universität koblenz
weiter:denken

Network Theory and Dynamic Systems
06. Graph-Based View of the Web
SOSE 2025

Dr. -Ing. Stefania Zourlidou

Institute for Web Science and Technologies
Universität Koblenz

WeST
Web Science and Technologies

Recap from Previous Lecture

universität koblenz
weiter:denken

- Triadic Closure
- The Strength of Weak Ties
- Tie Strength and Network Structure in Large-Scale Data Betweenness
- Tie Strength, Social Media, and Passive Engagement
- Closure, Structural Holes, and Social Capital
- Betweenness Measures and Graph Partitioning

Objectives of this Lecture

universität koblenz
weiter:denken

- Directed Networks
- The Web
- PageRank
- Weighted Networks
- Information and Misinformation Spread
- Co-Occurrence Networks
- Weight Heterogeneity

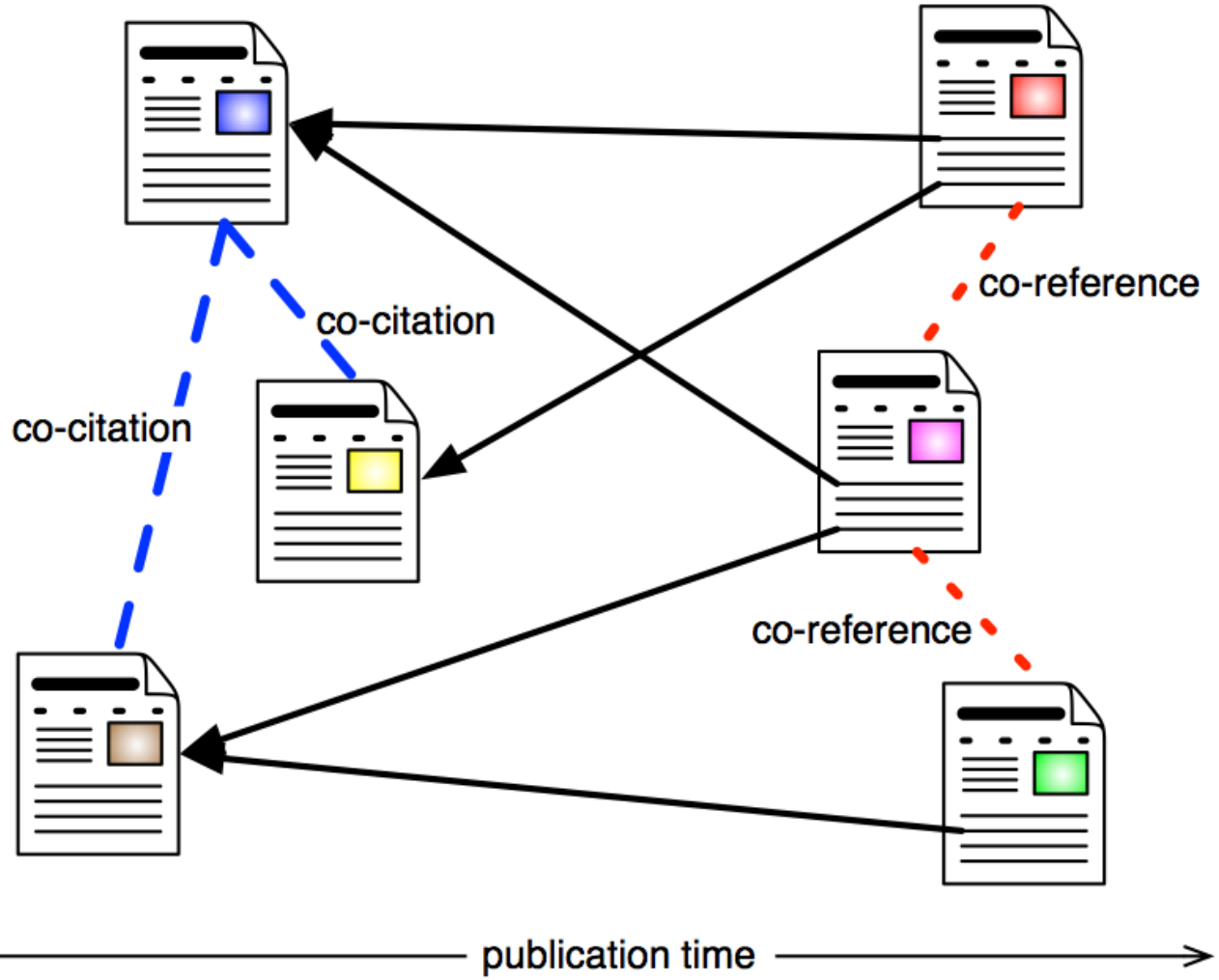
1. Directed Networks

Real-world Networks: Direction and Weight Matter

- Many real-world systems are best represented as directed and/or weighted networks
- We've seen multiple examples:
  - Food webs: links go from insect to bird, with weights indicating the number of individuals consumed
  - Web and Wikipedia: hyperlinks point from a source page to a target page, with weights possibly reflecting click frequency
  - Social networks: links may be weighted by interaction volume—such as retweets, mentions, likes, or comments

Directed Networks: Key Concepts

- Links point from a source node to a target node
- A node's degree is separated into:
  - In-degree: number of incoming links
  - Out-degree: number of outgoing links
- Paths must follow the direction of links
- Examples:
  - Email, Web, Wikipedia
  - Citation networks:
    - Nodes = scholarly articles
    - Links = citations of other works
    - Link direction follows publication time (older → newer is invalid)



2. The Web

Web Graph ≠ The Internet

universität koblenz
weiter:denken

- The Web graph (or Web link graph) is not the same as the entire Internet
- In the Web graph:
  - Nodes represent individual Web pages, identified by URLs
  - Edges are hyperlinks between pages
- The Host graph (or Site graph) is a higher-level abstraction:
  - Nodes are websites, identified by hostnames or domain names
  - An edge exists if there is any hyperlink between pages on the two sites
- Wikipedia forms a subnetwork within the Web graph

Web Graph vs. Host Graph

universität koblenz
weiter:denken

	Web Graph	Host Graph
Nodes	Individual **web pages** (e.g., https://example.com/page1)	**Websites** or **hosts** (e.g., example.com)
Edges	**Hyperlinks between pages**	**Aggregate links** between hosts: at least one hyperlink between any pages
Granularity	**Fine-grained** — shows exact page-to-page connectivity	**Coarser** — shows site-level or domain-level connectivity
Size	Much **larger**, due to many pages per site	**Smaller** — nodes represent hosts, not individual pages
Use case	Used for **detailed link analysis**, PageRank, crawling strategies	Used for **visualizing site-to-site relations**, host-level ranking
Example link	page1.html → page2.html	example.com → anotherdomain.org

Web Graph ≠ The Internet

universität koblenz
weiter:denken

	Internet	Web Graph
What it is	The global **network of computers and devices**	A **graph structure of webpages** connected by **hyperlinks**
Nodes	Physical **machines**: routers, servers, computers	**Web pages** (identified by URLs)
Links/Edges	**Physical/data connections** (e.g., Ethernet, fiber)	**Hyperlinks** (from one page to another)
Protocols	Includes TCP/IP, DNS, etc.	Uses **HTTP/HTTPS** protocols
Scope	Covers all digital communication infrastructure	Subset: Only about the **content** and **structure of the Web**
Examples	Email servers, DNS infrastructure, FTP servers	Google linking to Wikipedia, Wikipedia pages linking to each other

Web History

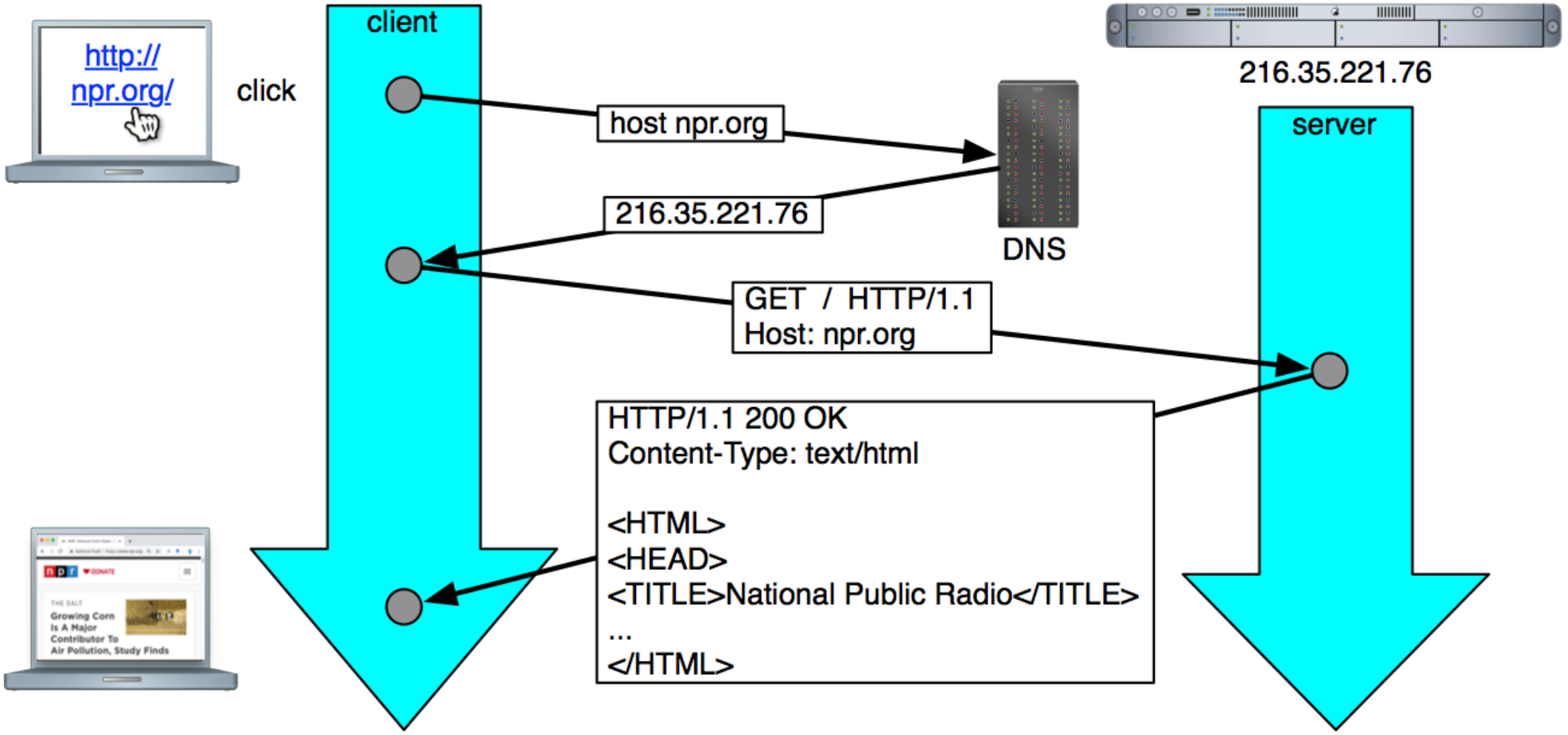
- The concept of hypertext predates the Web by decades, but bi-directional linking was hard to implement
  - It required a central authority to coordinate and maintain mutual links
- In the early 1990s, Tim Berners-Lee introduced the World Wide Web with a simpler model:
  - One-directional hyperlinks from the source page
  - This enabled a decentralized implementation
  - No need for the target page to acknowledge or maintain the link
  - Broken links were acceptable — no global consistency required
- Berners-Lee's key contributions included:
  - URL – Uniform Resource Locator: standardized way to reference a page
  - HTML – HyperText Markup Language: used to write and structure web content
  - HTTP – HyperText Transfer Protocol: enabled communication between browsers and web servers

