusters is “good”?

# Evaluating a Clustering – Modularity

EDGE FROM I TO J

1 IFF I AND J ARE IN SAME GROUP

$$Q = \sum_{ij} \left[ \left( \frac{1}{2m} A_{ij} - \frac{k_i}{2m} \times \frac{k_j}{2m} \right) \delta(x_i, x_j) \right]$$

ADJACENT NODES WITH FEW EDGES SHOULD BE IN THE SAME GROUP

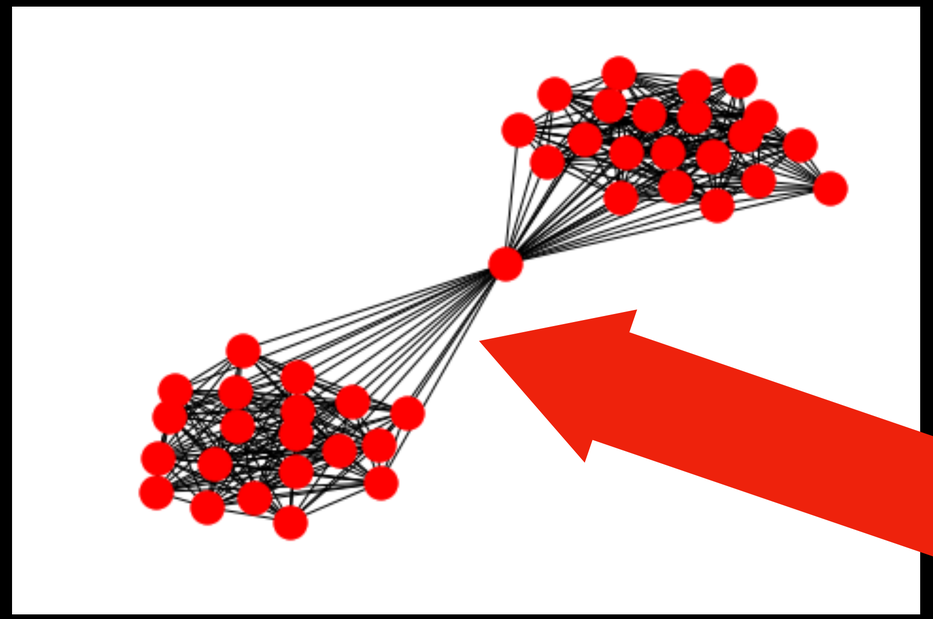
PROBABILITY OF A RANDOM EDGE FROM I TO J

