NETWORK SCIENCE OF ONLINE INTERACTIONS

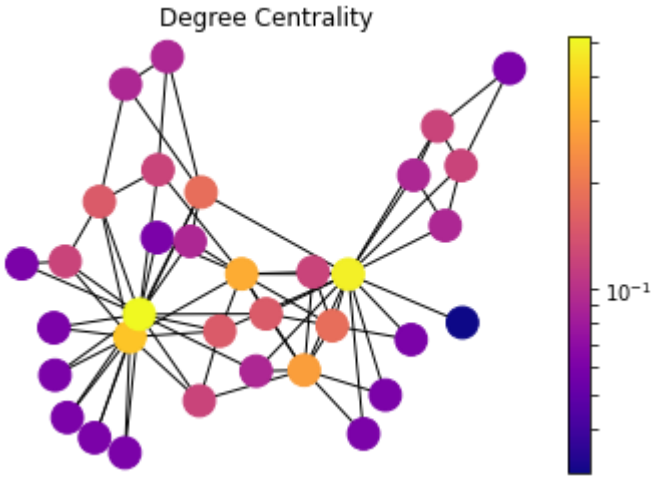
Chapter 4: Directions and Weights

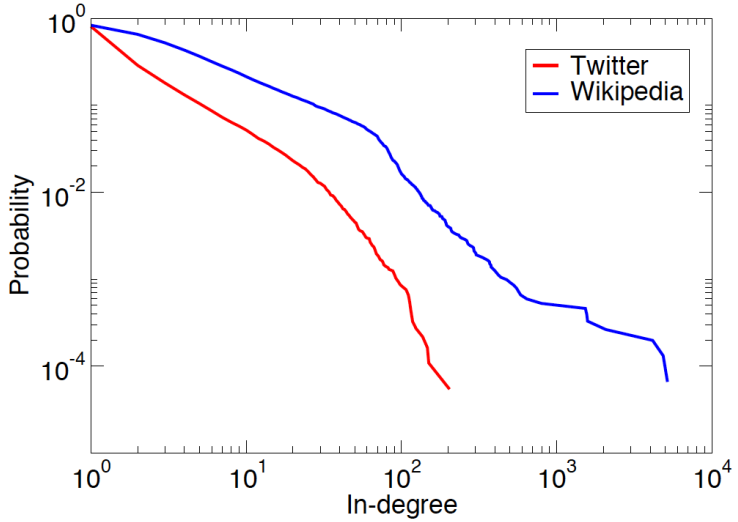
Joao Neto

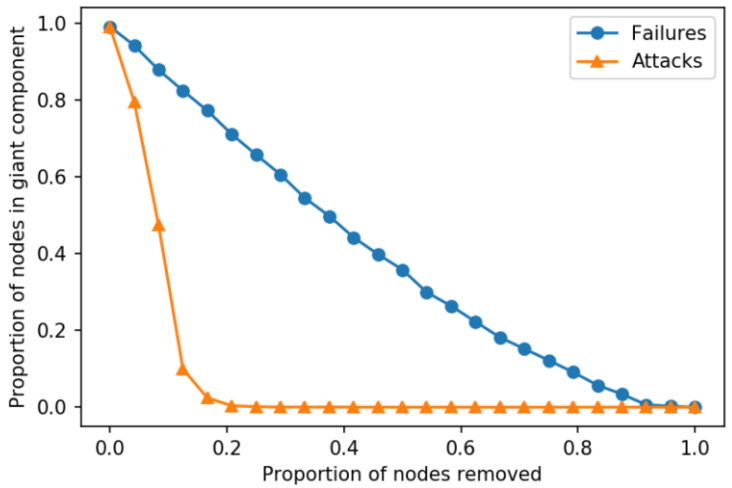
12/May/2023

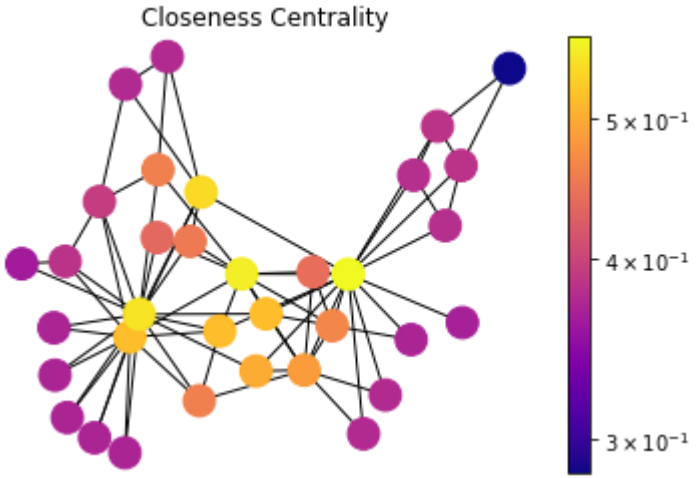
SUMMARY

- Different node/link centrality measures with different goals
  - Degree, closeness and betweenness centrality
- Statistical distributions of social networks are usually heavy-tailed
  - Careful when analysing/plotting it
- Networks can have non-intuitive properties (friendship paradox)
- Networks are robust against random failure, but weak against target attack