Lowering the Barriers to Publishing on the Web

universität koblenz
weiter:denken

- In the early days, only a few individuals and organizations had the expertise to set up servers and write web pages
- Over time, a series of developments dramatically lowered the technical and financial barriers to creating and sharing content:
  - Virtual hosting – made it easier and cheaper to host websites
  - Blogs – enabled non-technical users to publish regularly with minimal setup
  - Wikis – allowed collaborative editing and knowledge sharing
  - Web 2.0 – introduced platforms for user-generated content, tagging, and sharing
  - Social media – further simplified publishing and enabled viral content distribution
  - Microblogs – platforms like Twitter and Facebook allowed quick, informal updates with massive reach

How the Web Works



Web Crawlers (1/6)

universität koblenz
weiter:denken

- A **Web client** is any program that retrieves information from the Web by using the HTTP protocol to connect to a Web server
  - The most common example is the Web browser, used for interactive browsing by humans
- **Web crawlers** (also called *spiders* or *bots*) are automated programs that systematically download Web pages
  - Their main role is in search engines, where they gather and index content to make it searchable

Web Crawlers (2/6)

universität koblenz
weiter:denken

- **Web crawlers** gather information spread across billions of pages hosted on millions of servers worldwide
  - They bring this data to a central location for further analysis and mining
- A search engine processes the collected content to build an **index**
  - This index enables fast and efficient retrieval of pages matching user queries based on keywords and phrases
- Because the Web is constantly changing, crawlers continuously revisit pages to:
  - Detect new content
  - Remove deleted or outdated pages
  - Update moved or modified content

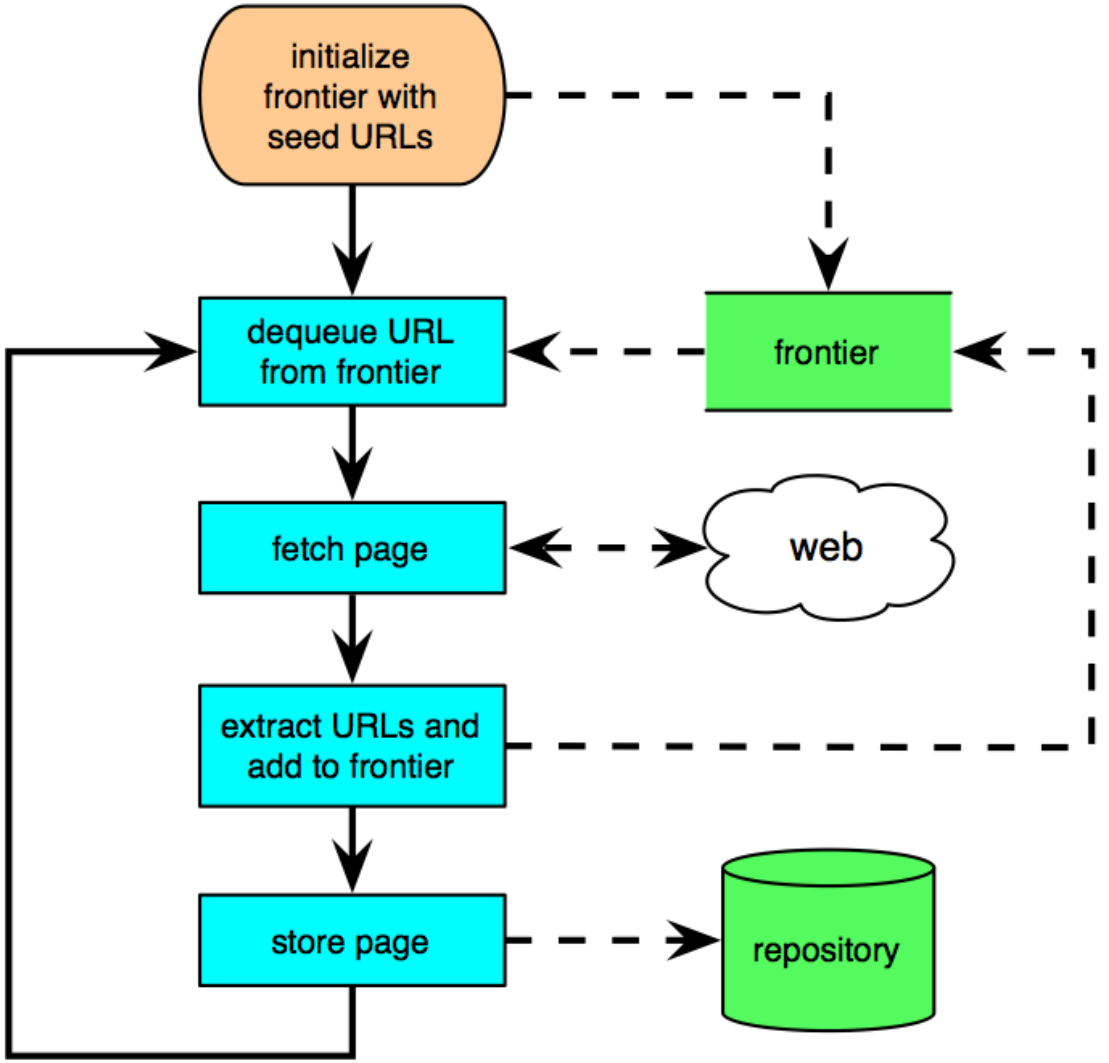
Web Crawlers (3/6)

- **Legitimate Uses:**
  - Business intelligence – analyzing market trends, competitors, and customer sentiment
  - Digital libraries – collecting and organizing online academic or archival content
  - Webometrics – studying the structure and impact of websites (e.g., link analysis, site rankings)
  - Research – gathering large-scale web data for studies in linguistics, social science, machine learning, etc.
- **Malicious Uses:**
  - Harvesting email addresses – to send unsolicited spam
  - Collecting personal information – for phishing, identity theft, or unauthorized profiling
  - Scraping protected content – violating terms of service or intellectual property rights

Web Crawler (4/6)

universität koblenz
weiter:denken

- A crawler performs a breadth-first traversal of the Web graph, similar to the BFS (Breadth-First Search) algorithm
- It begins with a set of high-quality seed pages
- It maintains a frontier — a queue of unvisited URLs
- Crawling process:
  - Dequeue a URL from the frontier
  - Fetch the corresponding Web page
  - Extract all hyperlinks from the page
  - Enqueue the new (unvisited) links into the frontier



Web Crawler (5/6)

universität koblenz
weiter:denken

- Although the basic crawling algorithm is simple, real-world crawling introduces many complications:
  - Scalability
    - Crawlers must operate across billions of pages and thousands of servers, often in distributed environments
  - Page revisit scheduling
    - Pages change over time — crawlers must decide when and how often to revisit them for freshness
  - Spider traps
    - Some sites generate infinite or near-infinite link paths (e.g., calendars), which can trap crawlers in loops

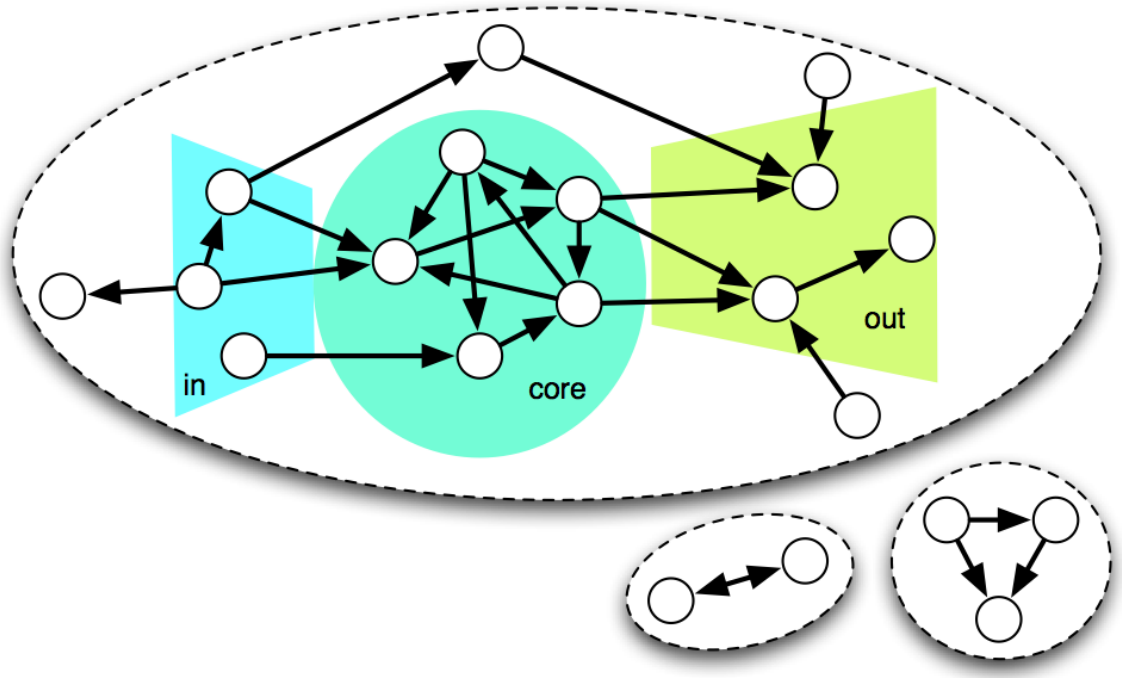
Web Crawler (6/6)

- Canonical URLs
  - The same content may appear under multiple URLs — crawlers must detect and consolidate duplicates
- Robust HTML parsing
  - Web pages often contain poorly structured or malformed HTML, requiring resilient parsers
- Ethical and legal concerns
  - Crawlers must respect robots.txt, avoid overloading servers, and handle privacy-sensitive content responsibly

The Bow-Tie Structure of the Web (1/2)

The Web graph has a characteristic "bow-tie" structure, consisting of several major components:

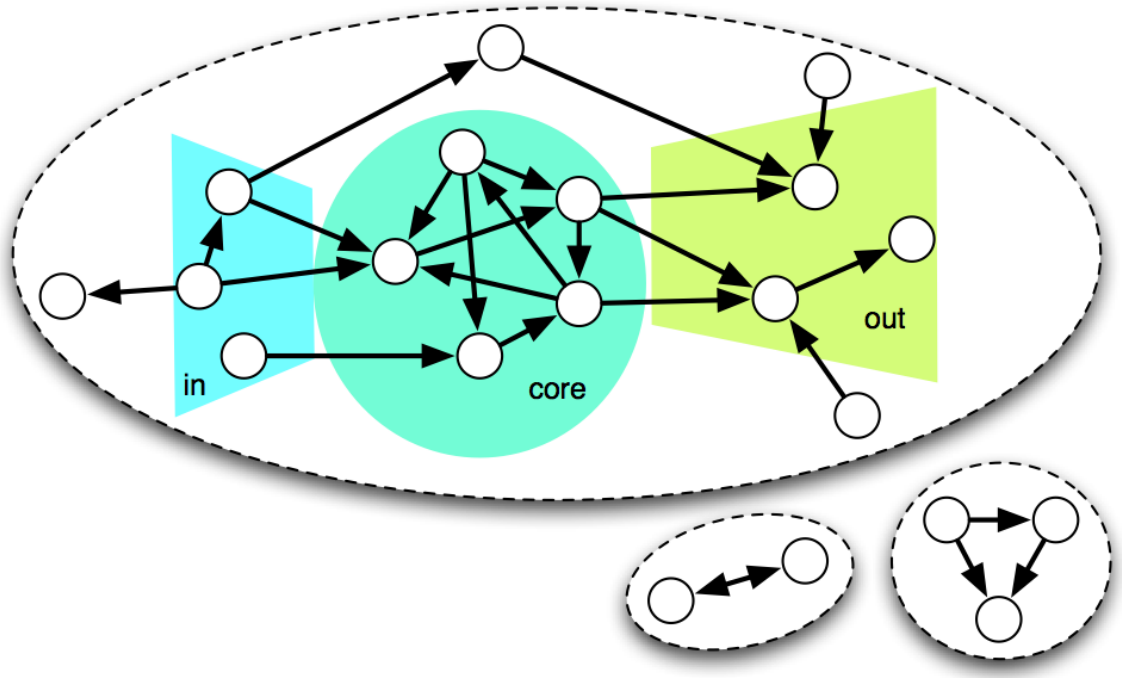
- **Core (SCC – Strongly Connected Component)**
  - A large strongly connected component where every page can reach every other page via directed links
  - Often referred to as the core of the Web
- **IN Component**
  - Pages that can reach the core, but cannot be reached from it
- **OUT Component**
  - Pages that can be reached from the core, but cannot reach back to it
- Together, the IN, SCC, and OUT make up over 90% of all pages — the giant weakly connected component