$k_i$ = degree of node k

m = number of edges

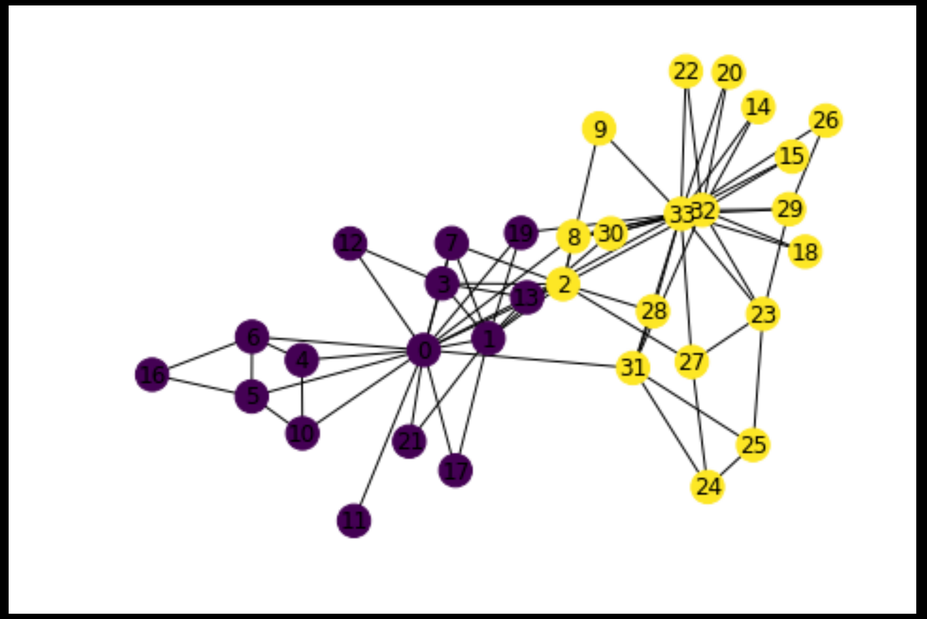
# Betweenness Centrality and Clustering

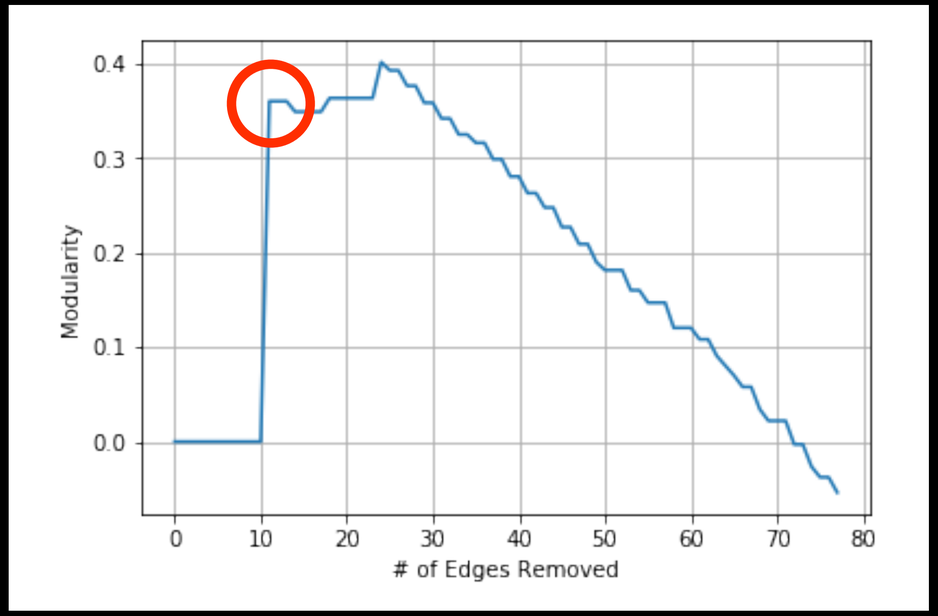
- Girvan-Newman algorithm for identifying clusters
- Iteratively remove ***edges*** with high betweenness centrality
- Stop removing edges when modularity stops increasing



