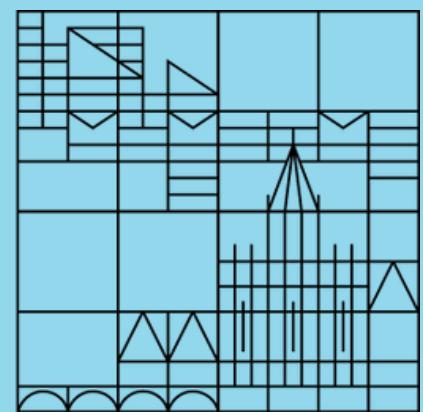


UNIVERSITÄT KONSTANZ

Measuring Nodes Centrality

Network Science of
Socio-Economic Systems
Giordano De Marzo

Universität
Konstanz



Recap

Power Law Probability Distributions

Many real life phenomena are characterized by extreme events described by power law probability distributions

Scale-Free Networks

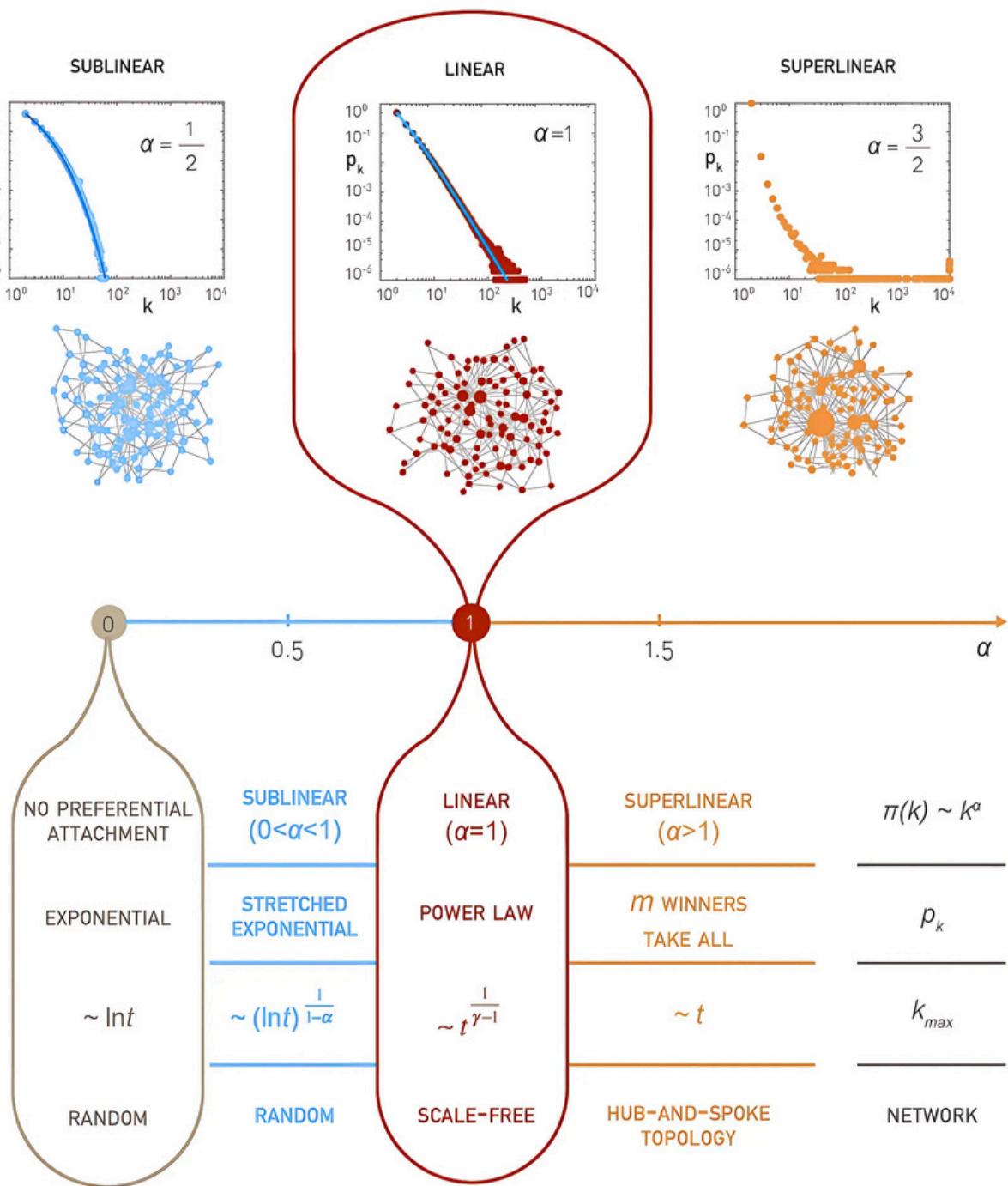
Real world networks tend to be scale-free, i.e. their degree distribution is a power law

Barabasi-Albert Model

Scale-free networks can be generated using linear preferential attachment

Robustness of Scale-Free Networks

Scale-free networks are more tolerant to failures, but more susceptible to attacks



Outline

- 1.The Quest for Online Search Engines
- 2.The PageRank
- 3.Centrality Measures
- 4.Analyzing Criminal Networks





The Quest for Online Search Engines

The Birth of the World Wide Web



- **1989** Tim Berners-Lee at CERN, proposes a system for sharing information among researchers.
- **1990** Development of the first web browser and web server. First website hosted at CERN: <http://info.cern.ch>
- **1991** The World Wide Web is made public. Open access to the first web server marked the beginning of the WWW as a global system.

1993 - The www Virtual Library

With the creation of the WWW a problem also emerged

- the rapid growth of websites made it difficult to find relevant content
- there were no search engines or tools to retrieve information

The WWW Virtual Library was the first attempt in solving this problem

- created in 1993 by Tim Berners-Lee
- manually curated by volunteers.
- organized links by topics

The WWW Virtual Library

- [Agriculture](#)
[Agriculture](#), [Beer & Brewing](#), [Gardening](#)...
- [Computer Science](#)
[Computing](#), [Graphics](#), [Languages](#), [Web](#)...
- [Communications and Media](#)
[Communications](#), [Telecommunications](#), [Journalism](#)...
- [Education](#)
[Education](#), [Cognitive Science](#), [Libraries](#), [Linguistics](#)...
- [Engineering](#)
[Civil](#), [Chemical](#), [Electrical](#), [Mechanical](#), [Software](#)...
- [Humanities](#)
[Anthropology](#), [Art](#), [Dance](#), [History](#), [Museums](#), [Philosophy](#)...
- [Information Management](#)
[Information Sciences](#), [Knowledge Management](#)...
- [International Affairs](#)
[International Security](#), [Sustainable Development](#), [UN](#)...
- [Law](#)
[Law](#), [Environmental Law](#)...
- [Business and Economics](#)
[Economics](#), [Finance](#), [Transportation](#)...
- [Recreation](#)
[Recreation](#), [Games](#), [Gardening](#), [Sport](#)...
- [Regional Studies](#)
[Asian](#), [Latin American](#), [West European](#)...
- [Science](#)
[Biosciences](#), [Medicine & Health](#), [Physics](#), [Chemistry](#)...
- [Society](#)
[Political Science](#), [Religion](#), [Social Sciences](#)...

Mirrors: [Stanford](#) (USA), [Penn State](#) (USA), [East Anglia](#) (UK) [Geneva](#) (CH), [Geneva-2](#) (CH), [Argentina](#).

[About the VL](#) | [Alphabetical listing](#) | [VL keyword search](#) | [What's New](#)

Last update Nov 23 1998

1994 - Yahoo!

Yahoo - A Guide to WWW

[[What's New?](#) | [What's Cool?](#) | [What's Popular?](#) | [Stats](#) | [A Random Link](#)]

[Top](#) | [Up](#) | [Search](#) | [Mail](#) | [Add](#) | [Help](#)

- [Art\(466\) NEW](#)
- [Business\(6426\) NEW](#)
- [Computers\(2609\) NEW](#)
- [Economy\(743\) NEW](#)
- [Education\(1487\) NEW](#)
- [Entertainment\(6199\) NEW](#)
- [Environment and Nature\(193\) NEW](#)
- [Events\(53\) NEW](#)
- [Government\(1031\) NEW](#)
- [Health\(367\) NEW](#)
- [Humanities\(163\) NEW](#)
- [Law\(163\) NEW](#)
- [News\(185\)](#)
- [Politics\(148\) NEW](#)
- [Reference\(474\) NEW](#)
- [Regional Information\(2606\) NEW](#)
- [Science\(2634\) NEW](#)
- [Social Science\(93\) NEW](#)
- [Society and Culture\(648\) NEW](#)

23836 entries in *Yahoo* [[Yahoo](#) | [Up](#) | [Search](#) | [Mail](#) | [Add](#) | [Help](#)]

yahoo@akebono.stanford.edu
Copyright © 1994 David Filo and Jerry Yang

The WWW Virtual Library was not very user friendly, Yahoo! brought the WWW to the people

- Founded by Jerry Yang and David Filo as a manually curated web directory.
- Organized websites into hierarchical categories
- Focused on quality over quantity, offering a user-friendly experience

Yahoo! became the go-to web directory of the mid-1990s and inspired the shift toward more structured approaches to web navigation.