The Bow-Tie Structure of the Web (2/2)

universität koblenz
weiter:denken

- **Tendrils and Tubes**
  - Pages connected to IN or OUT, but not directly part of the main flow
  - **Tubes** link IN to OUT without going through the core
- **Disconnected Components**
  - Many **small isolated clusters** or **single pages** not connected to the giant component



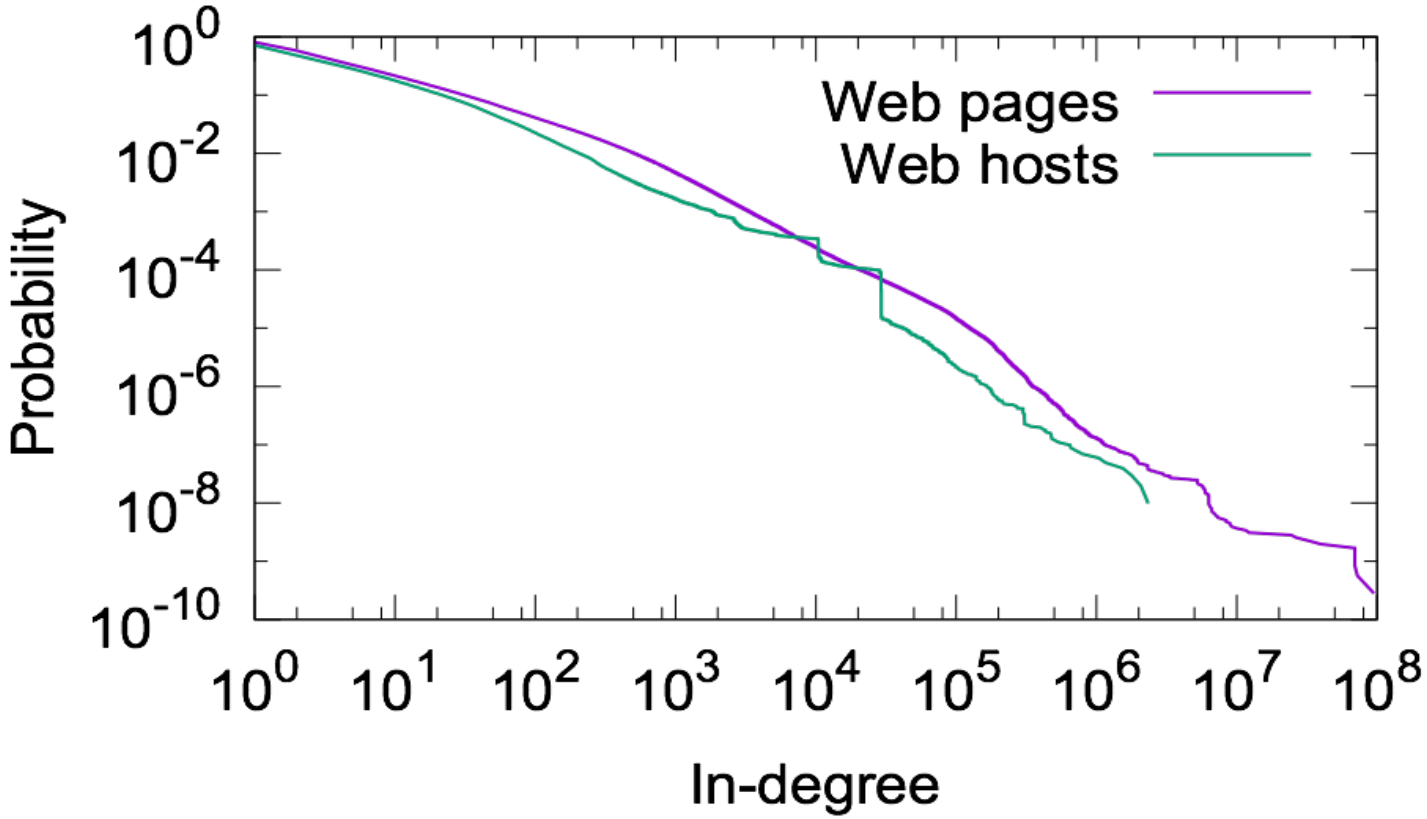
Web Degree Distributions

- We focus on **in-degree** distribution, which arises from collective behavior of authors
- Heavy tail with huge hubs:

$$\langle k_{in} \rangle \approx 10, \kappa \approx 40$$

- Out-degree distribution is less meaningful as it depends on individual content providers and can be gamed
  - E.g., **link farms** for **spamdexing**
- Short paths in core:

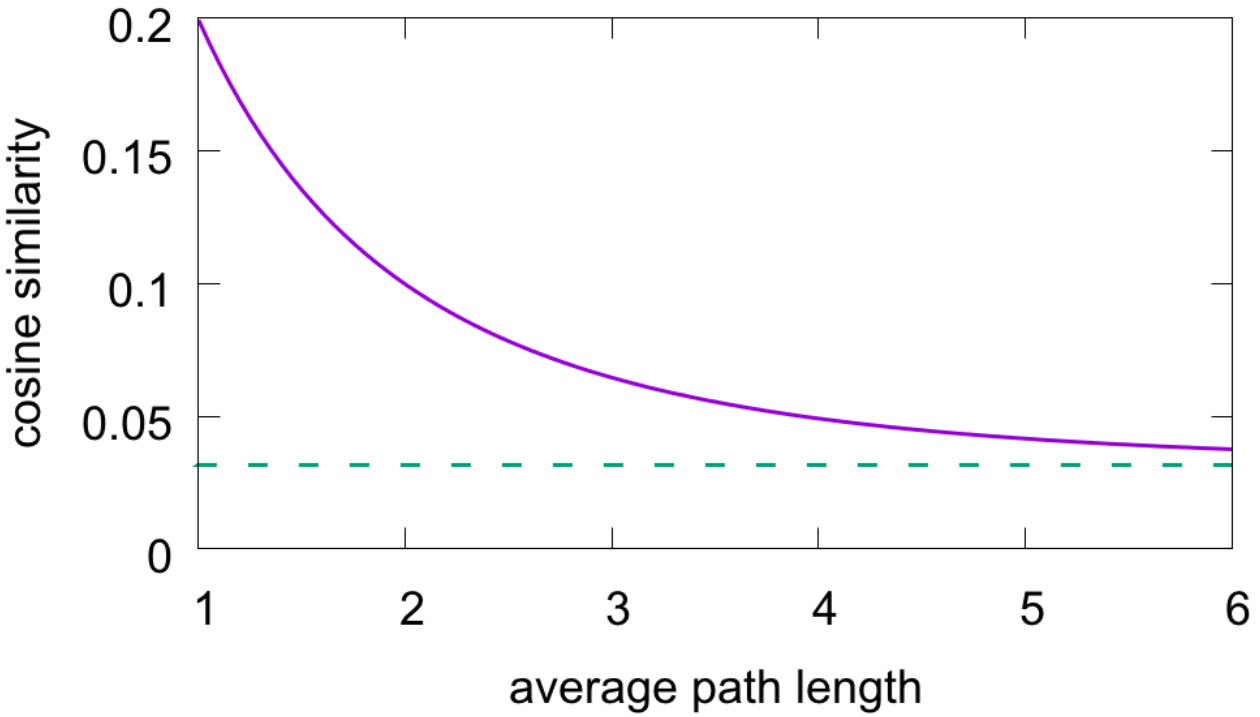
$$N = 3.6 \times 10^9, L = 1.3 \times 10^{11}$$



Cosine Similarity

- **Vector space model**: textual content of each page is represented as a vector
  - One dimension for each term in vocabulary (high-dimensional space!)
    - ✓ Remove or give lower weights to terms that are very common and therefore not meaningful
  - Or use deep learning to embed pages in lower-dimensional Euclidean space
- Measure similarity between two pages based on angle between vectors

$$\cos(\vec{d_1}, \vec{d_2}) = \frac{\vec{d_1}}{\| \vec{d_1} \|} \cdot \frac{\vec{d_2}}{\| \vec{d_2} \|} = \frac{\sum_t w_{d_1,t} w_{d_2,t}}{\sqrt{\sum_t w_{d_1,t}^2} \sqrt{\sum_t w_{d_2,t}^2}}$$

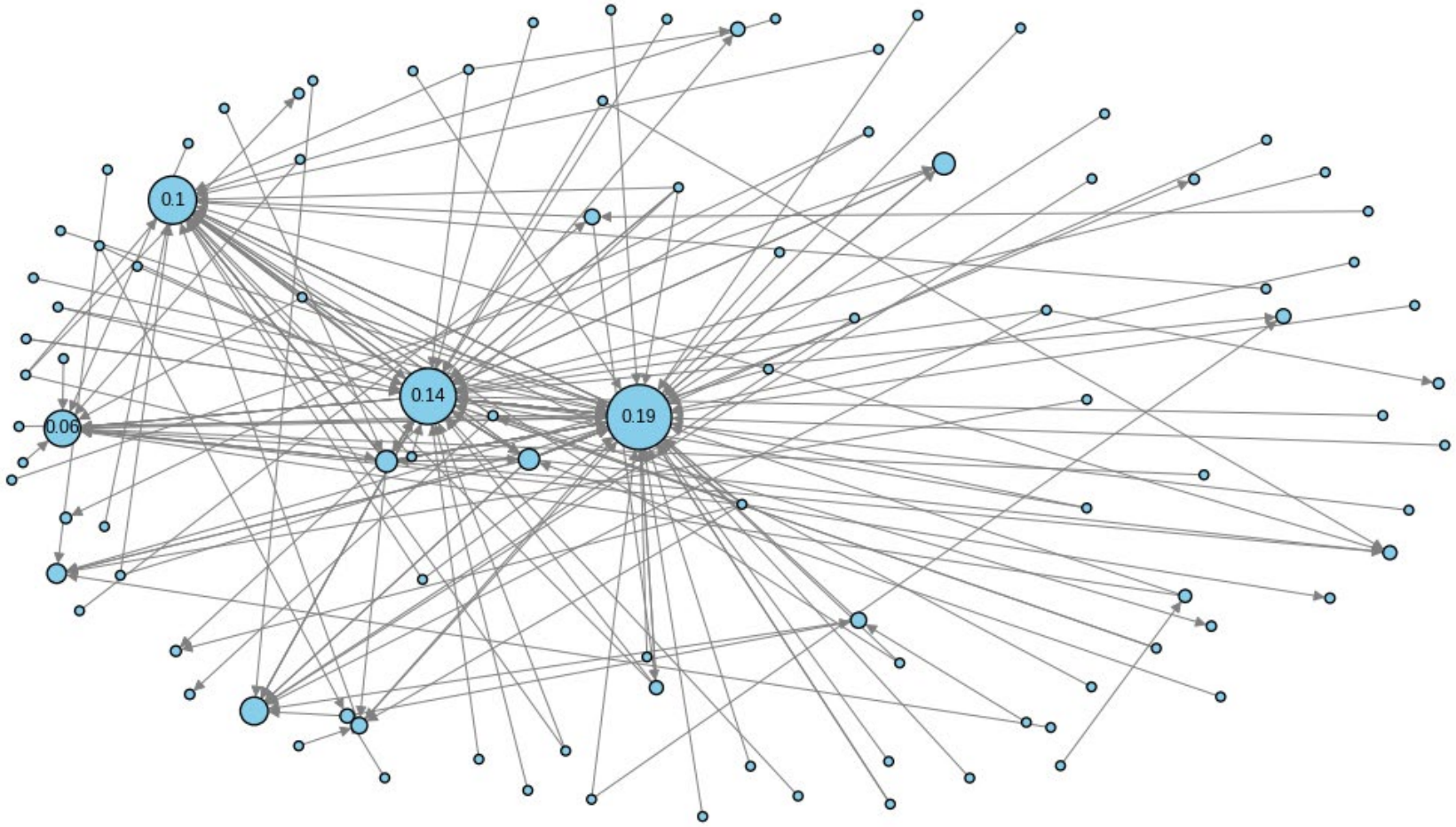


PageRank (1/4)

universität koblenz
weiter:denken

- **PageRank** is a centrality measure designed for directed networks, such as the Web graph
- It estimates the importance or prestige of nodes (e.g., Web pages), based on their link structure
- Especially useful when multiple pages are equally relevant by content — helps search engines rank them more effectively
- Introduced in 1998 by Sergey Brin and Larry Page as a core innovation behind Google's search engine
  - Pages gain importance by being linked to by other important pages
  - Captures the idea of recursive prestige
    - a page is important if it is linked to by other important pages