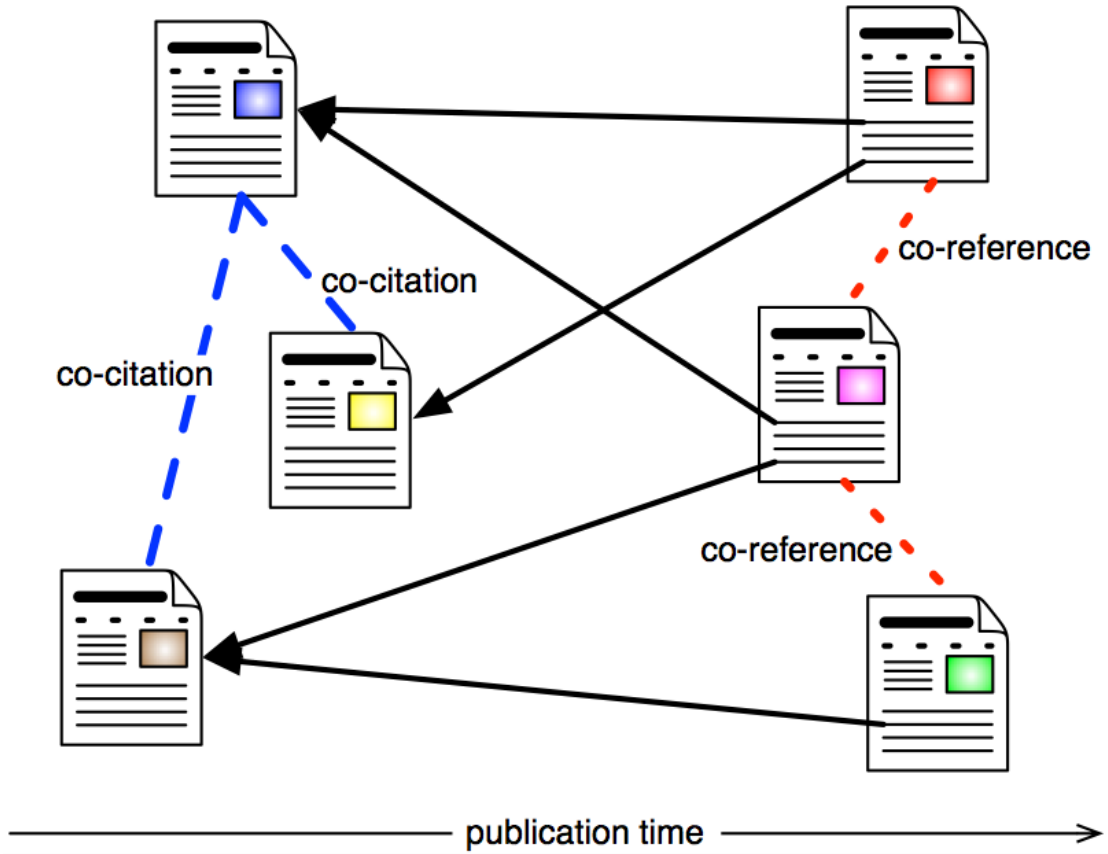






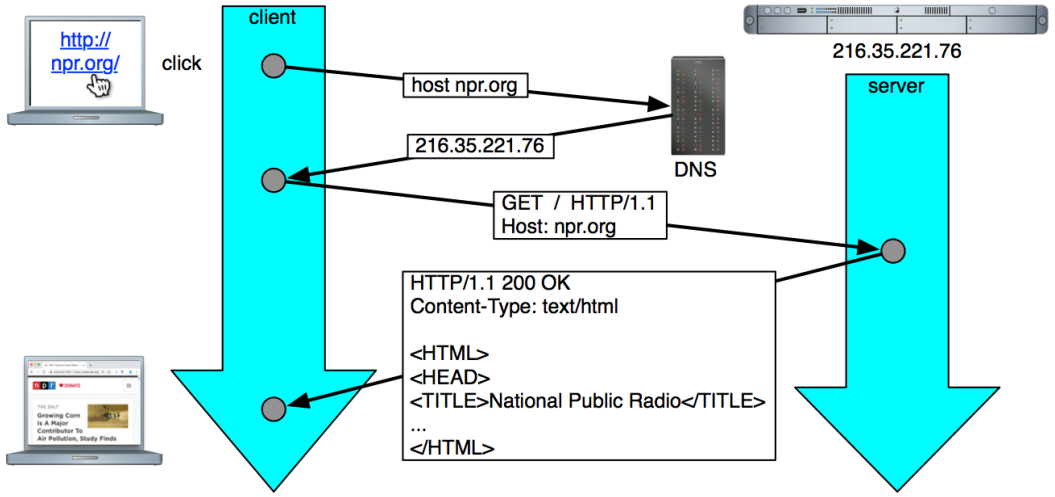
4.1 DIRECTED NETWORKS

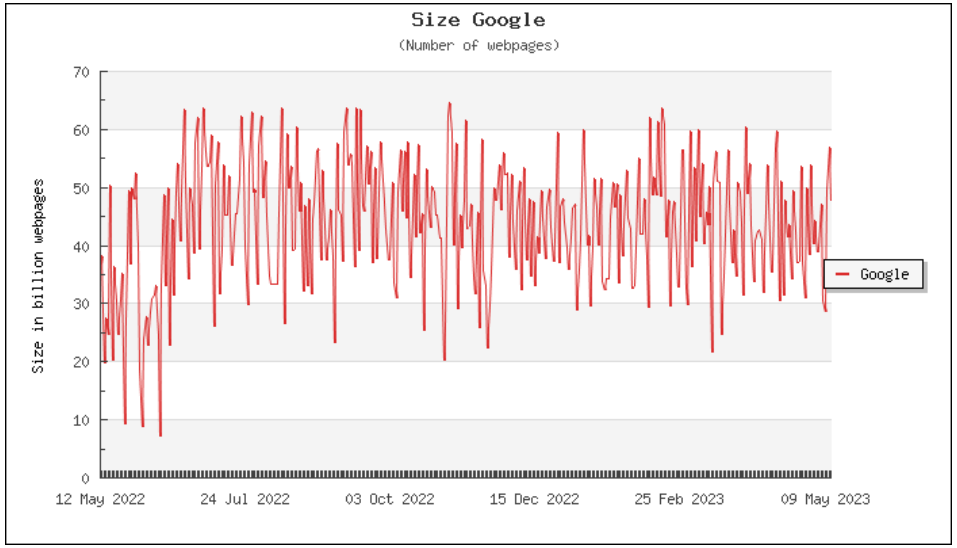
- Many real-world networks are directed
- Can be due to asymmetric, temporal interactions
- Many metrics split
- Degree
  - In-degree
  - Out-degree
- Component
  - In-component
  - Out-component