Explosive Growth of the WWW

If in '94 only a few thousands web pages existed, by 2 years this number was 100 times larger, making manually curated directories hard to scale



"Website" is defined as a unique hostname, i.e. a name which can be resolved, using a name server, into an IP Address.



@StatistaCharts

* As of August 1, 1991

** As of August 19, 2019 at 10:22 CET

Source: Internet Live Stats

statista

1995 - AltaVista

The exponential growth of web pages made manual curation impractical

- AltaVista was the first approach to solve this limit
- Launched in 1995 as one of the first search engines based on keyword search
- Introduced automated indexing and a crawler to scan and catalog the web.
- First search engine to catalog a large fraction of the WWW

The screenshot shows the AltaVista homepage. At the top, the AltaVista logo is displayed with the tagline "View Multimedia From Our Vantage Point". Below the logo is a banner for "AUTO.BY.TEL USA CANADA" advertising "Buy and insure new cars & trucks online". To the right of the banner, there's a section for "Car Buying & Car Insurance Pain Relief" featuring small illustrations of cars. A red banner across the middle of the page encourages users to "Click here for advertising information - reach millions every month!". Below this is a search bar with dropdown menus for "Search the Web" and "Display the Results in Standard Form". A "Submit" button is located to the right of the search bar. Further down, there are links for "Search with Digital's Alta Vista" and options for "Advanced Search" and "Add URL". Two promotional boxes are visible: one for "Contests" with the text "Make Me Laugh..." and another for "Creative Web" with the text "Create a Site...". At the bottom, there's a link to "Download free demo versions of AltaVista Technology software". The footer contains the AltaVista logo again, along with links for "[Creative]", "[Search]", "[Humor]", and "[Email]".

Limits of Early Search Engines

Early search engines had many issues making navigating the WWW very hard

Scalability Issues with Human Curation

- Manual directories (e.g., Yahoo!) couldn't keep pace with the rapid growth of the web.

Lack of Context Understanding in Keyword Search

- Early search engines matched keywords but failed to grasp context.
- Results often lacked relevance.

Difficulty in Quantifying Page Quality

- No effective metrics to assess the credibility or usefulness of web pages.
- All indexed pages were treated equally.

Vulnerability to Keyword Spamming

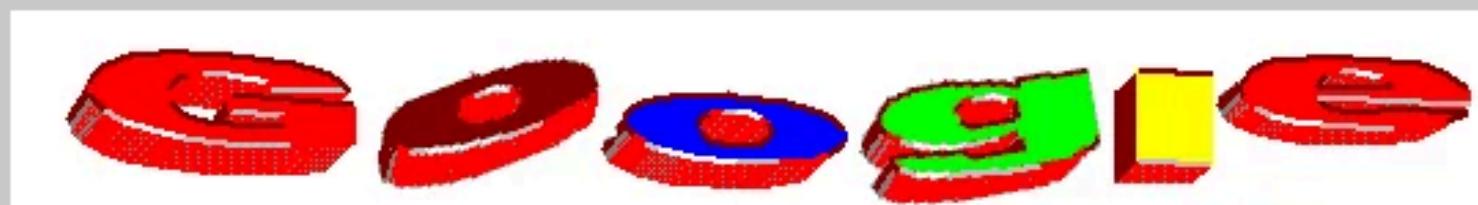
- Webmasters manipulated rankings by overloading pages with keywords, degrading search quality.

1998 - Google

Google Search Engine

This is a demo of the Google Search Engine. Note, it is research in progress so expect some downtimes and malfunctions. You can find the older [Backrub web page here](#).

Google is being developed by [Larry Page](#) and [Sergey Brin](#) with very talented implementation help by [Scott Hassan](#) and [Alan Steremberg](#).



Search Stanford

10 results ▾ clustering on ▾

Search The Web

10 results ▾ clustering on ▾

A better approach to web search was needed and Google was the solution

- Founded in 1998 by Larry Page and Sergey Brin at Stanford University.
- Developed initially as a research project called BackRub

Google experienced a rapid Growth

- By 2000, Google became the leading search engine, handling millions of queries daily.

A Radical Change of Perspective

Google represent a radical change of perspective from Content to Structure

- Previous Search Engines:
 - Focused primarily on content, matching keywords on web pages.
 - Treated pages as isolated entities without considering relationships.
- Google's Breakthrough:
 - Introduced the concept of structure and role within a network.
 - Viewed the web as a network of interconnected pages, where links represent citations.

A page's importance is not solely about its content but also about:

- How many other pages link to it.
- The importance of the pages providing those links.

The PageRank



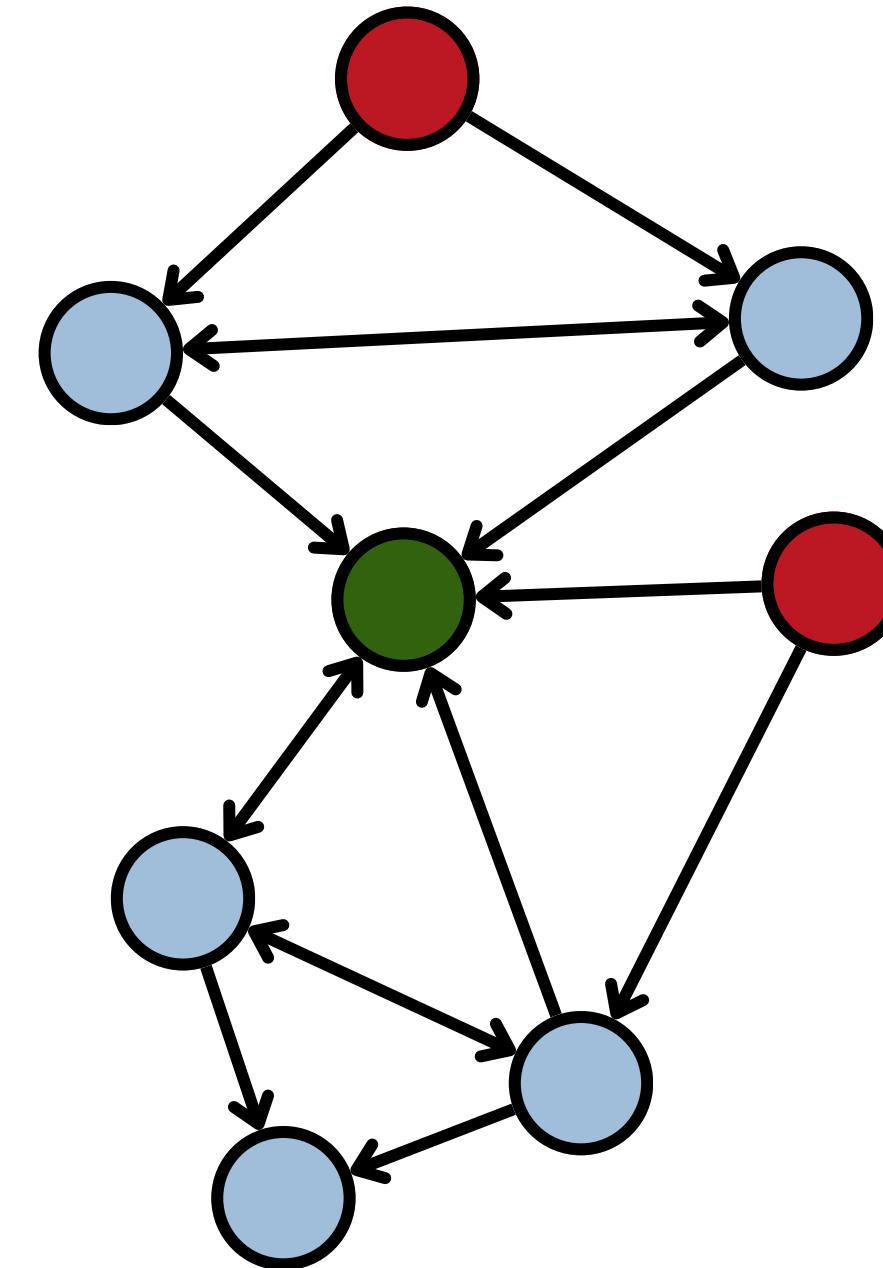
The PageRank

Google success is rooted in the PageRank algorithm, that is based on the following assumptions

- Incoming links can be seen as votes of confidence
- Not all votes are equally important
- Votes from pages with many incoming links themselves are more significant

This approach is very hard to manipulate

- Keywords can be manually added
- Incoming links are not directly influenceable



- **Low Reputation**
- **High Reputation**

Matrix Formulation

If we denote by $P(i)$ the PageRank of page i , this score will be

- the larger the more incoming links i has
- the larger the larger is the PageRank of the pages pointing to it
- the links coming from a page with many outgoing links should be valued less

Putting all these ingredients together we get the PageRank equation

$$P(i) = \sum_{j \rightarrow i} \frac{P(j)}{k_{out}(j)}$$

We can rewrite this expression in terms of the adjacency matrix A as

$$P(i) = \sum_j P(j) \frac{A_{ji}}{k_{out}(j)} = \sum_j P(j) M_{ji}$$

The matrix M is the adjacency matrix normalized by the out degree. In this way the sum of all the outgoing links from a node is always equal to one

Recursive Formulation

The PageRank of a page is influenced by the PageRank of the other pages, thus we don't have a closed equation, bu we can solve it recursively