PageRank (2/4)




```
PR_dict = nx.pagerank(D)      # D is a DiGraph
```

PageRank (3/4)

- **Random surfer model:** browse the Web at random
  - A random link is clicked from each page to get to the next
- Random walk model modified with random jumps (**teleportation**)
  - At each step, with probability α, stop browsing and start new session form a random page
- Recursive definition
- PageRank is conserved, neither created nor destroyed: $\sum_i R(i) = 1$

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

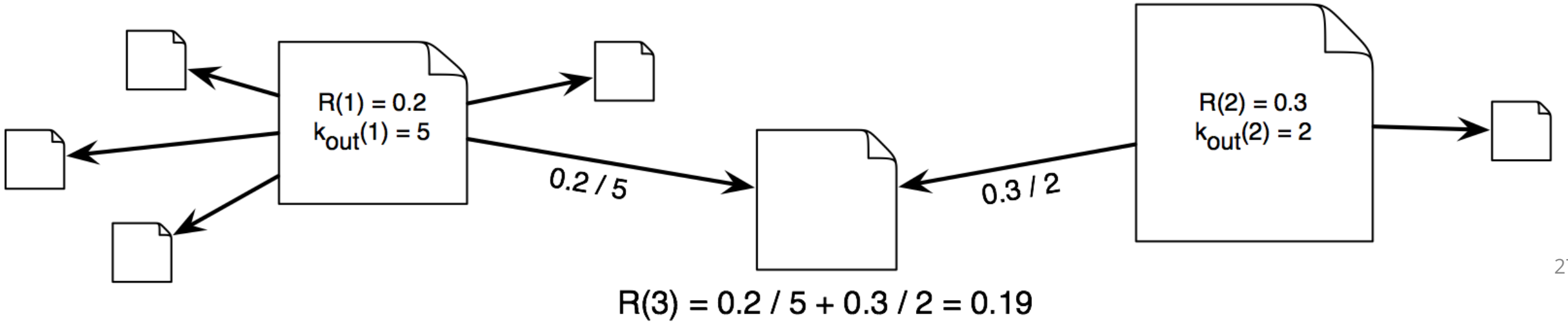
probability to land on *i* by random surfing

PageRank (4/4)

- α is the **teleportation** or **jump factor**, typically 0.15
  - PageRank converges quickly (in few iterations) if α > 0
- 1– α is the **damping factor**

Example with α = 0:

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



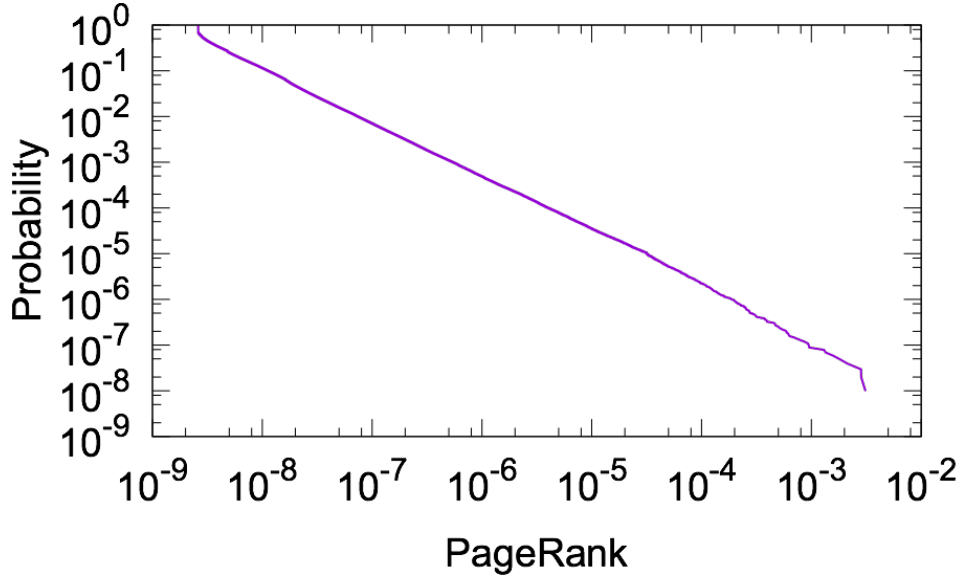
PageRank vs In-Degree (1/2)

universität koblenz
weiter:denken

- The distribution of PageRank across Web pages is highly skewed
  - A small number of pages accumulate very high PageRank, while most pages have very low values
- This distribution is similar to the in-degree distribution (number of incoming links), which also follows a power-law pattern
- To a first approximation, PageRank behaves like in-degree
  - Especially when all incoming links come from pages with equal PageRank
- However, the key distinction:
  - Links from more important pages carry more weight
    - A link from a high-PageRank page boosts your rank more than a link from a low-PageRank page
    - This reflects the idea of recursive prestige and quality over quantity

PageRank vs In-Degree (2/2)

- Search Engine Optimization (SEO) companies aim to improve a website's visibility and ranking in search engine results
  - One strategy is to boost the site's PageRank by increasing the number and quality of inbound links
- Some SEOs engage in unethical practices, such as:
  - Spamdexing – manipulating the link structure or page content to artificially inflate relevance or importance
  - Examples: link farms, keyword stuffing, hidden text/links
- Risks:
  - Search engines actively detect and penalize such behavior
  - If caught, a site may be heavily downranked or entirely removed (de-listed) from search results



universität koblenz
weiter:denken

3. Weighted Networks

Weighted Networks (1/2)

- Many real-world networks (directed and undirected) have link weights
  - Twitter diffusion: number of retweets between two users
  - Email: number of messages
  - Internet: data (bits) exchanged between routers
  - Air transportation: number of passengers
  - Brain: firing rate between neurons
  - Food webs: biomass

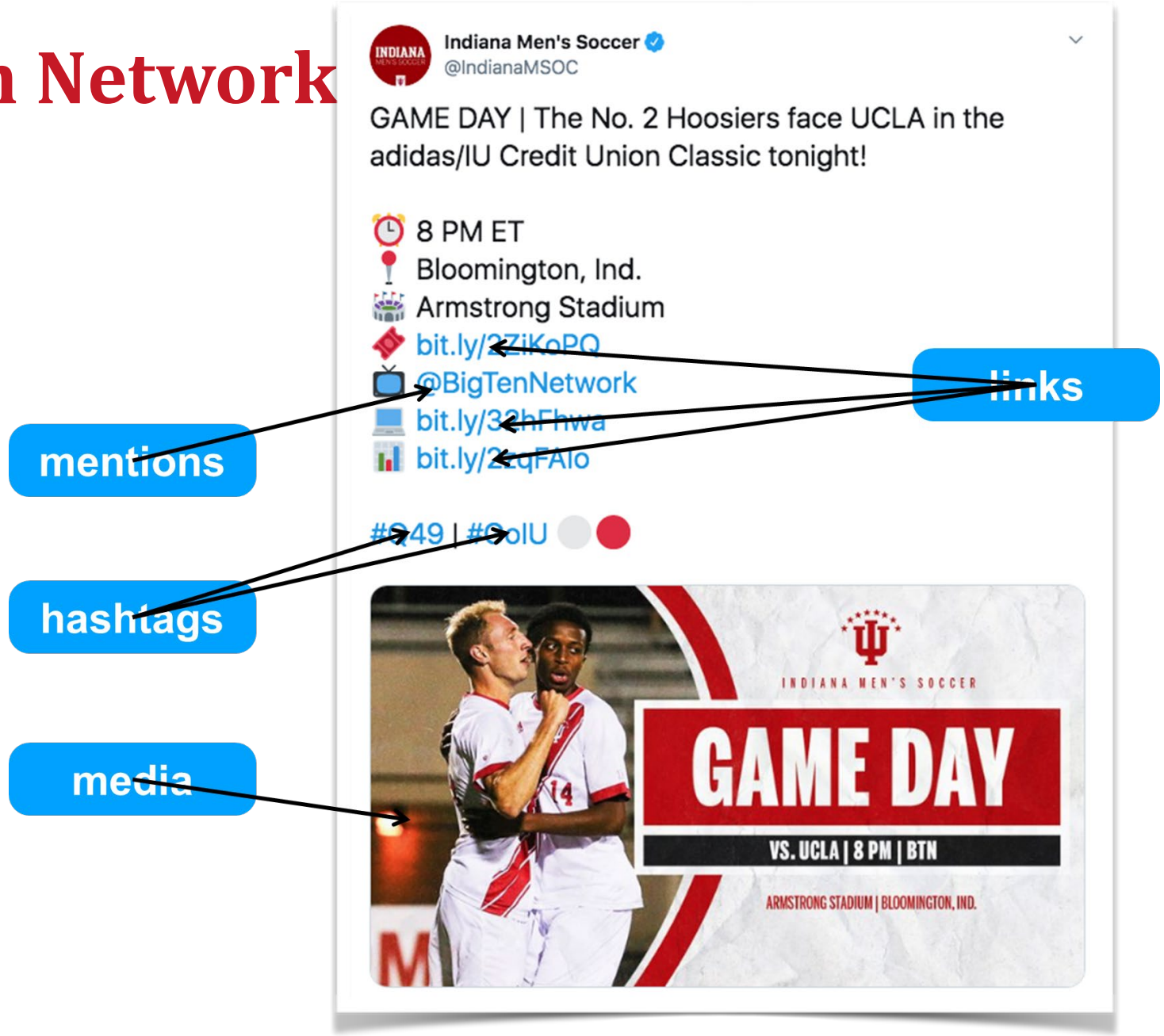
Weighted Networks (2/2)

- Even networks typically modeled as unweighted can be enriched with weights to reflect interaction intensity or relevance:
  - Facebook friendships
    - Weighted by number of interactions (e.g., likes, comments, tags) or mutual friends
  - Movie co-star network
    - Weighted by the number of films two actors have appeared in together
  - Wikipedia link network
    - Links weighted by click frequency or user navigation patterns
- In weighted networks, traditional degree measures extend naturally:
  - Degree → becomes strength (sum of weights on all links)
  - In-degree → in-strength
  - Out-degree → out-strength

4. Information and Misinformation Spread

Information Diffusion Network

- **Memes**: transmissible units of information, such as ideas, behaviors, news links, hashtags, and, yes, also images with captions (**image macros**)
- The definition of meme is due to Richard Dawkins, in analogy to genes transmitted from parent to offspring
- Like genes, memes can mutate and have fitness
- A tweet can carry several memes



Information Diffusion Networks (2/2)