4.2 THE WEB

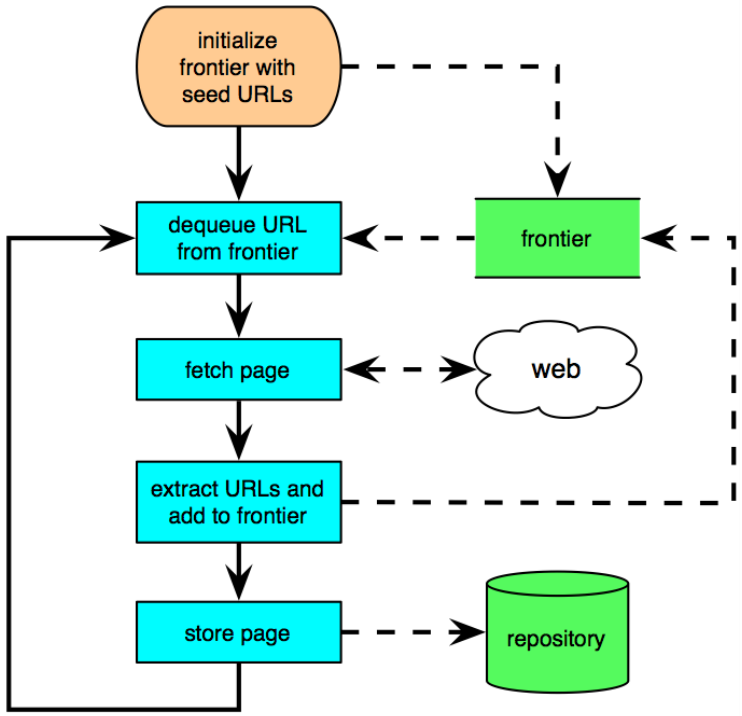
- World-Wide Web
  - nodes are pages with URLs
  - (directed) links are hyperlinks between pages
- How big is the web?
  - Lots of dynamic, ephemeral links
  - Estimate: web search indexes
  - Google: about 40-50B pages
  - Other estimates: 6-100B pages
- How to search that?



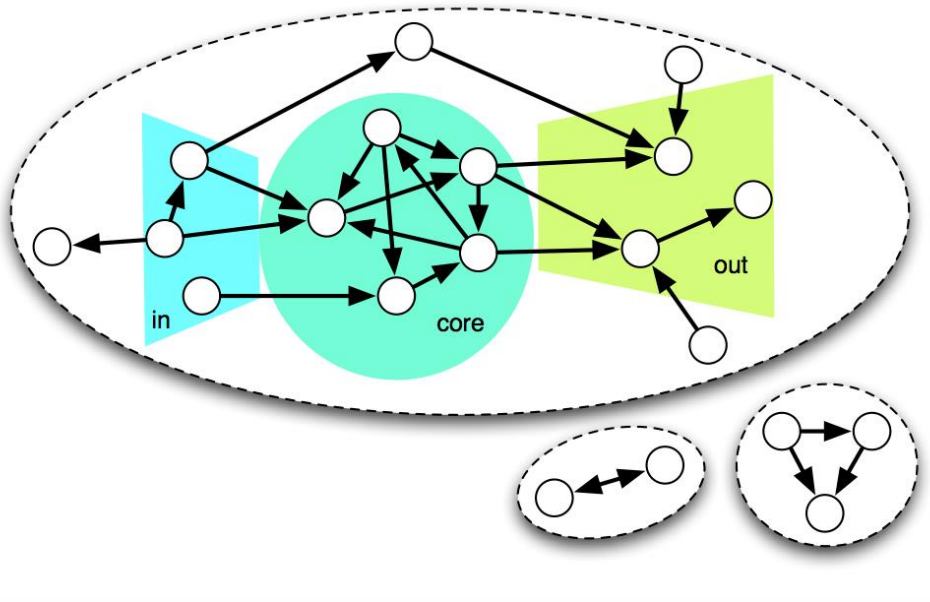


4.2 THE WEB

- Searching the web
  - Web crawler
  - Breadth-First Search

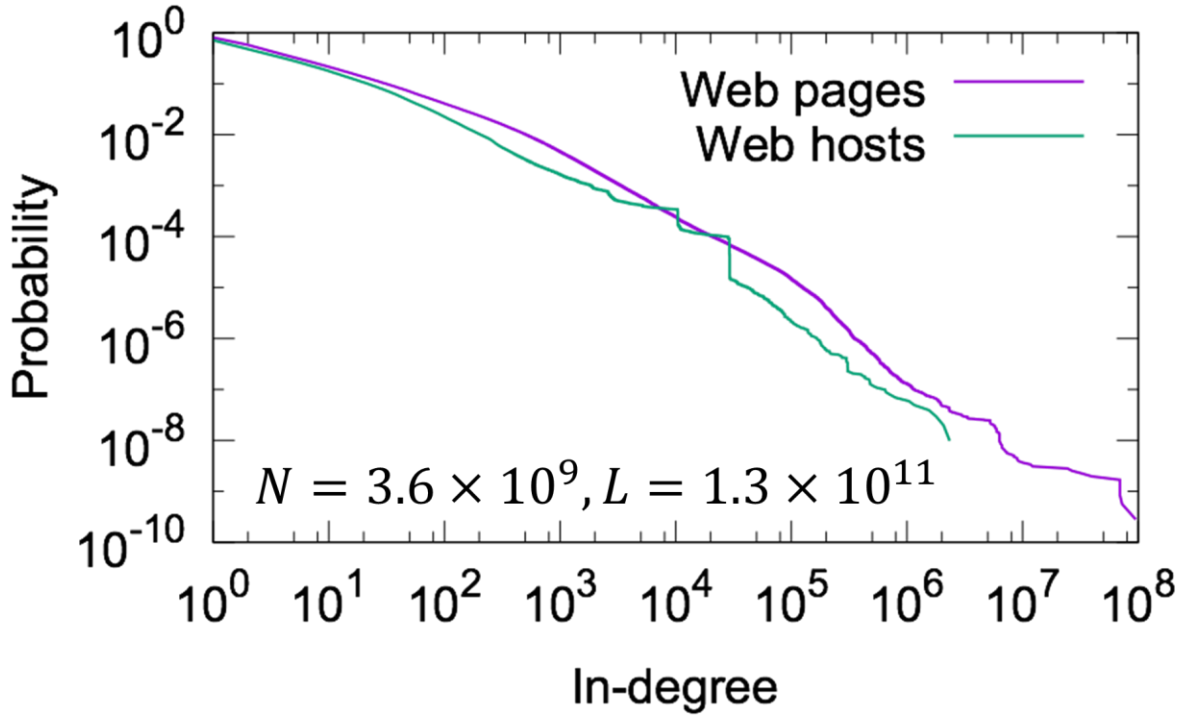


- Web graph structure
  - "Bow-tie" structure
  - Giant strongly connected component
  - Giant weakly connected component is >90% of the web