- we set all initial PageRank values to 1

$$P^{(0)}(i) = 1$$

- we use the recursive relation

$$P^{(n+1)}(i) = \sum_j P^{(n)}(j) M_{ji}$$

Let's see a simple example with a network with constant out degree equal to 1

$$P^{(1)}(i) = \sum_j P^{(0)}(j) \frac{A_{ji}}{k_{out}(j)} = \sum_j 1 \frac{A_{ji}}{1} = k_{in}(i)$$

$$P^{(2)}(i) = \sum_j P^{(1)}(j) \frac{A_{ji}}{k_{out}(j)} = \sum_j k_{in}(j) A_{ji}$$

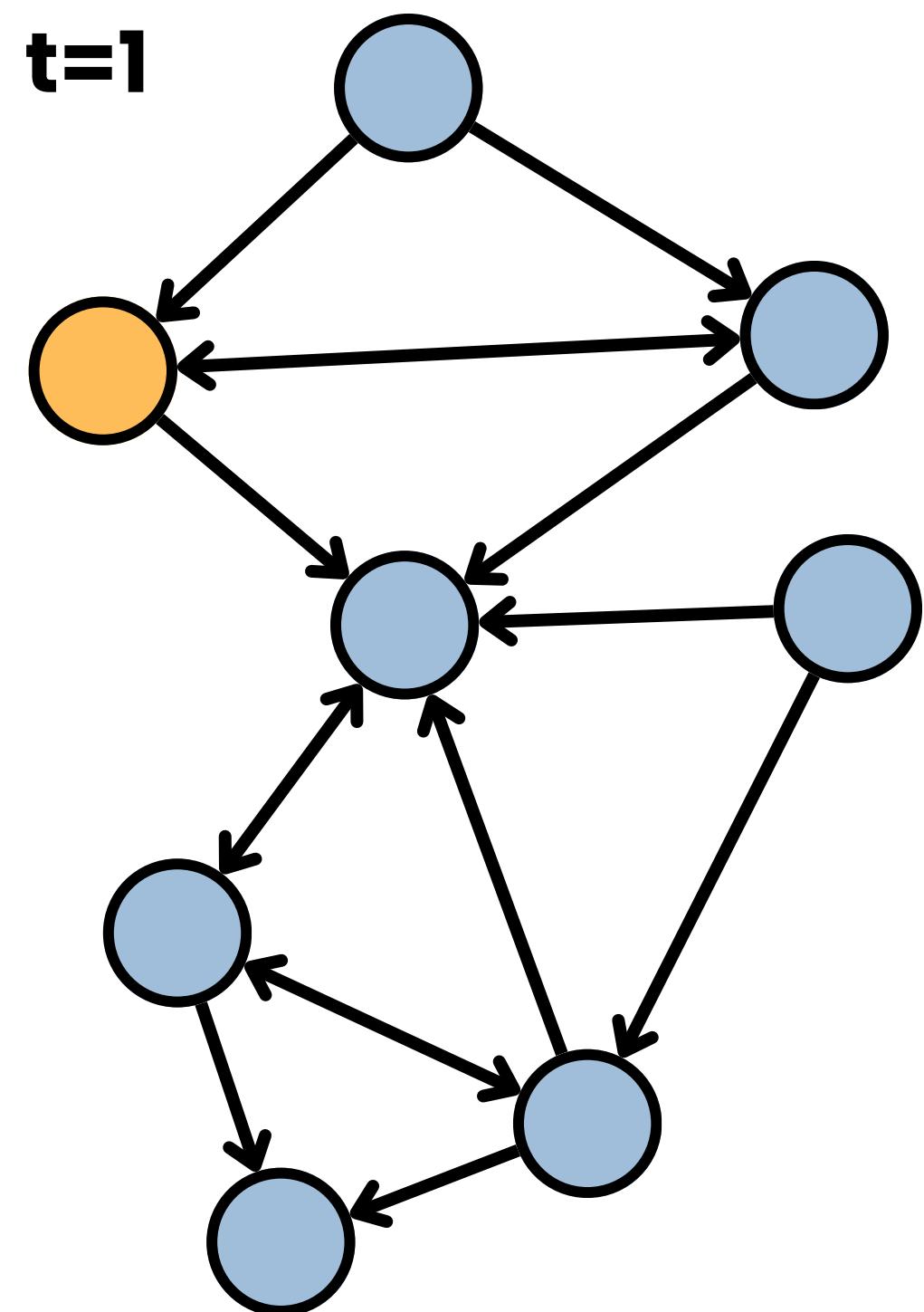
PageRank as a Random Walk

Let's look back at the PageRank equation

$$P(i) = \sum P(j) M_{ji}$$

We can give a different interpretation to this equation

- we consider a user randomly surfing the web, following out-links at random



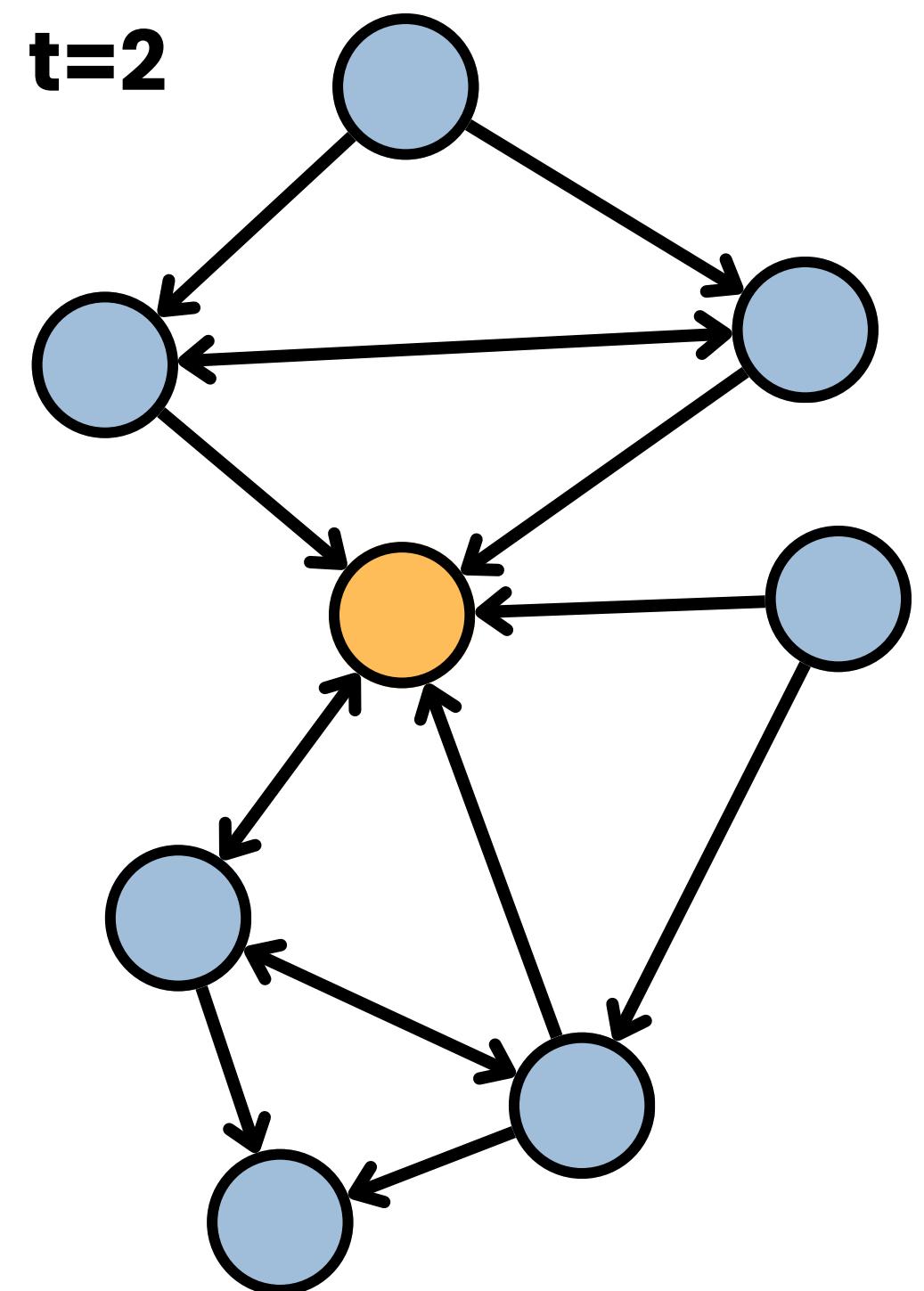
PageRank as a Random Walk

Let's look back at the PageRank equation

$$P(i) = \sum P(j) M_{ji}$$

We can give a different interpretation to this equation