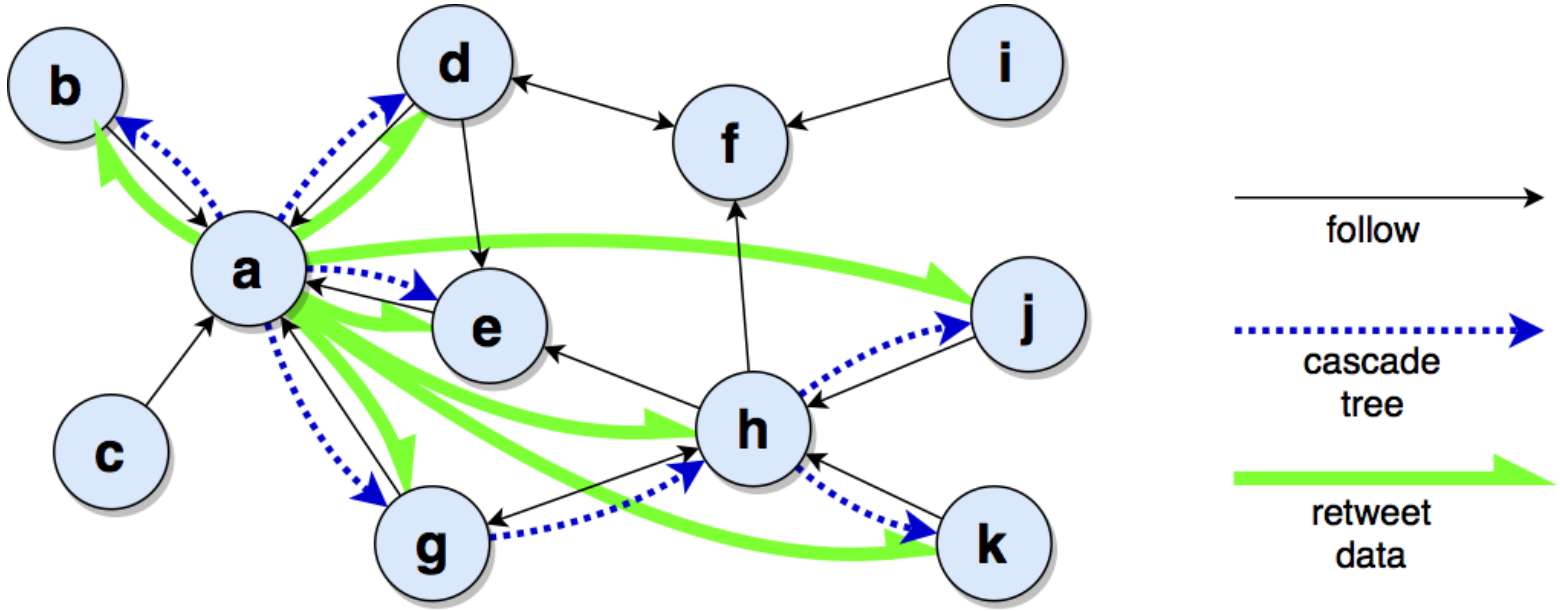
- We can track, map, and analyze the spread of memes on **Twitter**
  - **Retweet network**: link from retweeted user to retweeter user
  - **Mention/reply network**: link to user who replies or who is mentioned
- Tweets are time-stamped; we can aggregate the temporal networks
- Can focus on a particular meme (eg, a hashtag) or multiple ones (eg, a set of accounts or links to a news source)

Play with the interactive diffusion network tools at osome.iuni.iu.edu

Retweet Networks

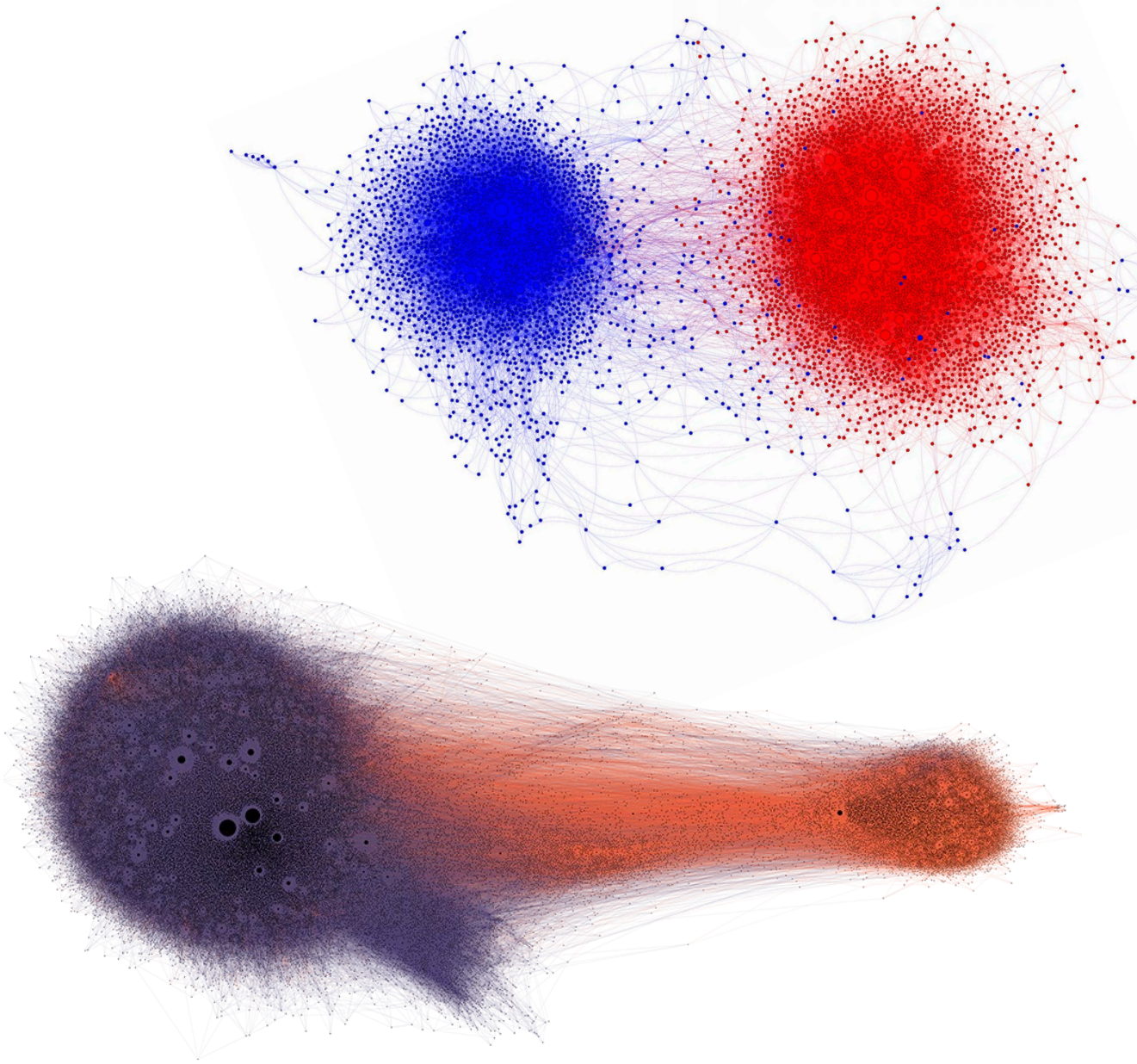
universität koblenz
weiter:denken

- In the data, each retweet cascade network is a star (all retweets point to original tweet)
- The actual cascade tree is difficult to reconstruct, but we can make some guesses based on the follower network and timestamps



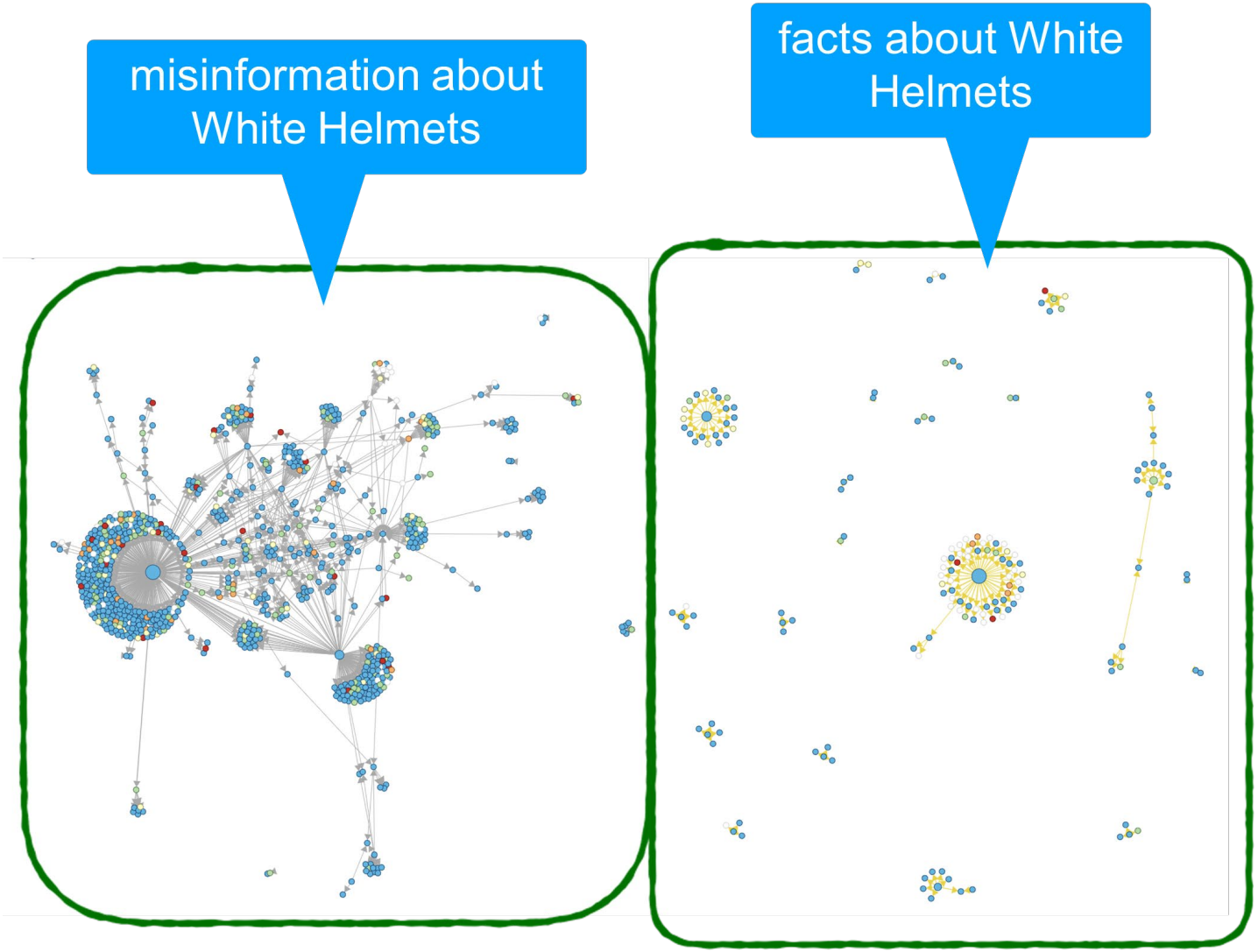
Echo Chambers

- Examples
  - Retweets of tweets with progressive (blue) and conservative (red) political hashtags during 2010 US election (k=3 core)
  - Retweets of tweets with links to low-credibility (purple) and fact-checking (orange) sources during 2016 US election (k=5 core)