4.2 THE WEB

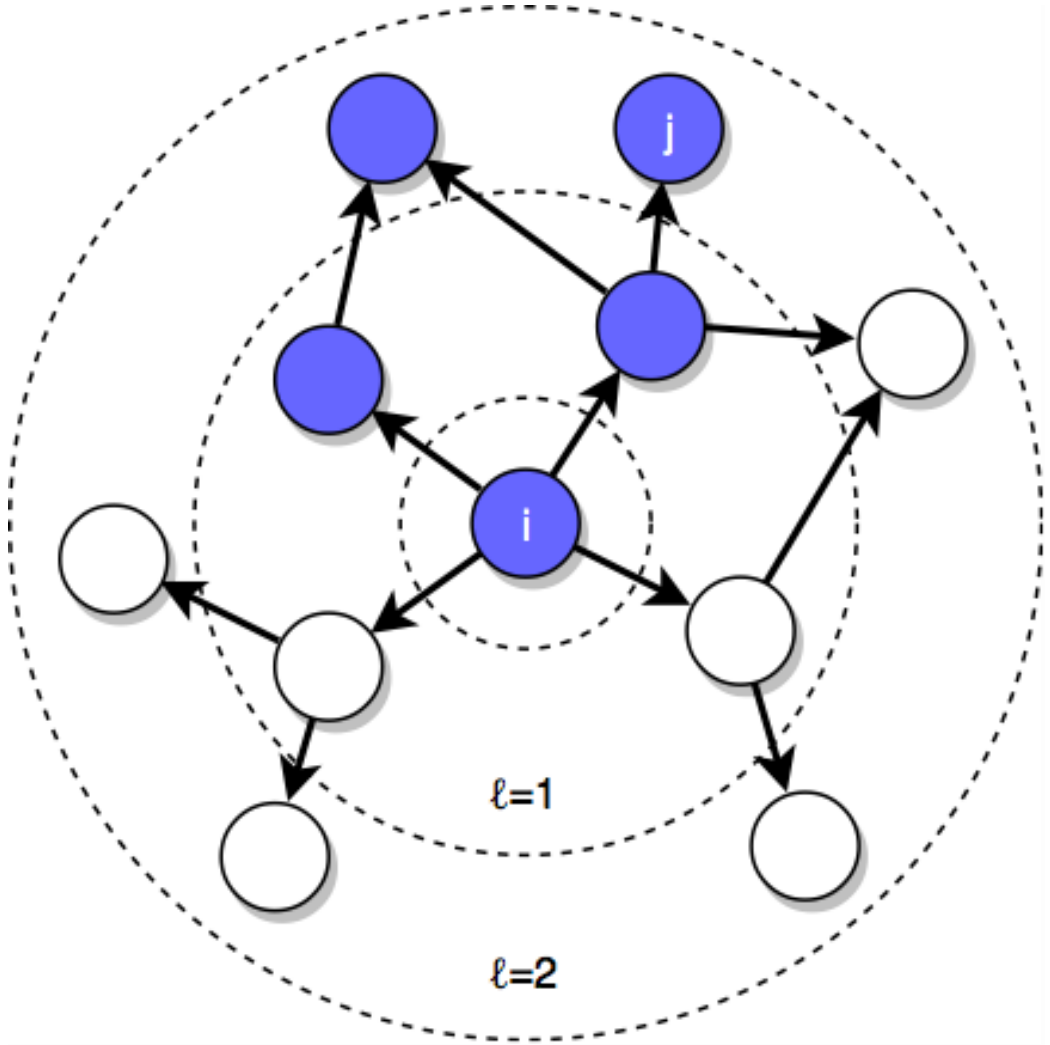
- Characterizing the web
  - In-degree: signature of popularity
  - Out-degree: signature of spam
- Heavy-tailed in-degree distribution
- 2012 web crawl
  - $N \approx 1.8 \times 10^9$
  - $\langle \ell \rangle \approx 13$
  - The web is an ultra small-world network



4.2 THE WEB

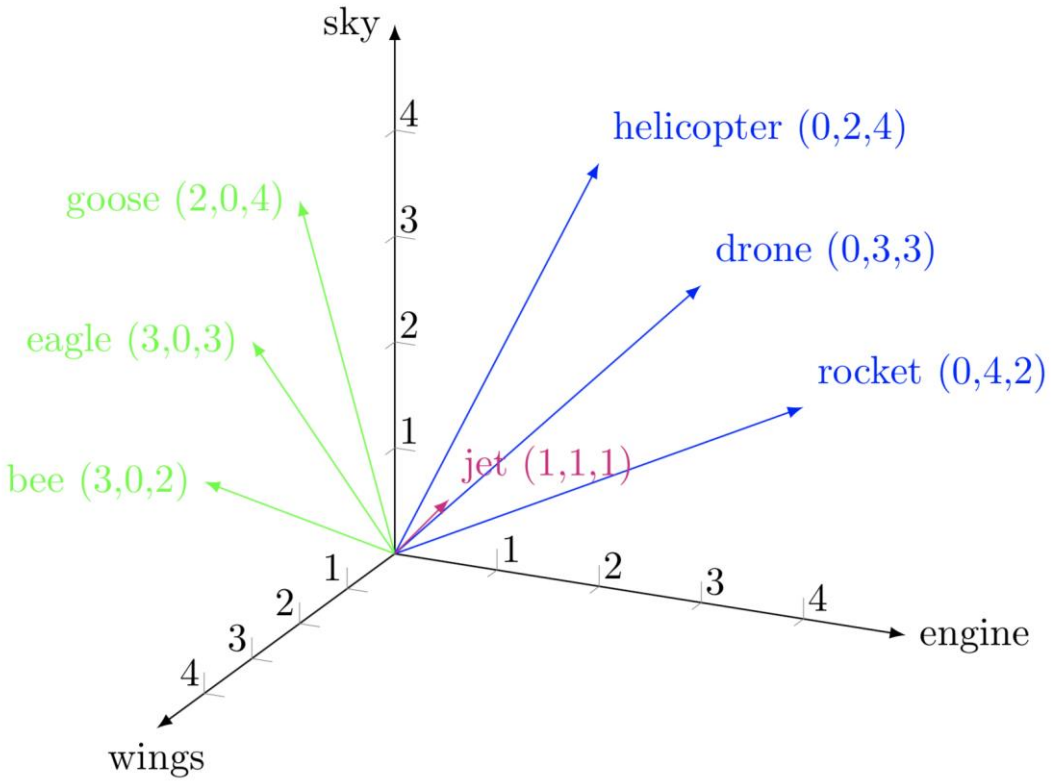
- Topical locality
  - Pages with related content tend to link (homophily)
  - Links are not random
    - Browsing is possible
    - BFS crawling is efficient