- we consider a user randomly surfing the web, following out-links at random



PageRank as a Random Walk

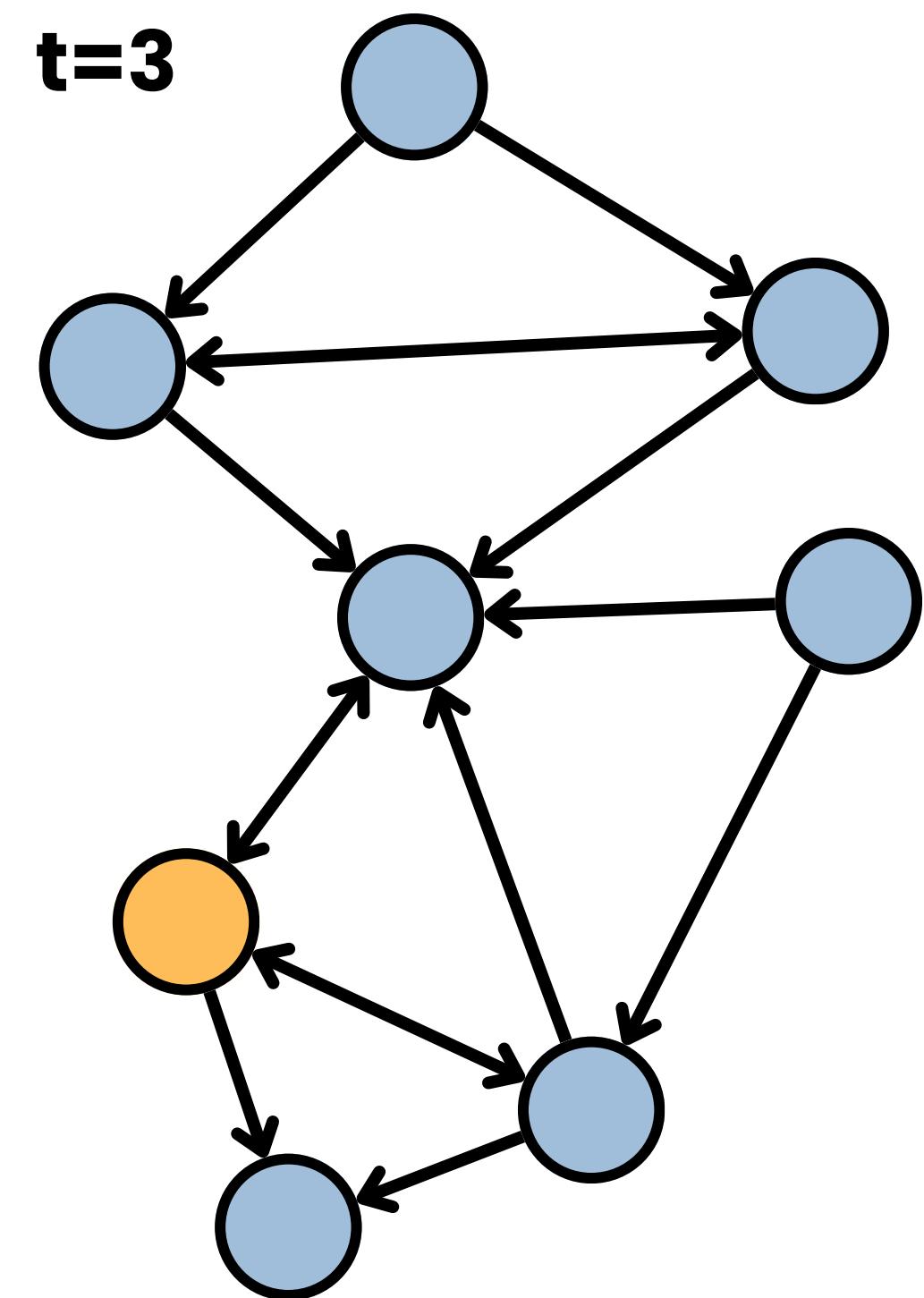
Let's look back at the PageRank equation

$$P(i) = \sum P(j) M_{ji}$$

We can give a different interpretation to this equation

- we consider a user randomly surfing the web, following out-links at random
- $P(i)$ gives the probability that the surfer is visiting page i after it has explored the WWW long enough

The equation says that the probability to be in page i is the probability to be in page j , times the probability to go from j to i , summed over all possible pages j



Dead Ends and Spider Traps

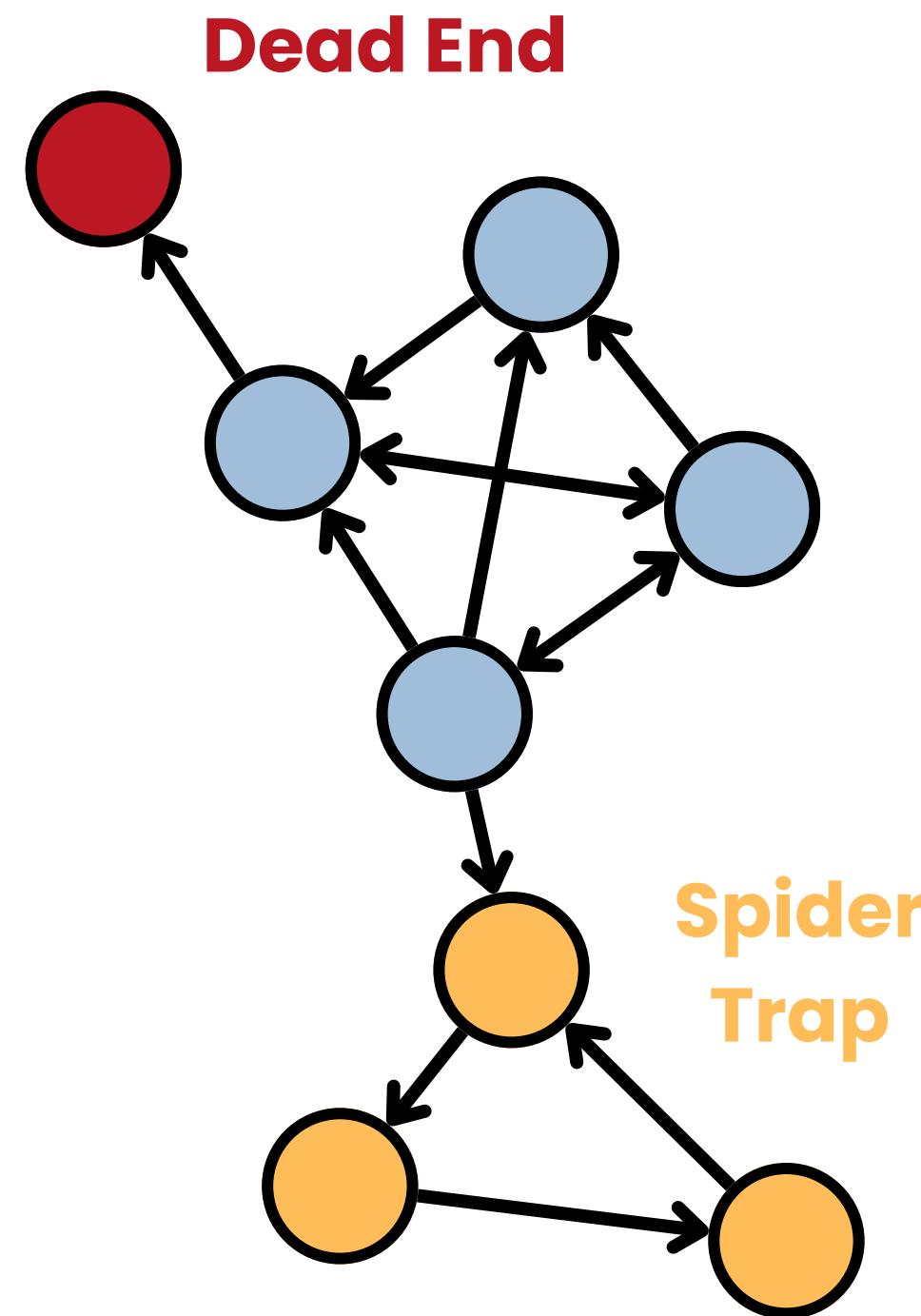
The interpretation of Google PageRank as a random walk makes it clear that we have some potential problems

Dead Ends

- There could be websites with incoming links, but no outgoing links, this would trap the random surfer