Virality

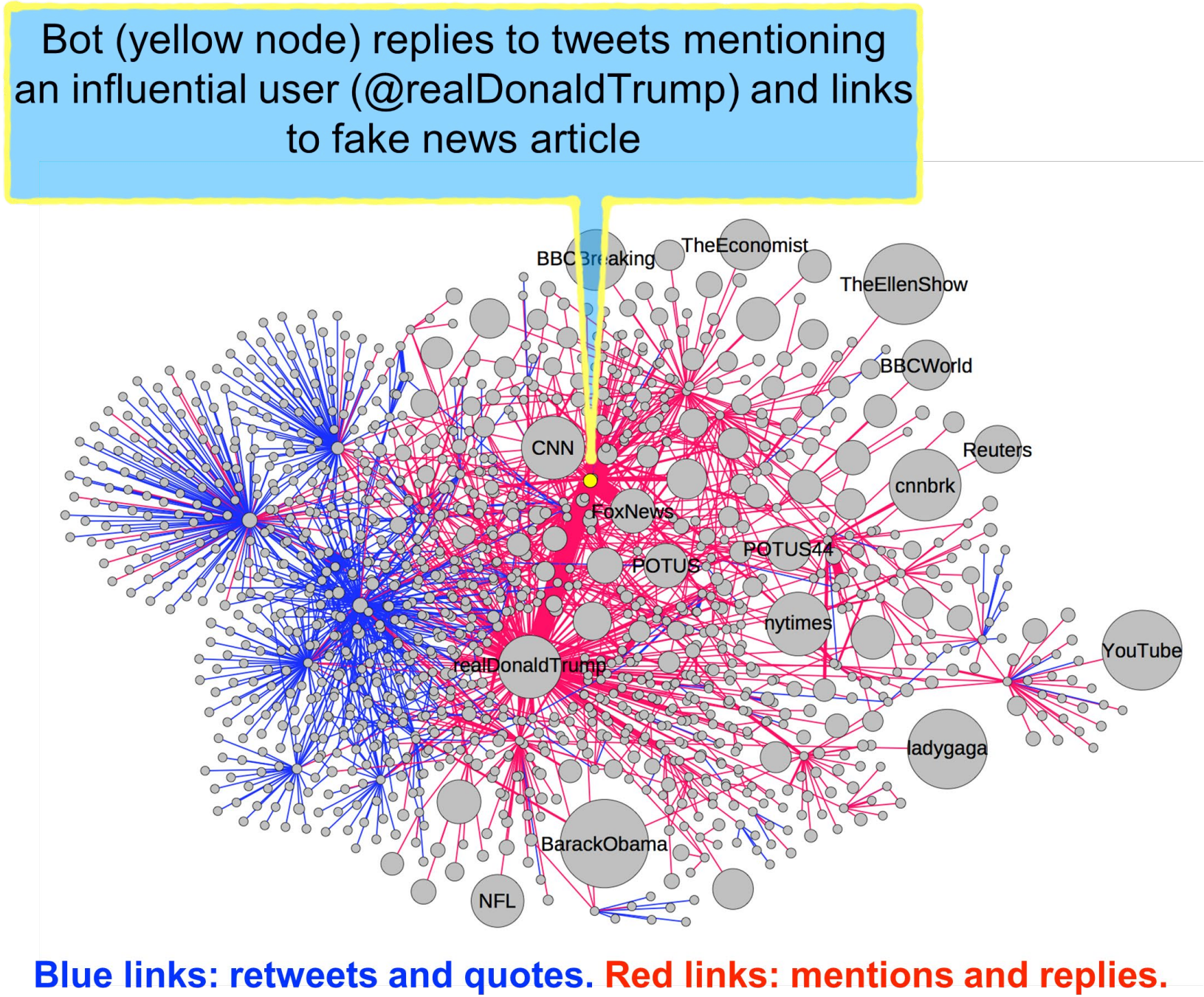
- Multiple ways to measure the virality of a meme:
  - Number of users exposed
  - Depth of diffusion tree
  - Fraction of users who retweet to users who are exposed
- Misinformation is often more viral than actual news reports



Source: hoaxy.iuni.iu.edu

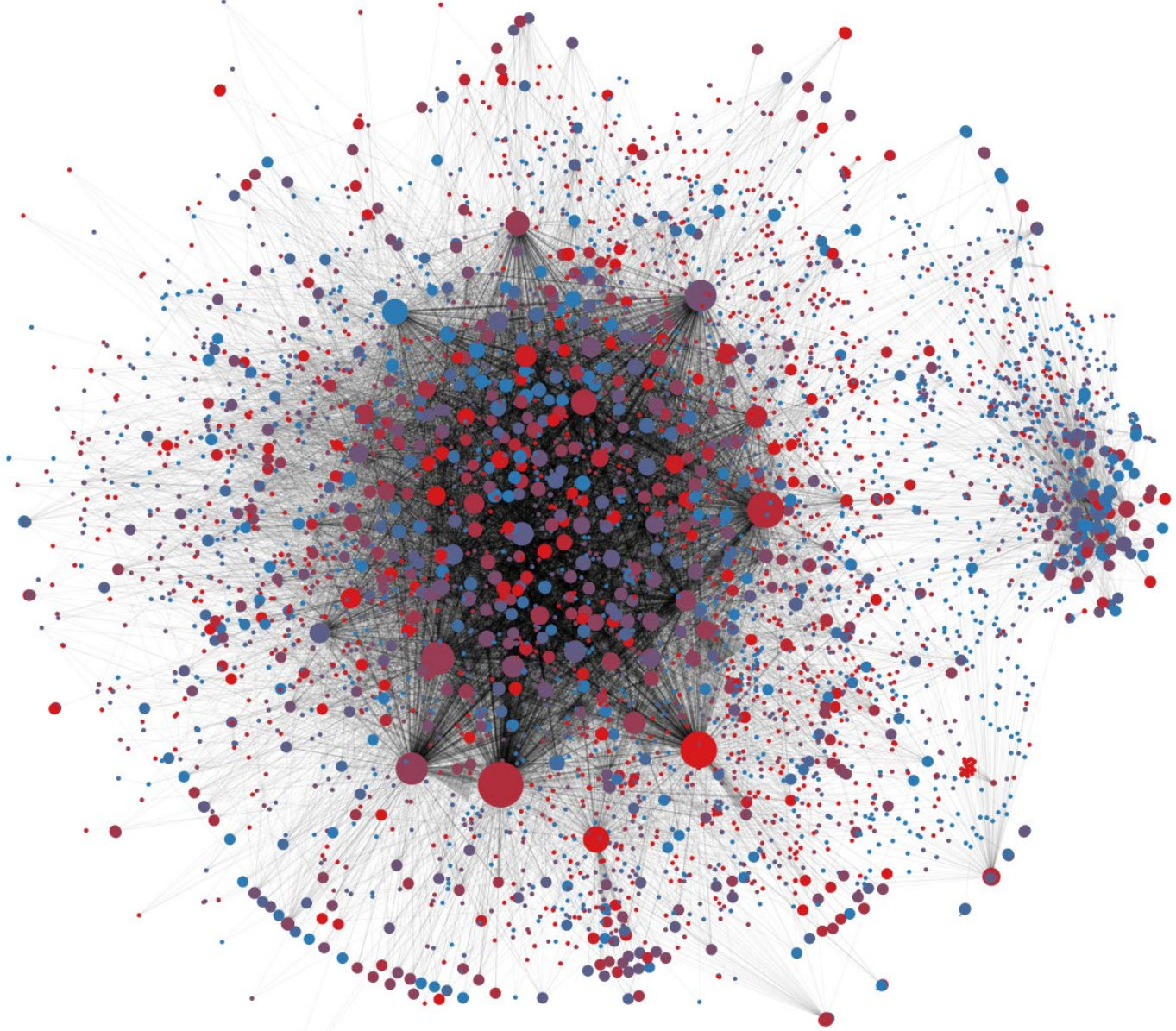
Influence

- Multiple ways to measure the influence of an account
  - 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
- Social bots can target influential accounts hoping for retweet



Social Bots

- Accounts controlled by an entity via software
- Malicious social bots can impersonate humans, deceive, and manipulate diffusion networks
  - Fake followers
  - **Amplification**: fake retweets
- All social media platforms and users are vulnerable



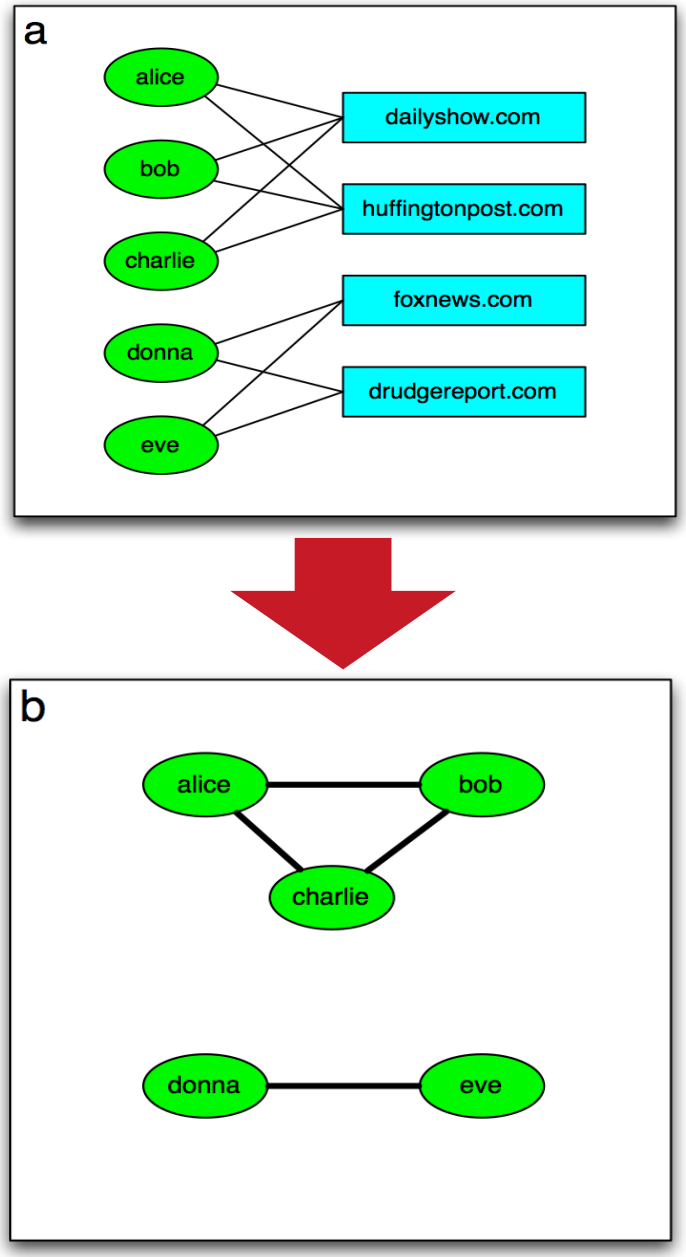
Large red nodes: influential bots manipulating online debate about vaccination policy

universität koblenz
weiter:denken

5. Co-Occurrence Networks

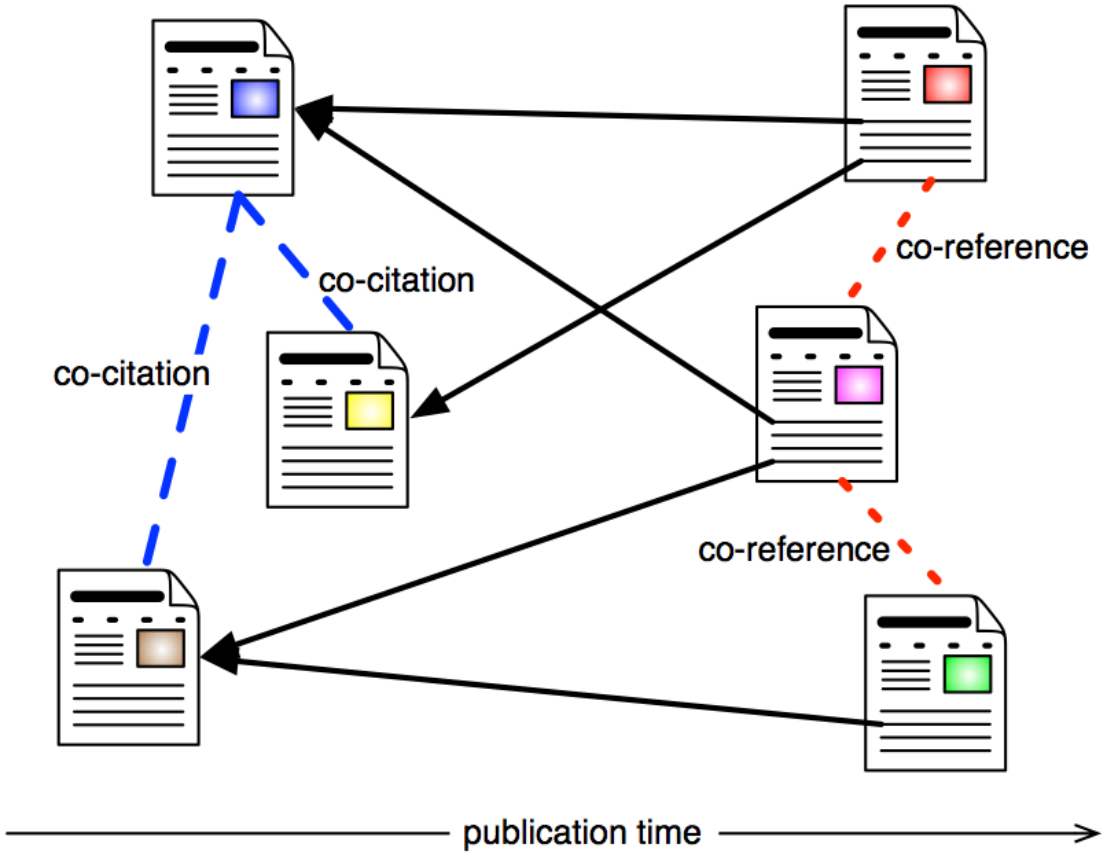
Co-occurrence Networks

- Weight can originate from relationships between more than one type of entity
- Nodes in a bipartite network **co-occur** if they share common neighbors
- **Projection**: keep nodes from either group of bipartite network and add edges among co-occurring nodes
- Examples
  - From star-movie network to co-star network
  - From 'like' data to friend recommendations
  - From music listening data to similar-song recommendations
  - From purchase data to product advertising ("people who bought this also bought…")