# Girvan-Newman Clustering

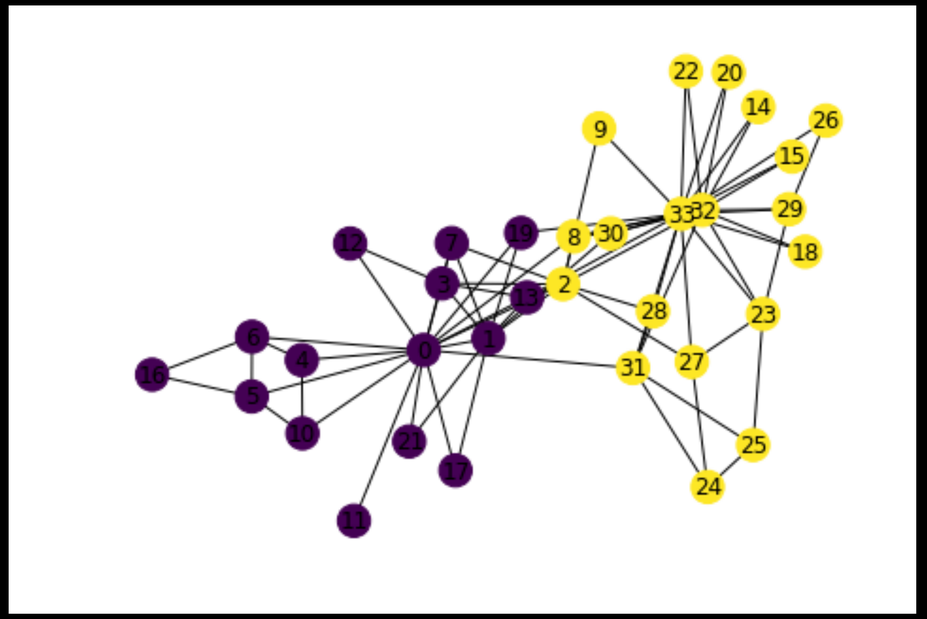
## Zachary's Karate Club Graph



# Girvan-Newman Clustering

## Zachary's Karate Club Graph




# Girvan-Newman Clustering

- Implemented in NetworkX

**girvan_newman**(*G, most_valuable_edge=None*) [source]

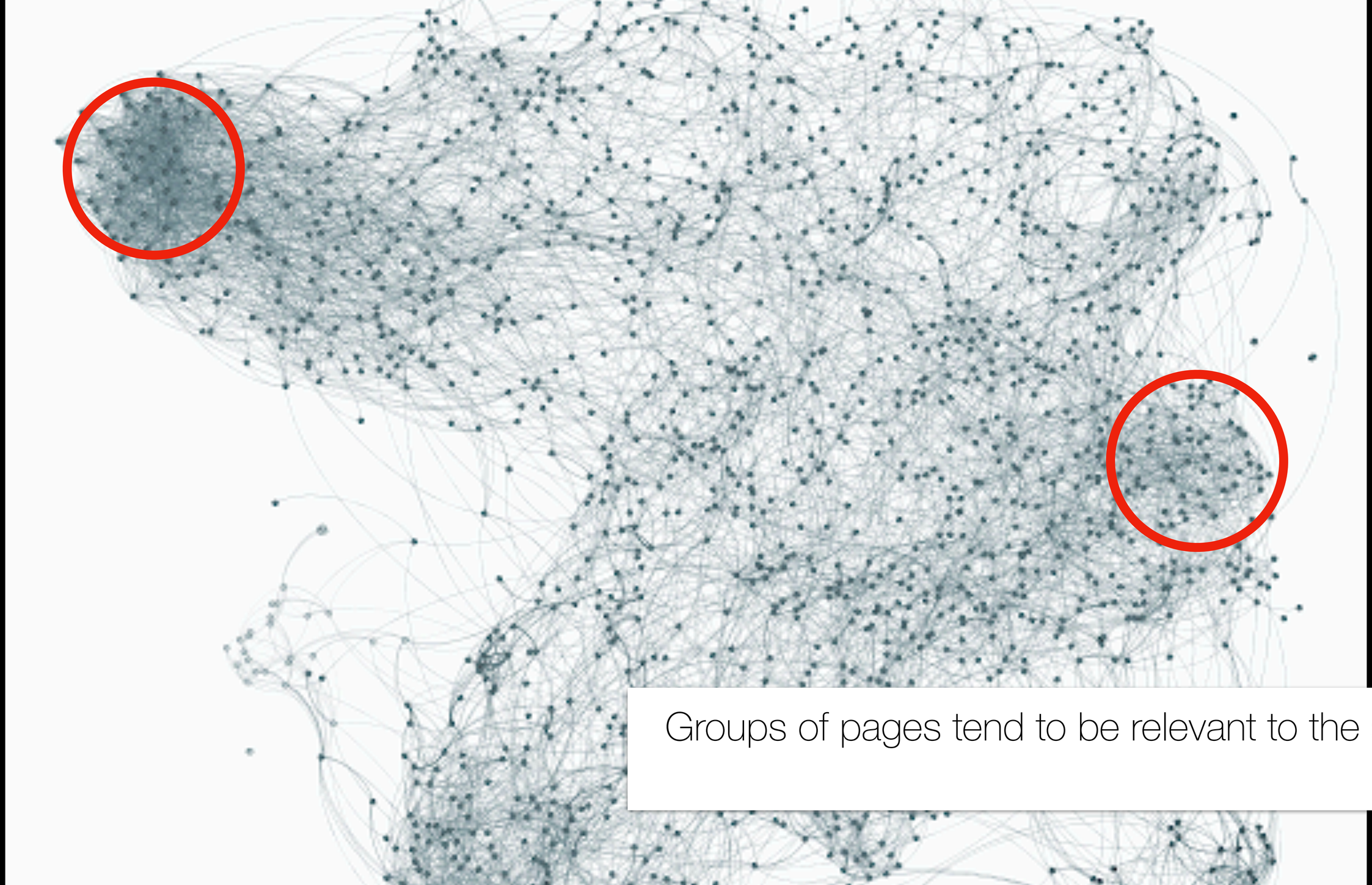
Finds communities in a graph using the Girvan–Newman method.