Spider Traps

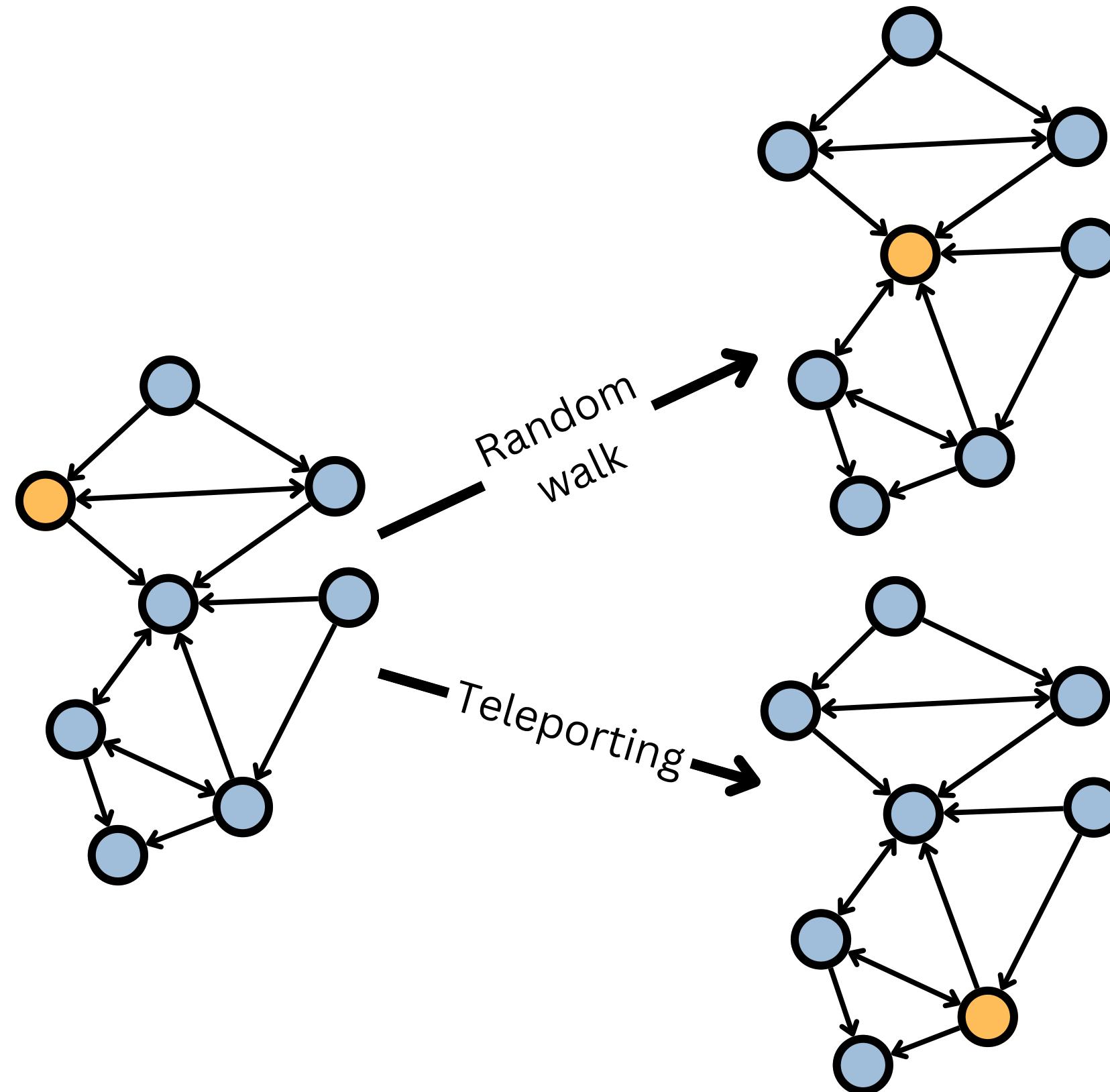
- There could be a set of websites with a one-way connection to the rest of the web. Once reached them, the random surfer would be trapped within them



Teleporting

Page and Brin had a brilliant solution to this problem: introducing teleportation

- at each time step, the random surfer can
 - follow an outgoing link at random
 - perform a teleportation to a random page
- teleportation occurs with probability β between 0.1 and 0.2
- teleportation occurs with probability one from dead ends



Google Billion Dollars Equation

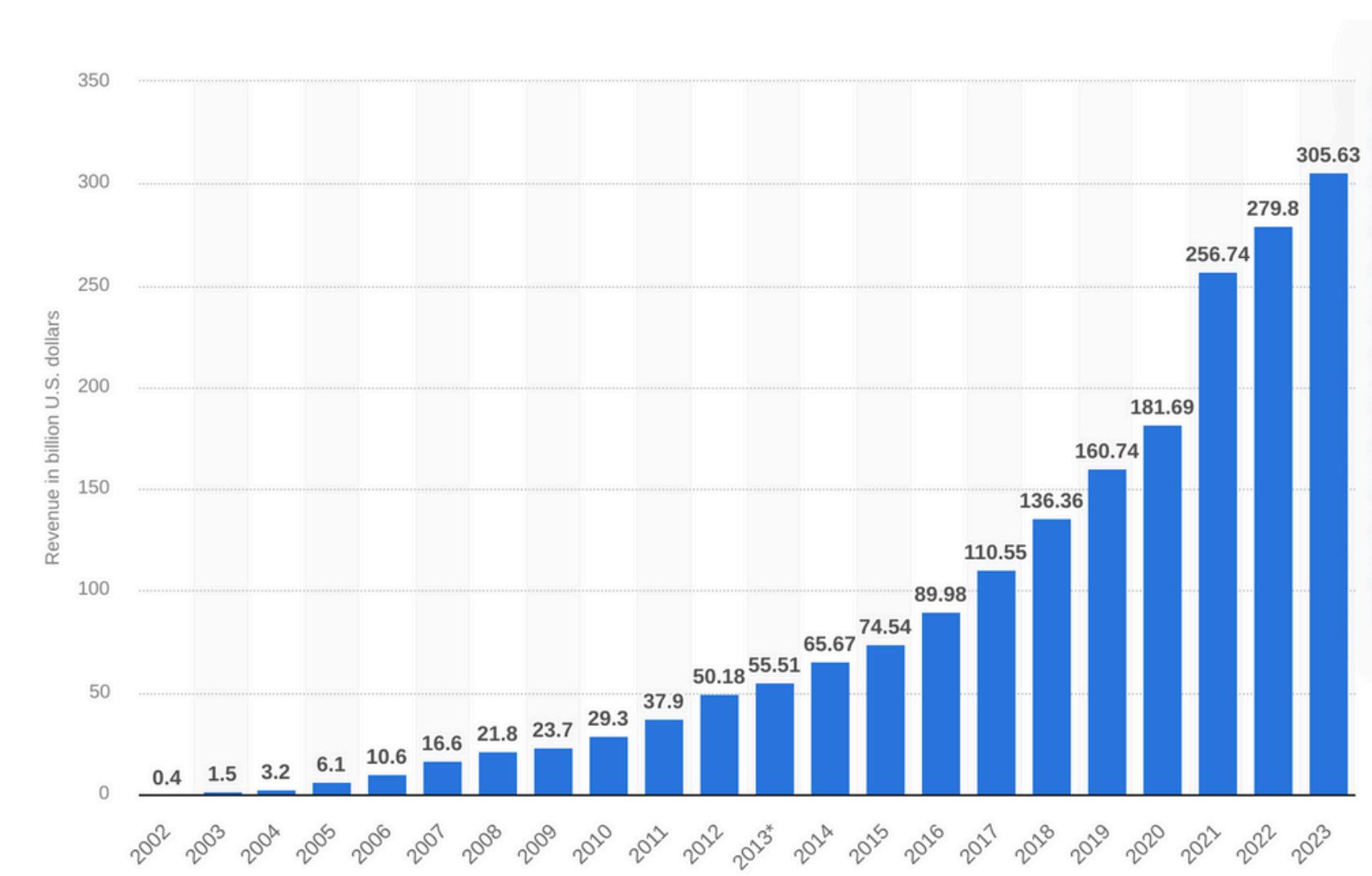
We have to modify the PageRank equation in order to include the teleport

- we add a probability to jump from any page to the target page

$$P(i) = \beta \sum_j \frac{P(j)}{N} + \sum_j (1 - \beta) P(j) M_{ji}$$

- we use the fact that $P(j)$ is a probability so it sums to one

$$P(i) = \frac{\beta}{N} + \sum_j (1 - \beta) P(j) M_{ji}$$



Topic-Specific PageRank

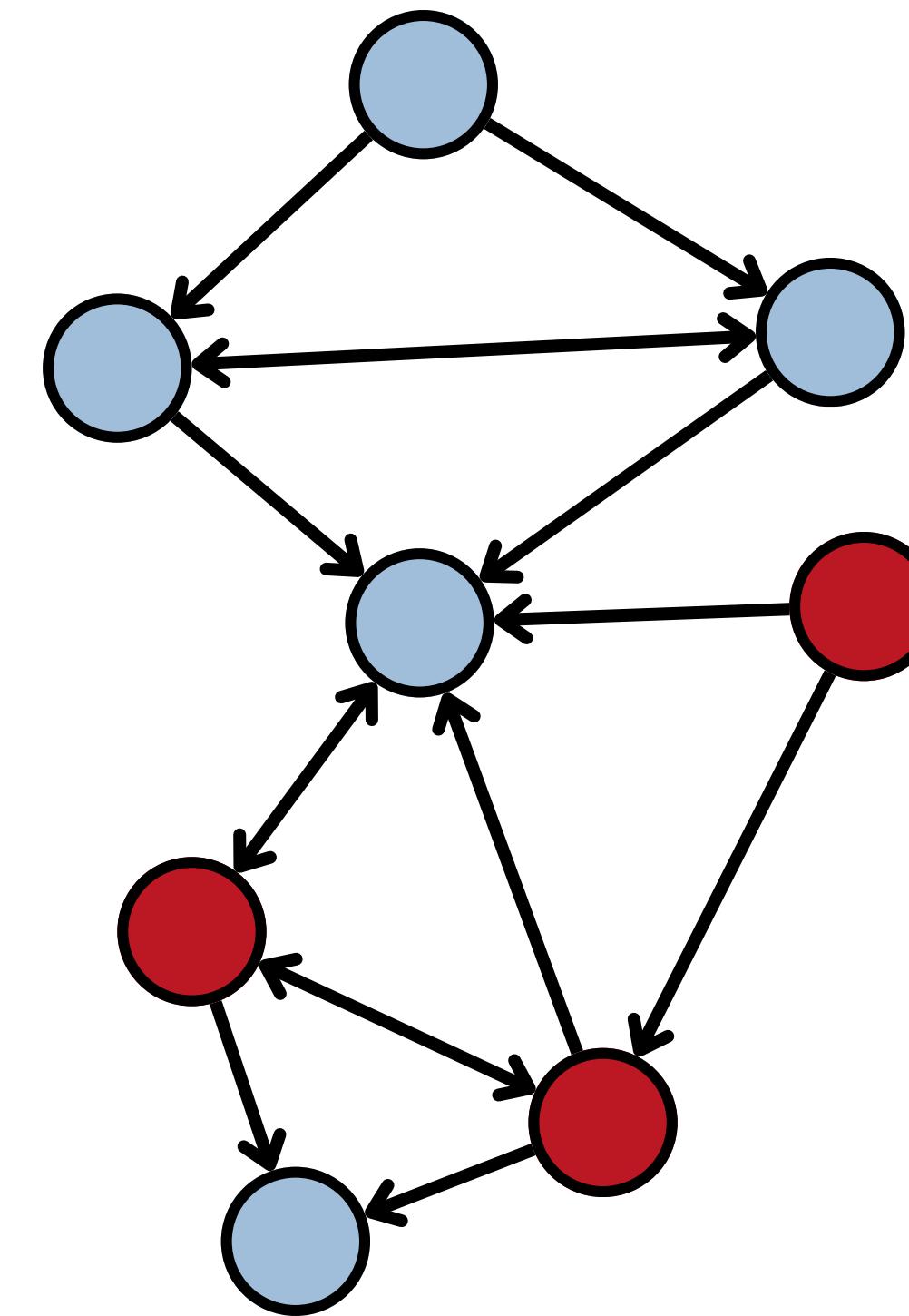
In the current formulation PageRank assigns a reputation score to each page