- **Topic drift**: as distance increases, similarity decreases



4.2 THE WEB

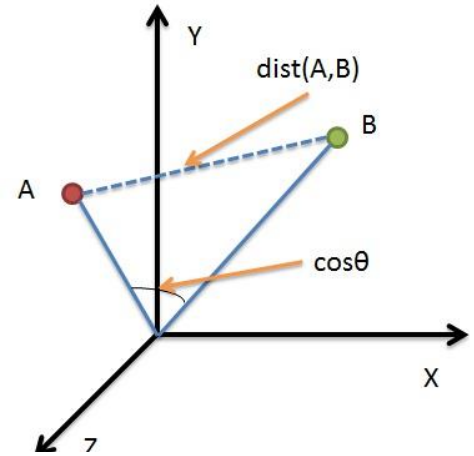
- How to measure similarity?
- **Word embedding**
  - Text content is represented as a vector in some space:
    - $\vec{d} = \{w_{d,1}, \dots, w_{d,n_t}\}$
  - One dimension for each term in your **dictionary**
  - Dictionary can be the entire language (~171K for English)
  - Smaller, expertly designed dictionaries are often more useful
  - **Dimensionality-reduction** techniques can help
- Similar content stays close in that space
- Weight (norm) of a term usually proportional to frequency

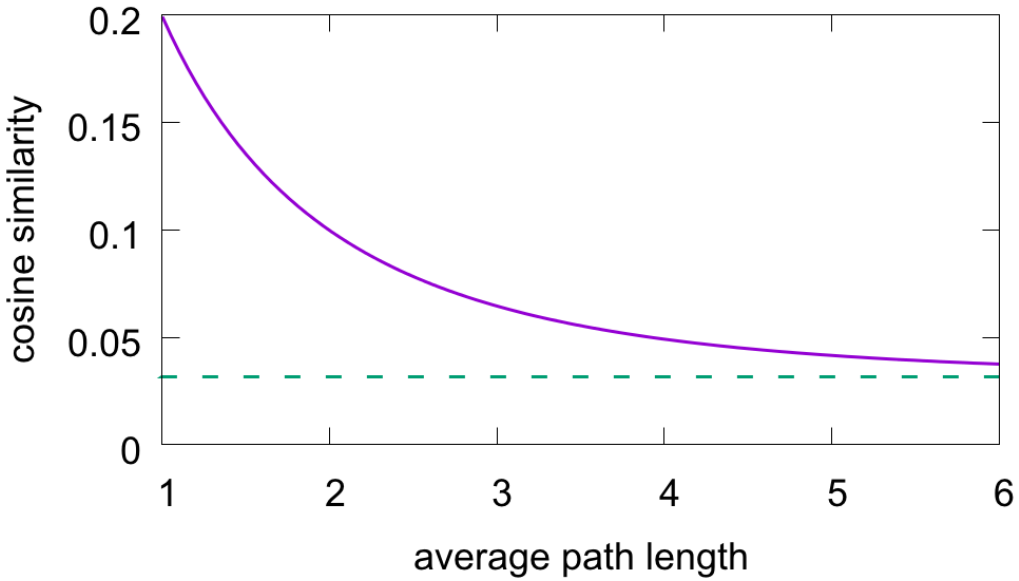


4.2 THE WEB

- Cosine similarity
  - $\cos\left(\overrightarrow{d_1}, \overrightarrow{d_2}\right) = \frac{\overrightarrow{d_1}}{\|d_1\|} \cdot \frac{\overrightarrow{d_2}}{\|d_2\|}$

- What is the average cosine similarity of webpages?
  - Decays with distance
  - Average of 100 seeds x 100 topics x 6 depths of a web search
  - Shows topic locality

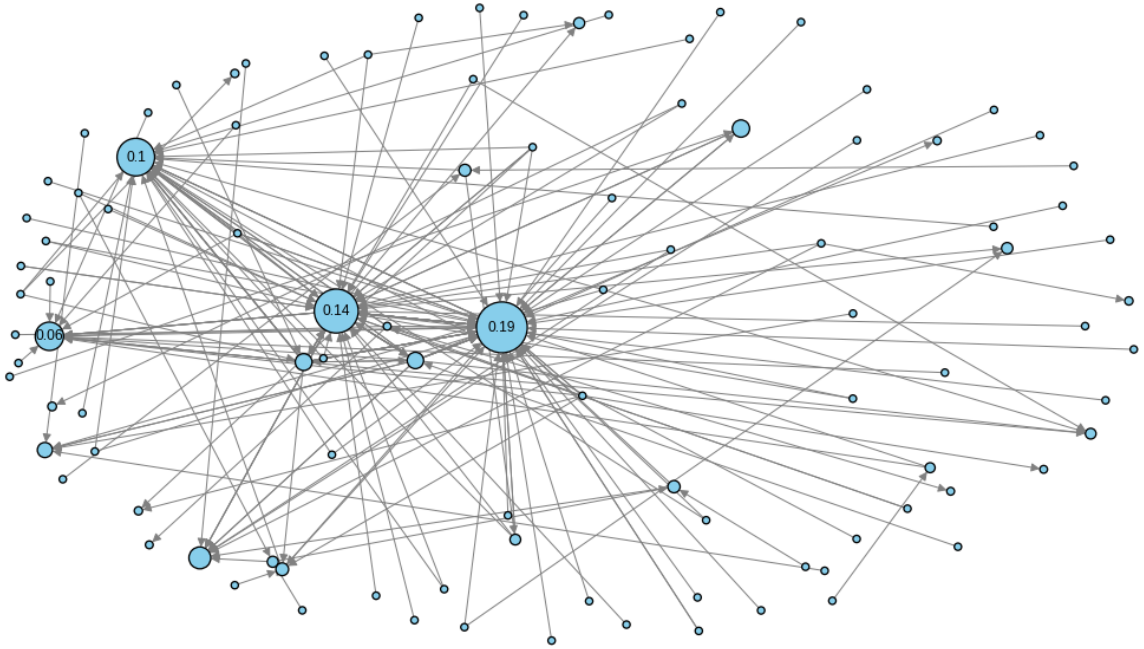




4.3 PAGERANK

- How to measure centrality of webpages?
- Centrality metrics so far
  - Easily gamed (spam)
  - Not iterative (doesn't scale)
  - Don't reproduce browsing behaviour