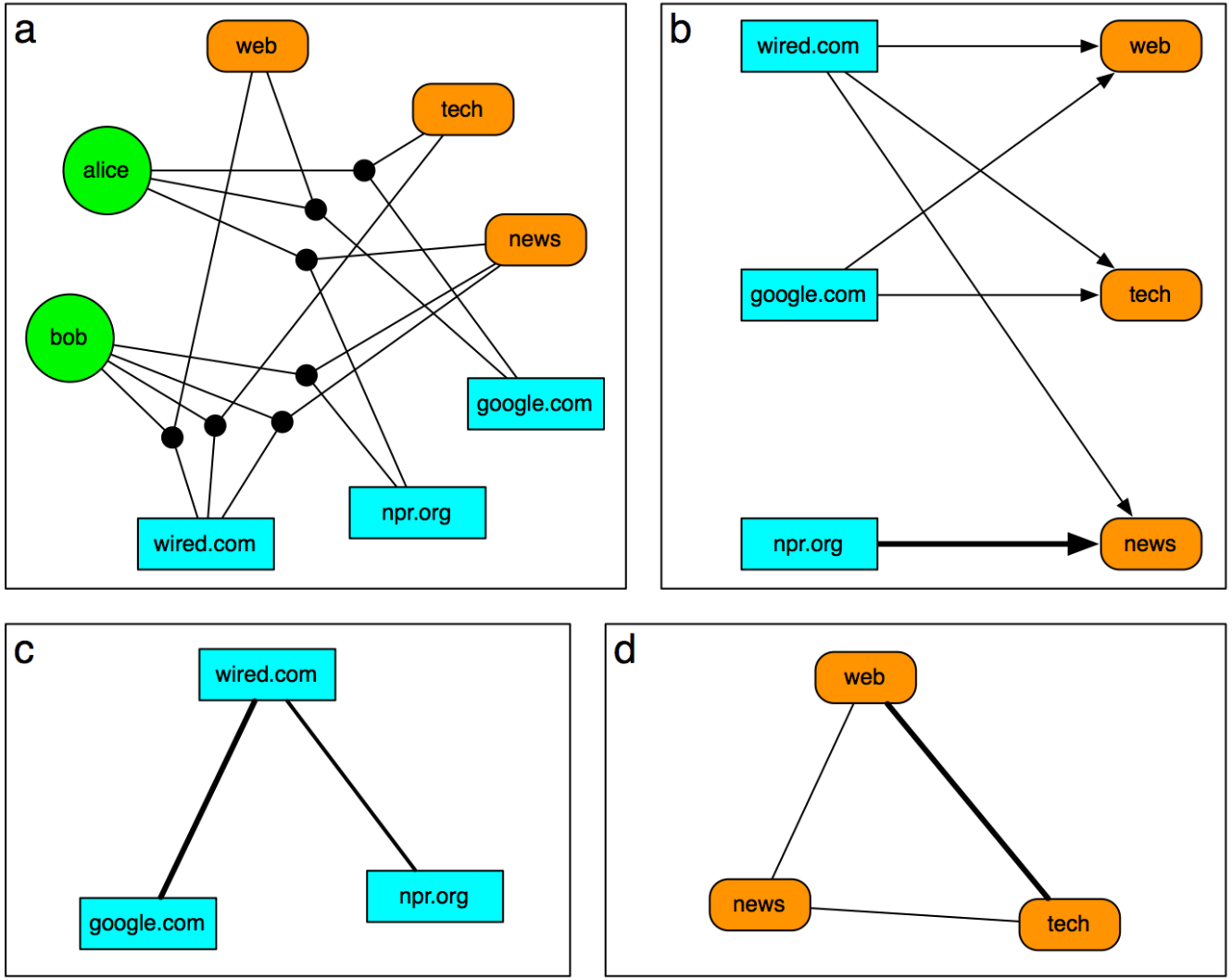
Co-citation and Co-reference

- Think of any directed network as a bipartite network with sources and targets as two groups of nodes
  - nodes with both incoming and outgoing links can appear in both groups
- Count the number of shared successors or shared predecessors and use it as a weight between nodes in the same group
  - **Co-citation network**: shared predecessors (cited by same papers)
  - **Co-reference network**: shared successors (citing same papers)



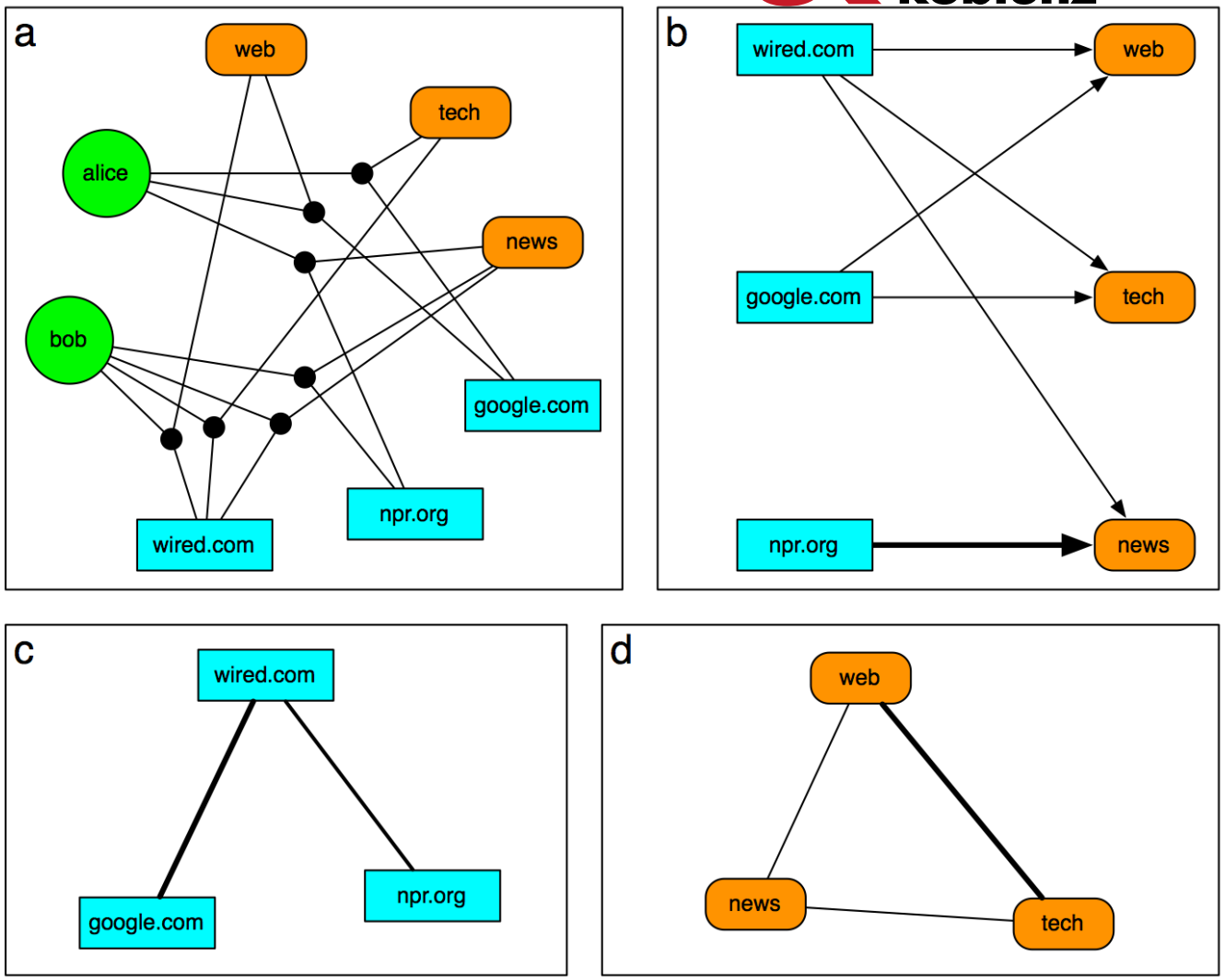
Social Tagging (1/2)

- **Folksonomy**: set of triples (*user, resource, tag*) where *user* assigns *tag* to *resource*
  - Resource can be any URL, e.g., photos on Flickr
  - Implicit in social media, e.g. #hashtags on Twitter
  - **Tripartite network**: each link connects three nodes from distinct sets



Social Tagging (2/2)

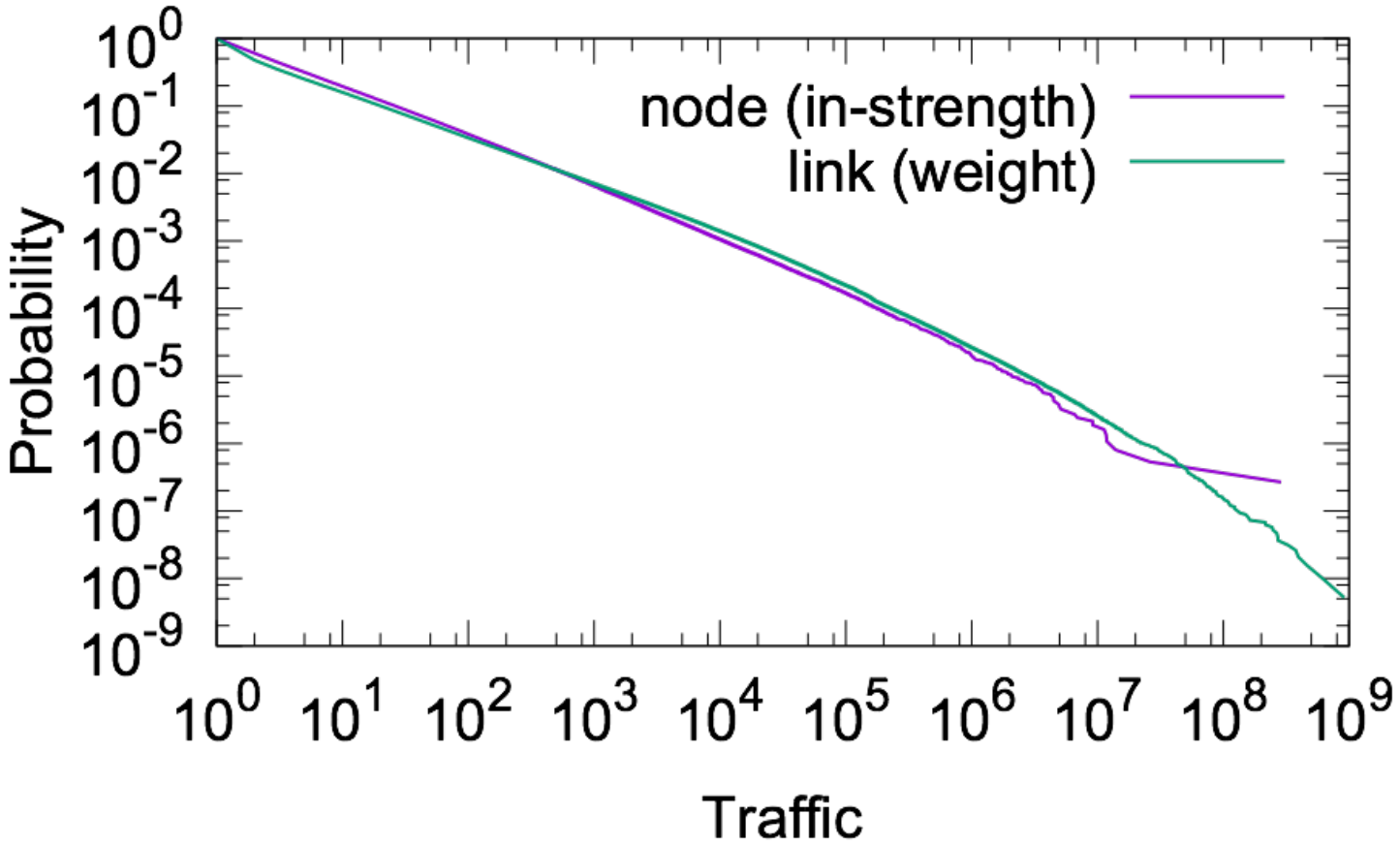
- Can project based on common neighbors
  - First stage: project to weighted bipartite network
  - Second stage: project to weighted network; cosine similarity is one way to compute weights
- Check out osome.iuni.iu.edu/tools/networks/ for interactive Twitter hashtag co-occurrence networks