- it is not sensitive to specific queries
- it does not allow users to search by topics

We can overcome this limit introducing the Topic-Specific PageRank

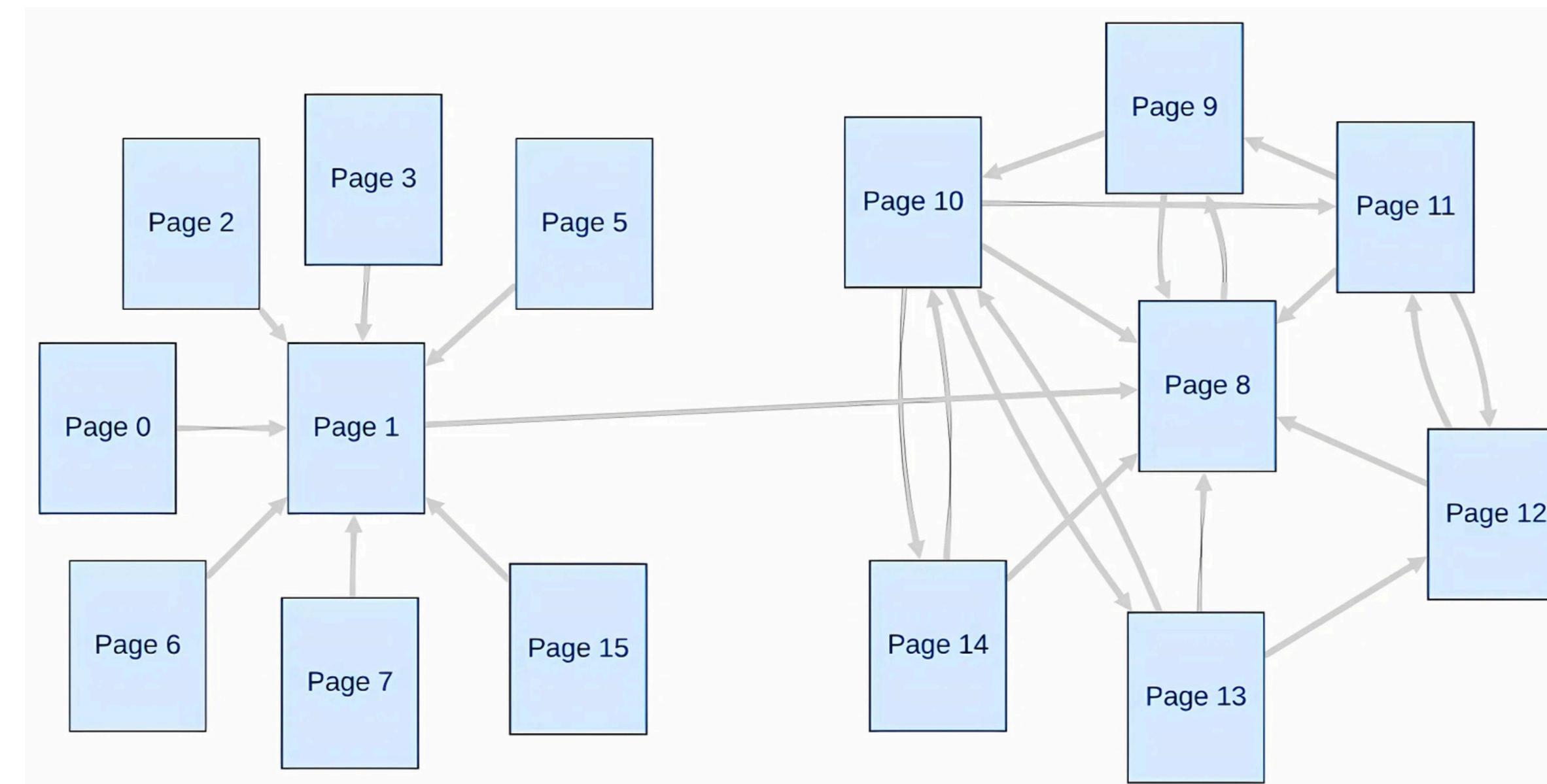
- given a set of keywords we select all websites containing them
- when we perform a teleportation, we only teleport on a page belonging to the selected subset

In this way we bias the random walk toward the pages we are interested to



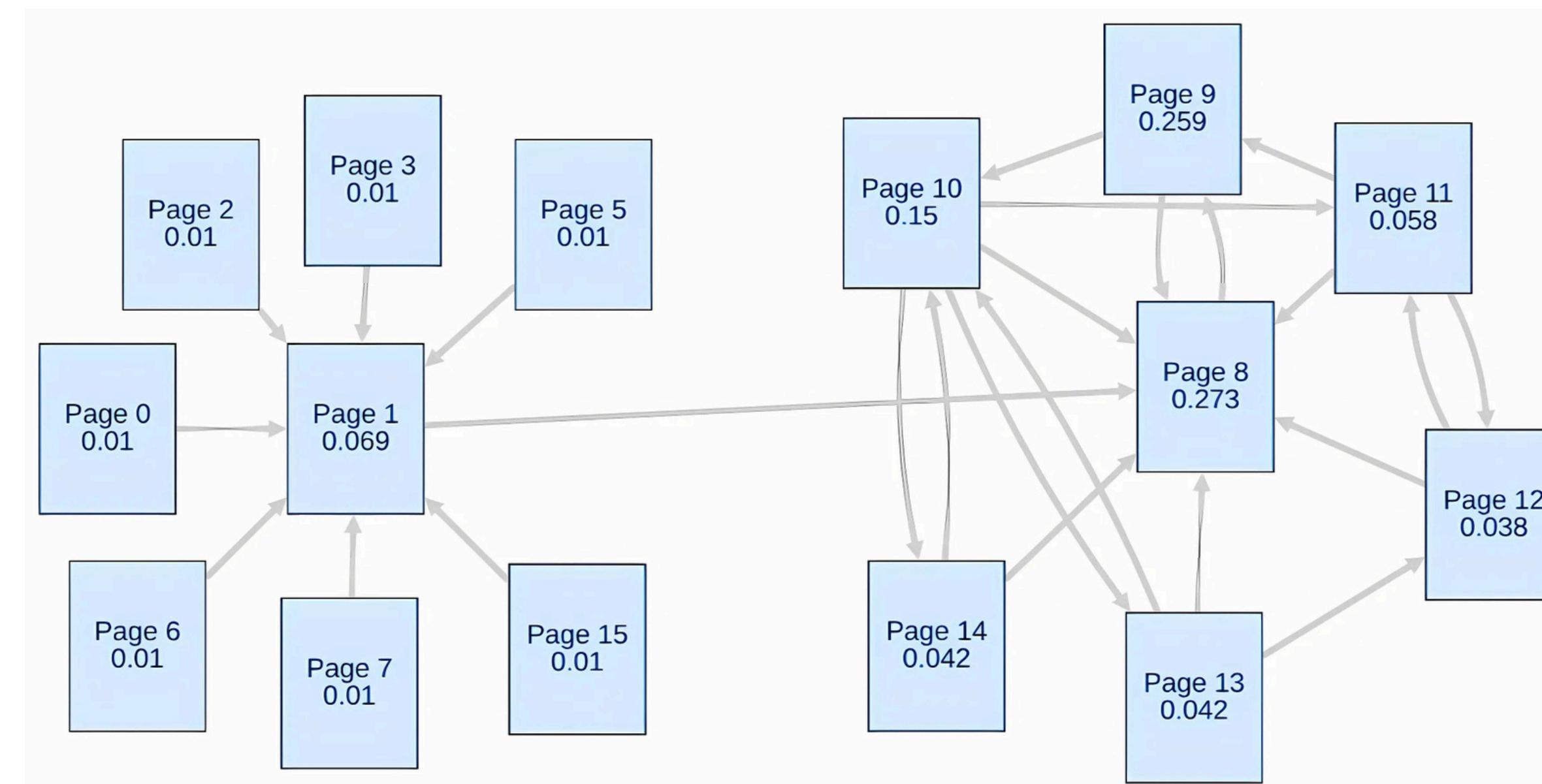
PageRank Example

Let's see a practical example of PageRank. Which page in the network has the highest PageRank?