- **PageRank**
  - Created in 1998, backbone of Google
  - Ranks pages
  - Measures page prestige




```
PR_dict = nx.pagerank(D)
```

4.3 PAGERANK

- **Random surfer model:**
  - random links are clicked
  - Browsing can stop and start somewhere else
- **Random walk** model modified with random jumps (**teleportation**)
- PageRank of node i is given by $R(i)$
  - $\sum_i R(i) = 1$
- Defined recursively
- Free parameter: teleportation α
  - Typically, $\alpha = 0.15$
  - Dual role: dampens random walk, prevents rank sinks

Power method to calculate PageRank:

- Initialize each node with

$$R_0 = 1/N$$

- At each iteration *t*, loop over nodes and update PageRank of each node *i* via this recursive equation:

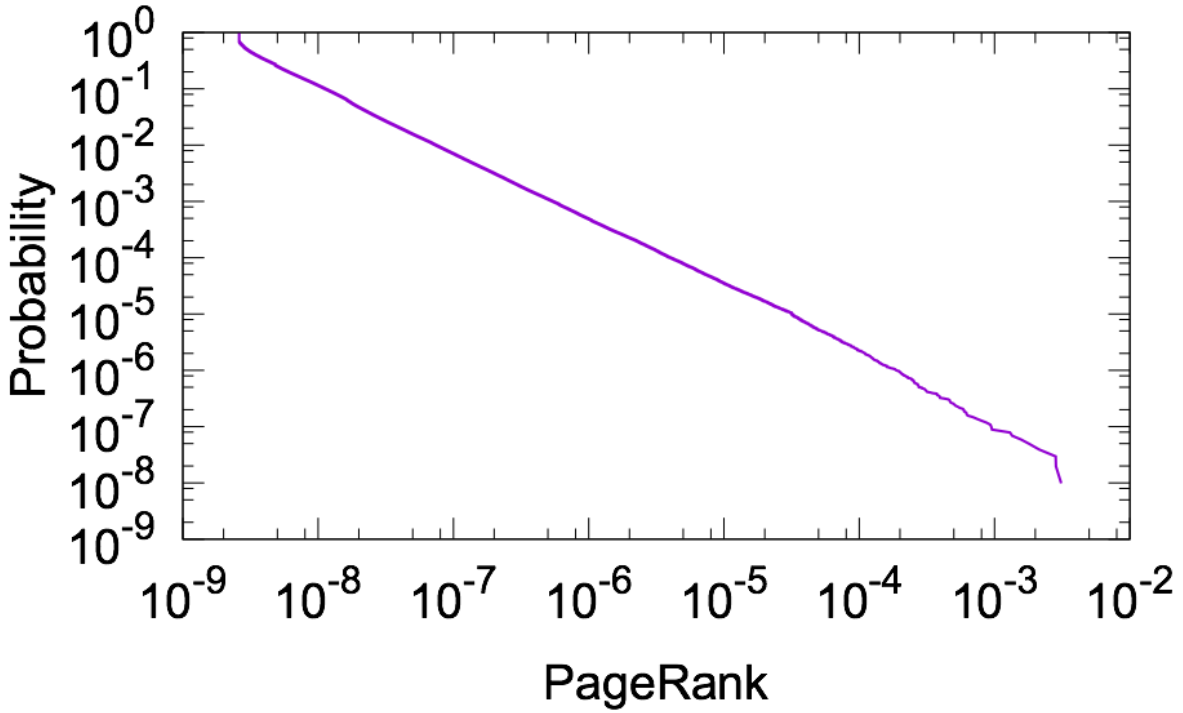
$$R_t(i) = \frac{\alpha}{N} + (1-\alpha) \sum_{j \in pred(i)} \frac{R_{t-1}(j)}{k_{out}(j)}$$

probability to land on *i* by teleportation

probability to land on *i* by random surfing

4.3 PAGERANK

- What is the PageRank distribution?
  - Again, power-law
- How does it differ from e.g. in-degree?
  - Depends on the *prestige* of who links to you
  - Harder to game by spam
- PageRank exemplifies the benefit of having a targeted model (browsing) when choosing a model
  - Better sorting -> $1.5T market cap