6. Weight Heterogeneity

Weight Heterogeneity

- Weights can span many orders of magnitude
- Example: Web traffic
  - Click data collected at IU
  - ~100k anonymous users
  - ~1B clicks
  - $N \approx 4$M sites
  - $L \approx 11$M weighted directed links
- Both in-strength (s_{in}) and weight (w) distributions have very heavy tails

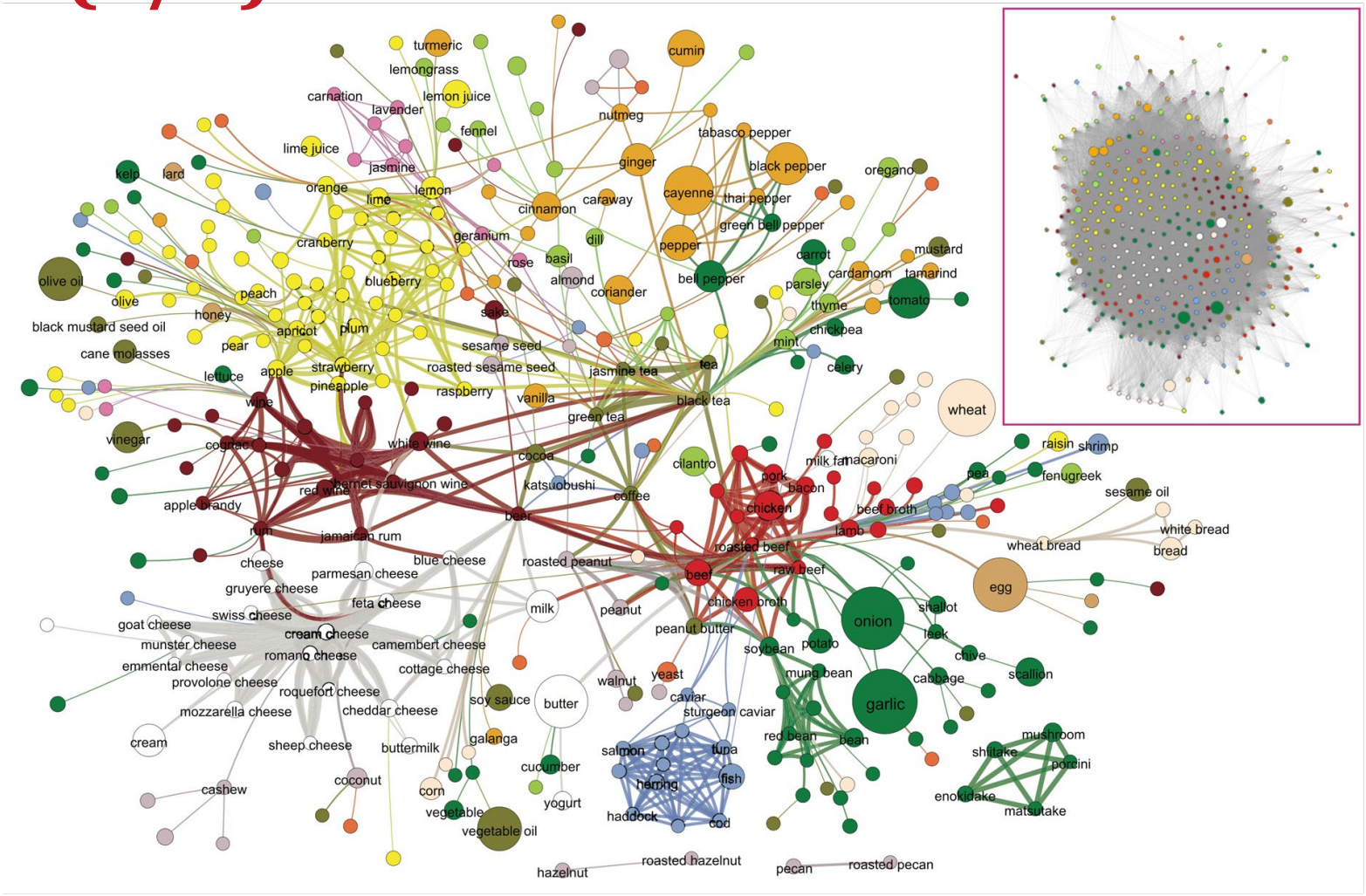


Link Filtering

- Weighted networks often contain many low-weight links, making them too dense to visualize or analyze effectively
- If these weak links represent low-importance connections or noise, we can filter them out to create a sparser, more meaningful network
- Example:
  - In a text similarity network, documents might have small but non-zero similarity scores due to common words like "the" or "and", which don't reflect actual content similarity
- Common Solution:
  - Apply a weight threshold:
    - Remove all links with weights below a certain value
    - This reduces noise and simplifies analysis
    - Often sufficient for practical purposes

Network Backbone (1/2)

- How to set the threshold to filter links?
- In some cases, due to weight heterogeneity, a lot of information is lost by setting a threshold such that enough links are removed



Backbone flavor network with α = 0.04

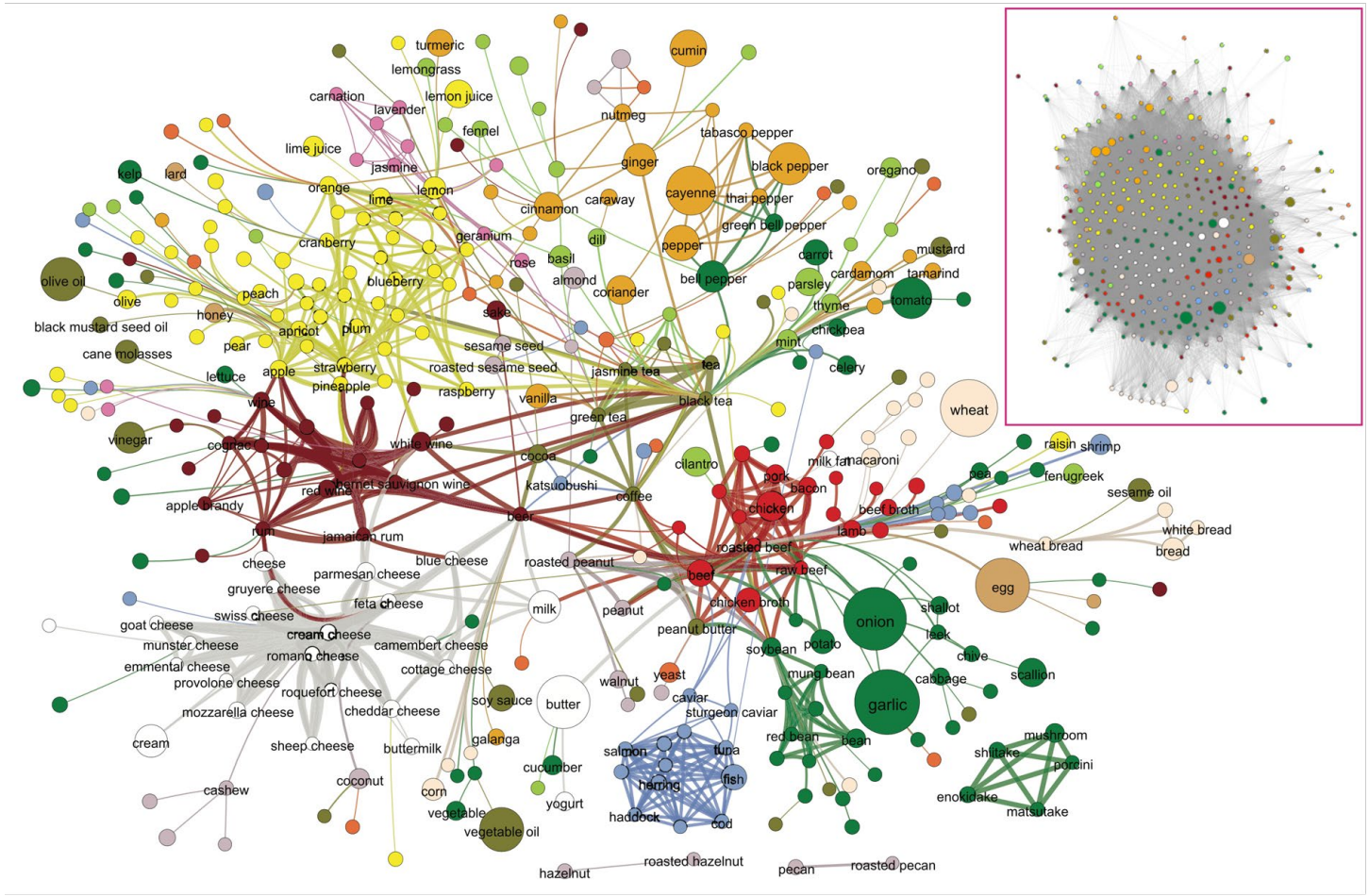
Image DOI:10.1038/srep00196 CC BY 4.0

Network Backbone (2/2)

- **Backbone**: set a different threshold for each node, so that we remove links with weights that are not significant compared to the other links of that node
- Compare probability of link weight, given degree and strength of node, to a **significance threshold**: keep links if

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

Dense flavor network



Backbone flavor network with α = 0.04

7. Summary

Summary (1/3)

universität koblenz
weiter:denken

1. Web Structure & Graphs

- The Web forms a vast **network of pages connected by hyperlinks**
- Web graphs analyzed using **web crawlers**
- Key properties:
  - **Heavy-tailed in-degree distribution**
  - **Short paths** via popular hub pages
  - Presence of a **large strongly connected component**

2. Document Similarity & Representation

- Pages as **high-dimensional word vectors**
- **Cosine similarity** measures content closeness
- Linked pages often **semantically related** (topical locality)

Summary (2/3)

universität koblenz
weiter:denken

3. PageRank

- Measures **importance** via a **random walk model**
- Links from **important pages** boost rank more
- Core of **early Google search algorithm**

4. Information Diffusion in Social Networks

- Content spreads via **retweets, shares, likes**
- **Cascade networks** trace spread of ideas and misinformation
- **In-/out-strength** used to find **key influencers**
- Beware of **social bots** manipulating visibility

Summary (3/3)

5. Weighted Networks

- Derived from **bipartite data** (e.g., co-purchases, co-citations)
- Weights reflect **interaction strength**
- Can reveal **underlying structures**, but may be too dense

6. Pruning Dense Networks

- **Remove low-weight edges** to reduce noise
- Use **thresholding** to extract **core structures**
- Reveals **backbone** of meaningful connections

References

[1] Menczer, F., Fortunato, S., & Davis, C. A. (2020). A First Course in Network Science Cambridge: Cambridge University Press.

- Chapter 4 Directions and Weights

[2] OLAT course page: https://olat.vcrp.de/url/RepositoryEntry/4669112833