PageRank Example

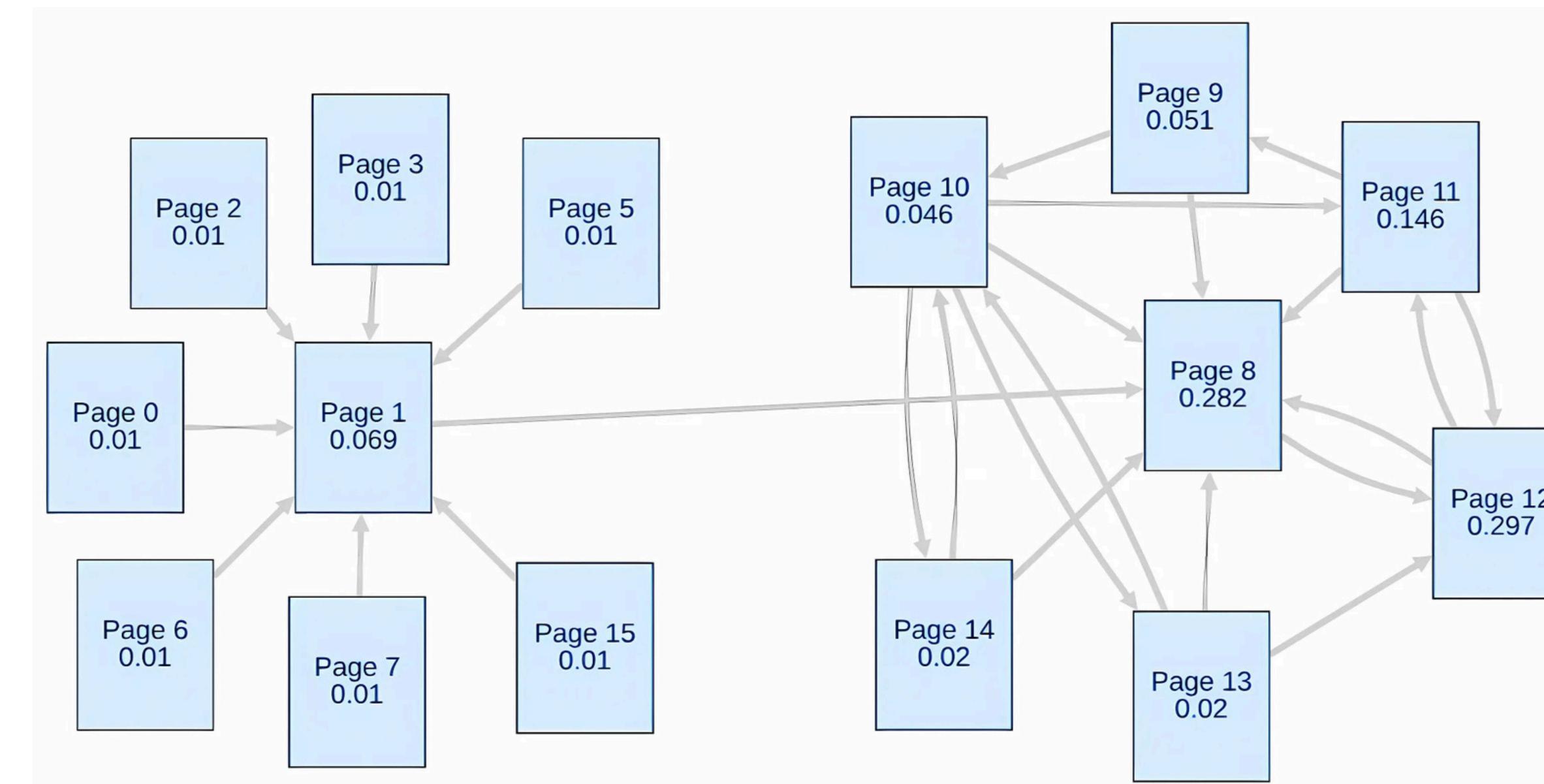
Even if Page 1 and Page 8 have the same degree, Page 8 has a much higher PageRank. Also Page 9, despite a much lower degree, has an higher PageRank



PageRank Example

A small change in the network, can strongly influence the PageRank values

<http://computerscience.chemeketa.edu/ascholer/cs160/WebApps/PageRank/>



Centrality Measures



The Concept of Node Centrality

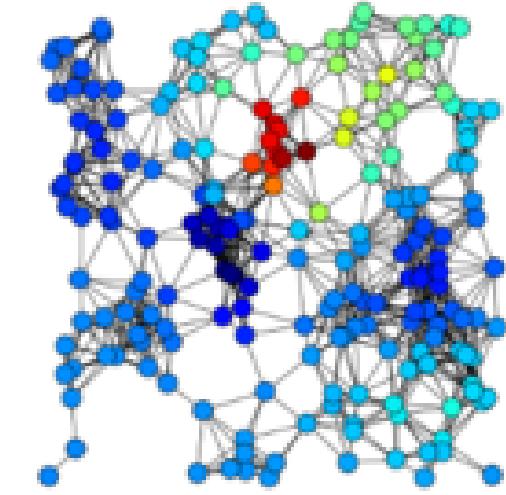
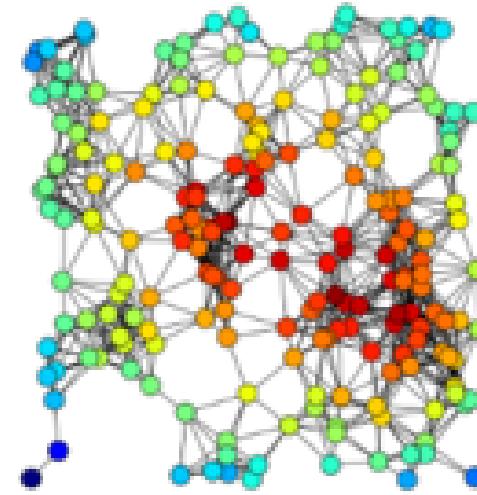
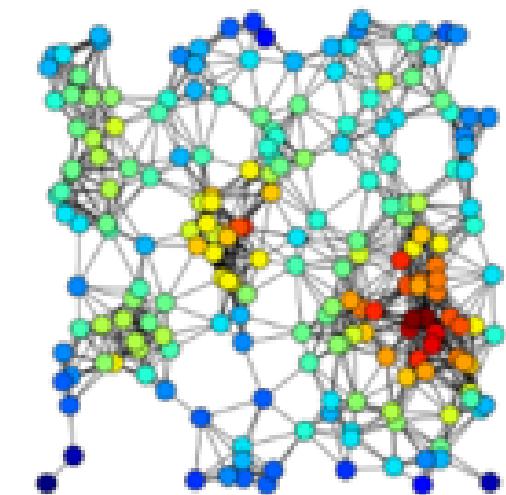
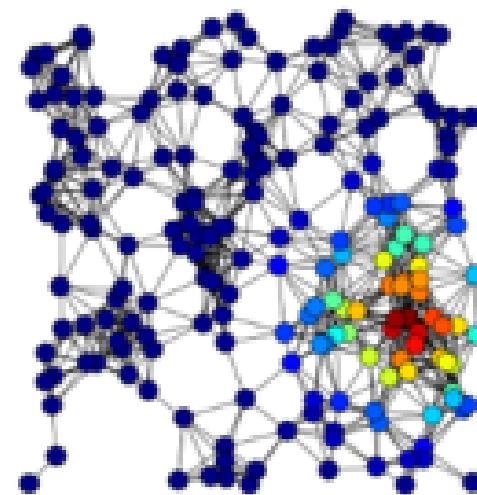
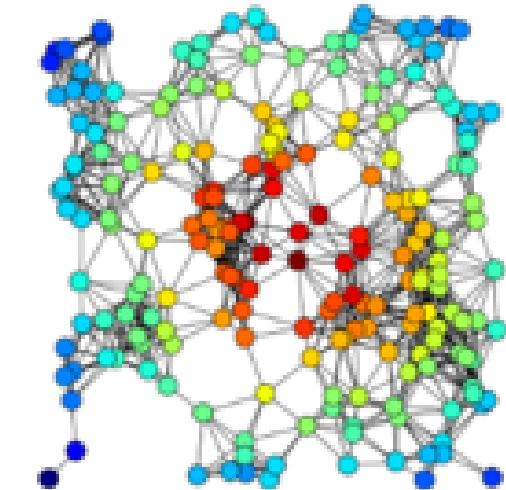
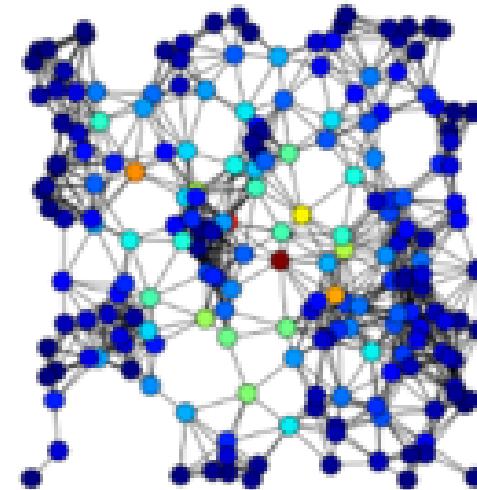
Centrality measures quantify the importance or influence of a node in a network.