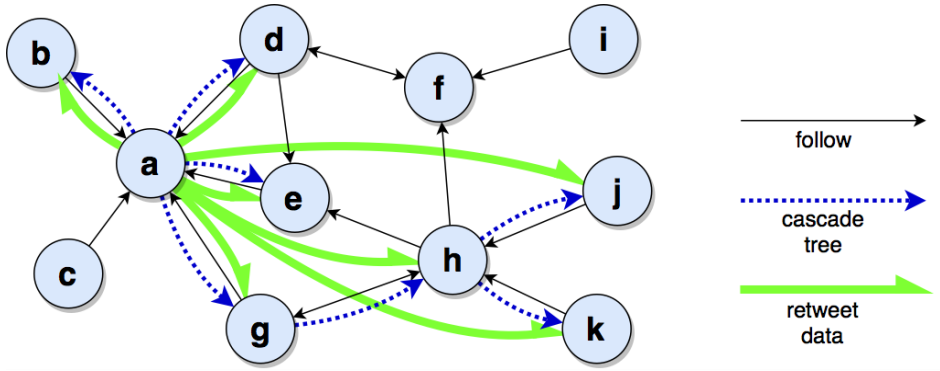


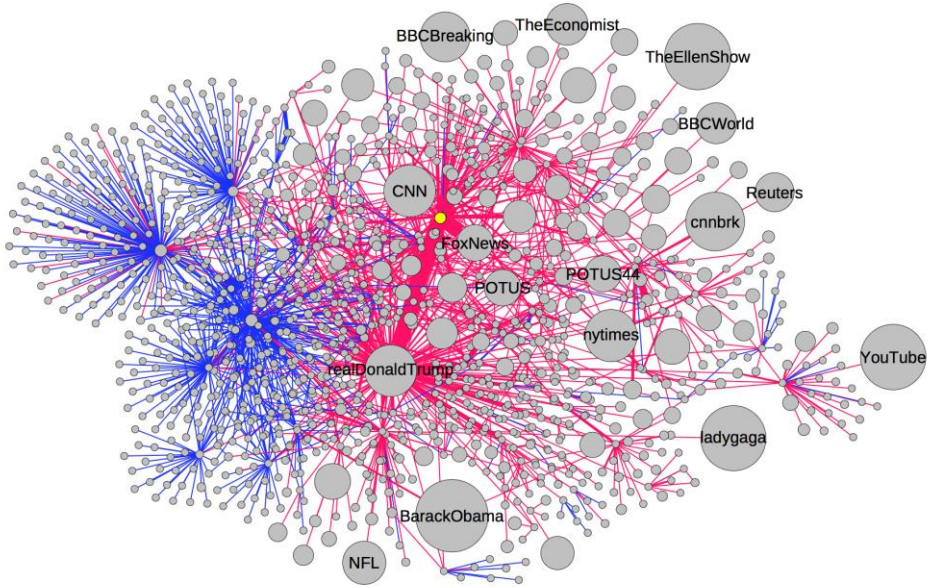
4.4-4.6 INTERESTING WEIGHTED NETWORKS

- Many real-world networks are weighted

- Information diffusion networks
  - How does content spreads/diffuses/percolates in a network?
  - How can we measure the *influence* of the nodes in this diffusion?

- Best-studied case: Twitter
  - Retweet/reply processes
  - User networks





CONTENT VIRALITY

- How to measure content virality?
  - Number of users exposed
  - Tree properties
- Tree properties
  - **Structural virality**
  - $v = \frac{1}{n(n-1)} \sum_{i,j} d_{ij}$
  - Average path length of the diffusion tree

Figure 1 **A Schematic Depiction of Broadcast vs. Viral Diffusion, Where Nodes Represent Individual Adoptions and Edges Indicate Who Adopted from Whom**

Goel et al. 'The Structural Virality of Online Diffusion'. *Management Science*. doi:10.1287/mnsc.2015.2158 (2015)

NODE INFLUENCE

- Measuring influence
  - Number of followers (in-degree in follower network)
  - Number of users exposed (out-degree in retweet network)
  - Number of retweets (out-strength in retweet network)
  - Fraction of retweets to followers
  - **New**: number of impression

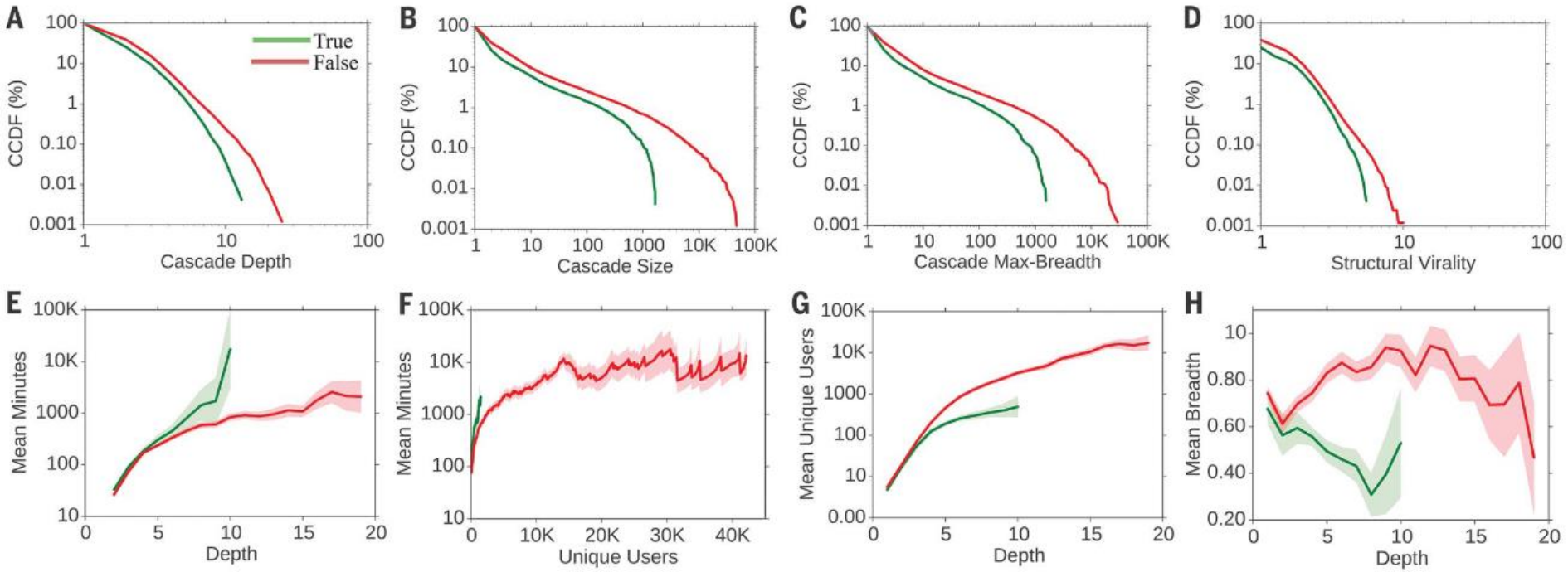
MEASURING VIRALITY

- Measuring virality is tricky
- Widely-cited paper from 2018
  - **"fake news spread 6x more than true news"**
  - Fake news spread **more** and **more** broadly