**Parameters:**
- **G** (*NetworkX graph*)
- **most_valuable_edge** (*function*) – Function that takes a graph as input and outputs an edge. The edge returned by this function will be recomputed and removed at each iteration of the algorithm.

  If not specified, the edge with the highest `networkx.edge_betweenness_centrality()` will be used.

**Returns:** Iterator over tuples of sets of nodes in `G`. Each set of node is a community, each tuple is a sequence of communities at a particular level of the algorithm.

**Return type:** iterator



Groups of pages tend to be relevant to the same topic

New Motivation: How do we find these *groups* of nodes?

Betweenness Centrality and Girvan-Newman

How else?

networkx.org/documentation/stable/reference/algorithms/community.html

NetworkX
Network Analysis in Python

Install Tutorial Reference Gallery Developer Releases Guides

- Centrality
- Chains
- Chordal
- Clique
- Clustering
- Coloring
- Communicability
- Communities
- Components
- Connectivity
- Cores
- Covering
- Cycles
- Cuts
- D-Separation
- Directed Acyclic Graphs
- Distance Measures
- Distance-Regular Graphs
- Dominance
- Dominating Sets
- Efficiency
- Eulerian
- Flows
- Graph Hashing
- Graphical degree sequence
- Hierarchy
- Hybrid
- Isolates

# Communities

Functions for computing and measuring community structure.

The functions in this class are not imported into the top-level `networkx` namespace. You can access these functions by importing the `networkx.algorithms.community` module, then accessing the functions as attributes of `community`. For example:

```
>>> from networkx.algorithms import community
>>> G = nx.barbell_graph(5, 1)
>>> communities_generator = community.girvan_newman(G)
>>> top_level_communities = next(communities_generator)
>>> next_level_communities = next(communities_generator)
>>> sorted(map(sorted, next_level_communities))
[[0, 1, 2, 3, 4], [5], [6, 7, 8, 9, 10]]
```

On this page
- Bipartitions
- K-Clique
- Modularity-based communities
- Tree partitioning
- Label propagation
- Louvain Community Detection
- Fluid Communities
- Measuring partitions
- Partitions via centrality measures
- Validating partitions

## Bipartitions

Functions for computing the Kernighan–Lin bipartition algorithm.

| | |
|---|---|
| `kernighan_lin_bisection`(G[, partition, ...]) | Partition a graph into two blocks using the Kernighan–Lin algorithm. |

## K-Clique

`k_clique_communities`(G, k[, cliques]) Find k-clique communities in graph using the

## Modu

Functions for detecting communities based on modularity.

NetworkX has several community detection techniques

# This Lecture's Learning Objectives

Describe at least three network centrality metrics

Use the Girvan-Newman method to identify communities in graphs

# Questions?

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