SOCIAL SCIENCE

The spread of true and false news online

Soroush Vosoughi,[1] Deb Roy,[1] Sinan Aral[2]*

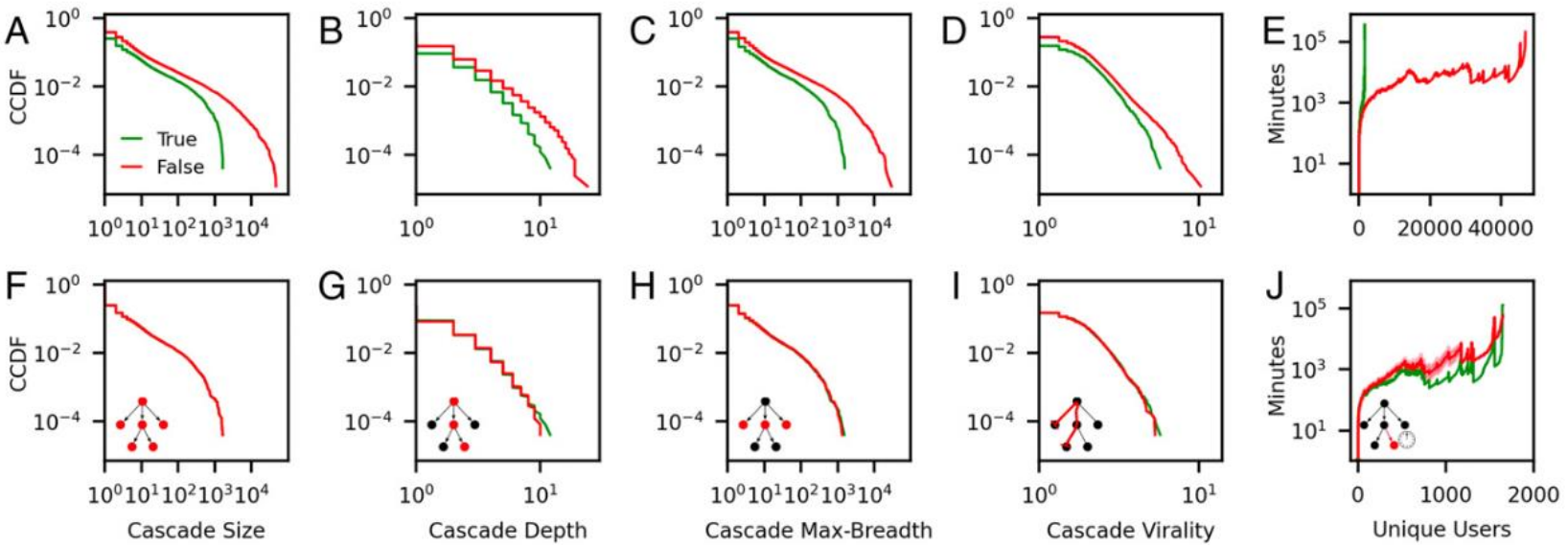


MEASURING VIRALITY

- 2021 re-analysis
- Cascade match:
  - If we match the **size** of the diffusion cascades, what happens to the other metrics?
  - The fake news tweets spread *more*, but not *more broadly*
  - May be due to data sampling bias
  - Requires processes to be extremely similar to work

Comparing information diffusion mechanisms by matching on cascade size

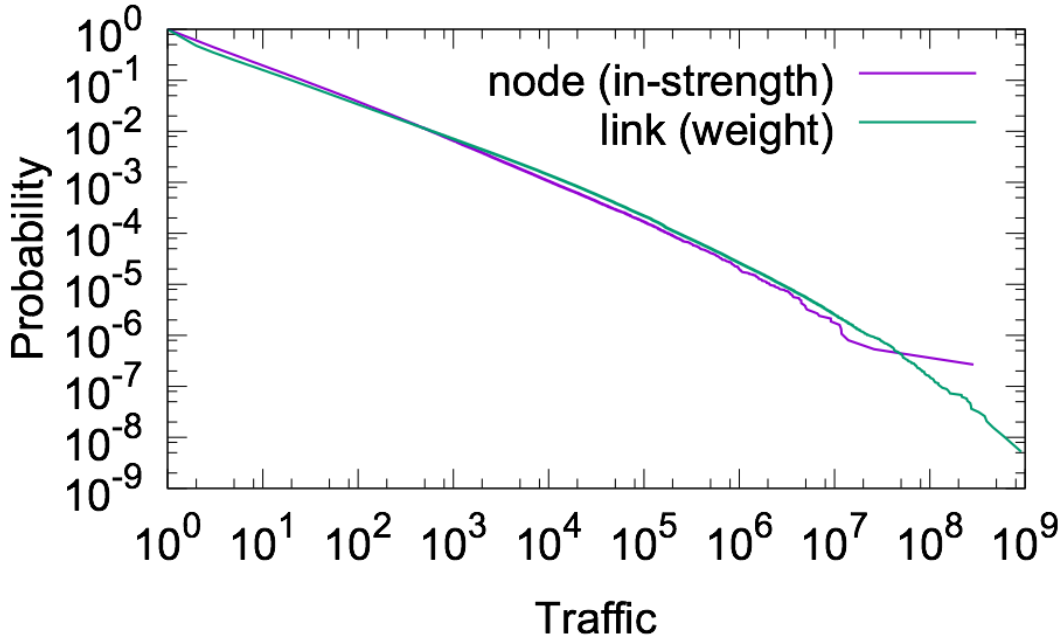
Jonas L. Juul[a,1] and Johan Ugander[b,1]



Juul & Ugander. '*Proceedings of the National Academy of Sciences* 118, no. 46. doi:10.1073/pnas.2100786118.

4.7 WEIGHT HETEROGENEITY

- Weights can span many orders of magnitude
- Do we need all of them?
- Thresholding: prunes weak links
  - Makes it more manageable
  - Can more easily reveal important structure
  - **Dangerous**
- Simple pruning:
  - Removes links with $w_{ij} < \alpha$
  - Can quickly break and mischaracterize the network if heavy-tailed
- Per-node pruning
  - Per-node threshold, removes less-important links of each node



$$p_{ij} = \left(1 - \frac{w_{ij}}{s_i}\right)^{k_i - 1} < \alpha$$

SUMMARY

- Lots of directed networks are heavy-tailed both in degree and weights
- Word embedding is a powerful tool to study content networks (e.g. social media)
- Diffusion models are very useful
  - Metrics that model diffusion can excel (PageRank)
  - Can be used to study spread of misinformation in social media