- Importance depends on a node's position, not just on its degree
- Many features can be relevant
 - Pathways
 - Bridging
 - Influence

There are two main types of centralities

- degree based
- shortest-path based

PageRank is a degree based centrality measure



Eigenvector Centrality

Eigenvector Centrality is a measure of node importance where connections to influential neighbors increase a node's score

- It's computed using the principal eigenvector of the adjacency matrix A:
 $x = \lambda \cdot A \cdot x$ here x is the centrality vector and λ is the largest eigenvalue of A

There are some similarities with the PageRank

- Both consider the importance of connections, not just their number.
- Influence propagates through the network.

However there are also differences

- Eigenvector centrality doesn't include the random surfing model used in PageRank.
- Influence spreads equally across all connections, unlike PageRank which adjusts for out-degrees.

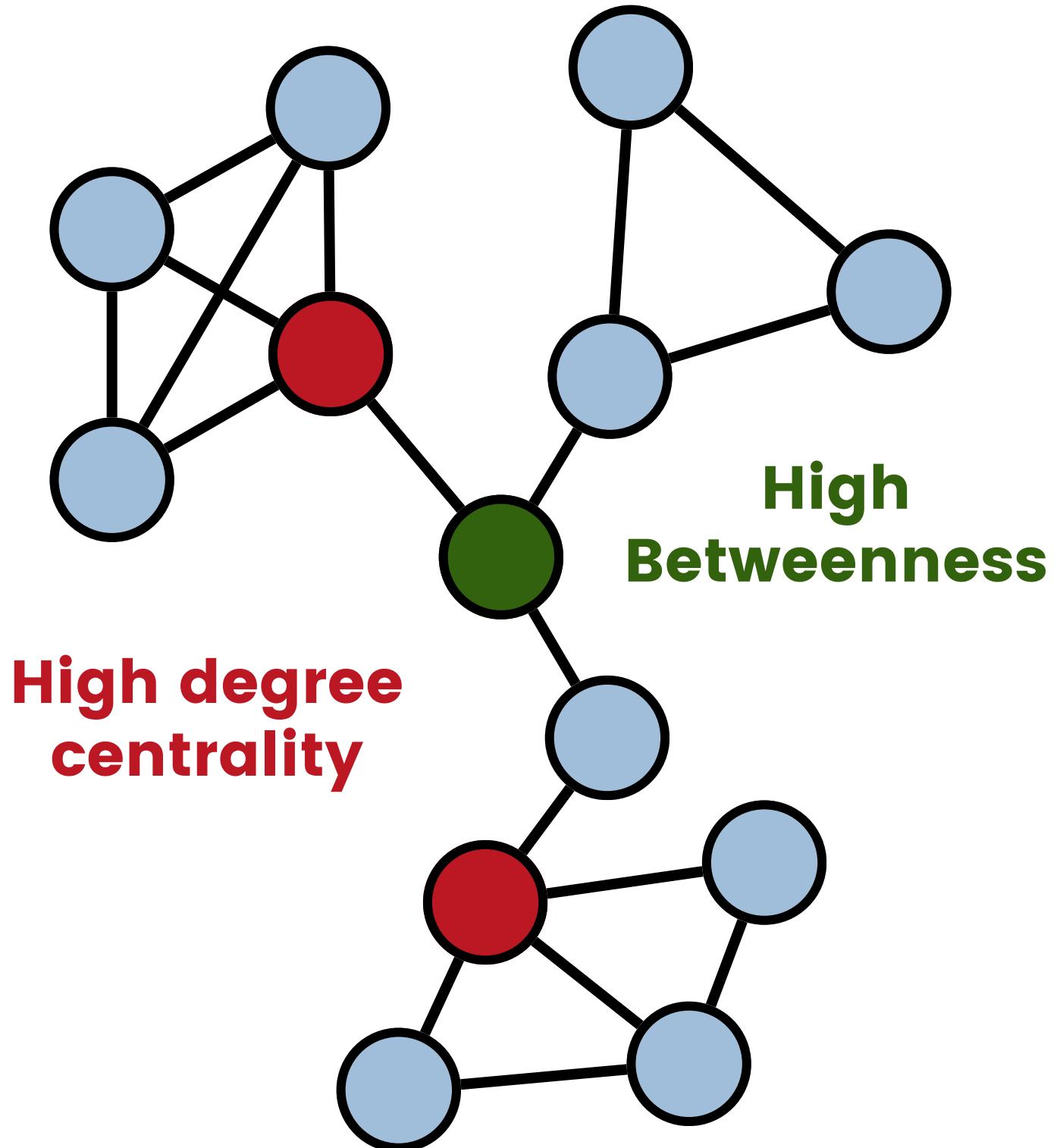
Betweenness Centrality

Betweenness Centrality measures a node's importance based on how often it lies on the shortest paths between other nodes

- Nodes with high betweenness act as bridges in the network, controlling the flow of information.

How It Works

1. Calculate all shortest paths in the network.
2. Count the number of these paths that pass through each node.
3. Normalize the value based on the size of the network.

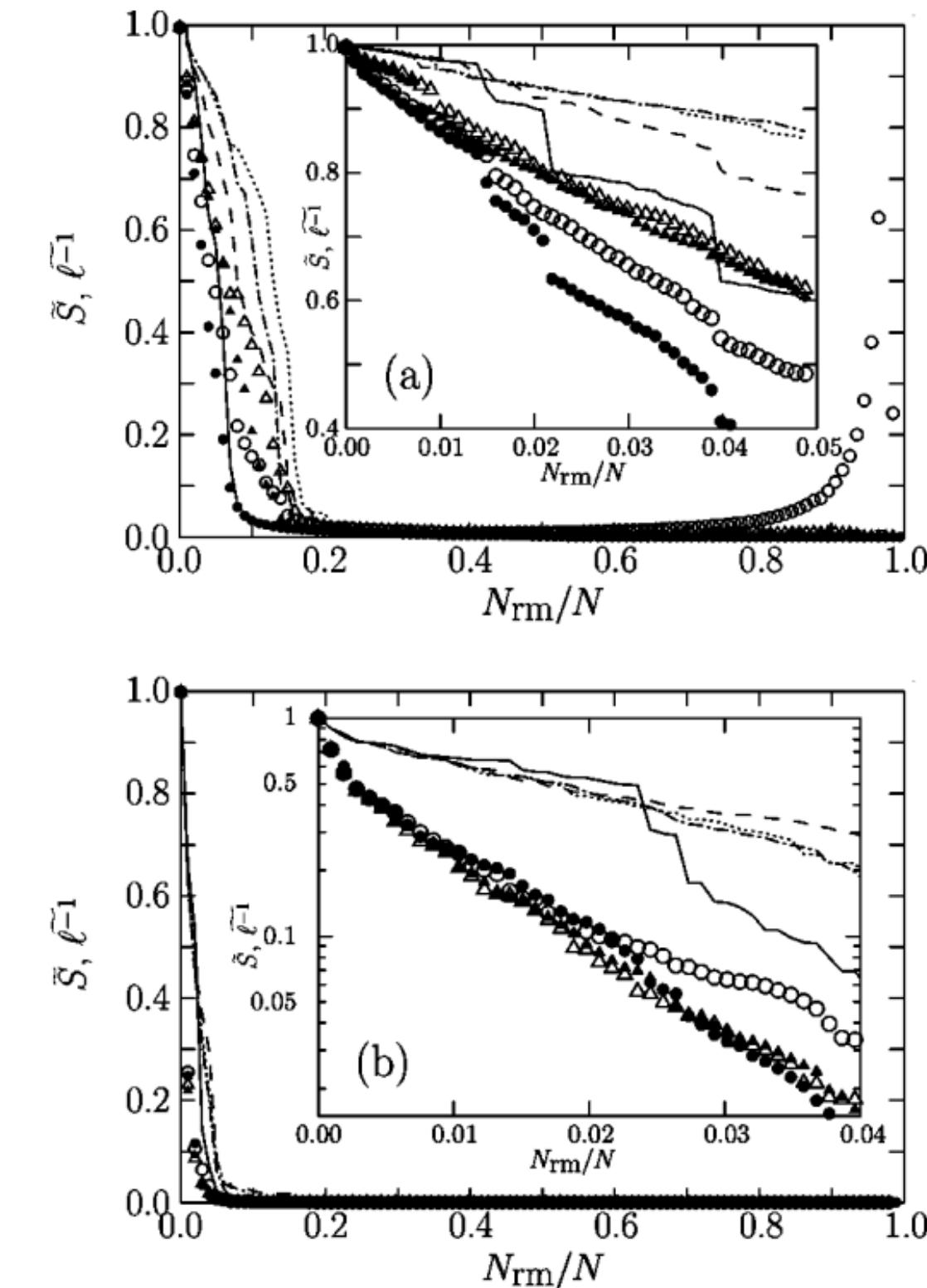


Betweenness and Attacks

Betweenness centrality plays an important role in Network Vulnerability

- Nodes with high betweenness centrality are critical for network connectivity.
- Their removal disrupts shortest paths, fragmenting the network and reducing its efficiency.

The plot shows the results of attacks to a internet network and a citation network. Solid lines show the size of the giant component using betweenness to remove nodes



Holme, Petter, et al. "Attack vulnerability of complex networks." Physical review E 65.5 (2002): 056109.

Closeness Centrality

Closeness Centrality is a measure of how close a node is to all other nodes in a network

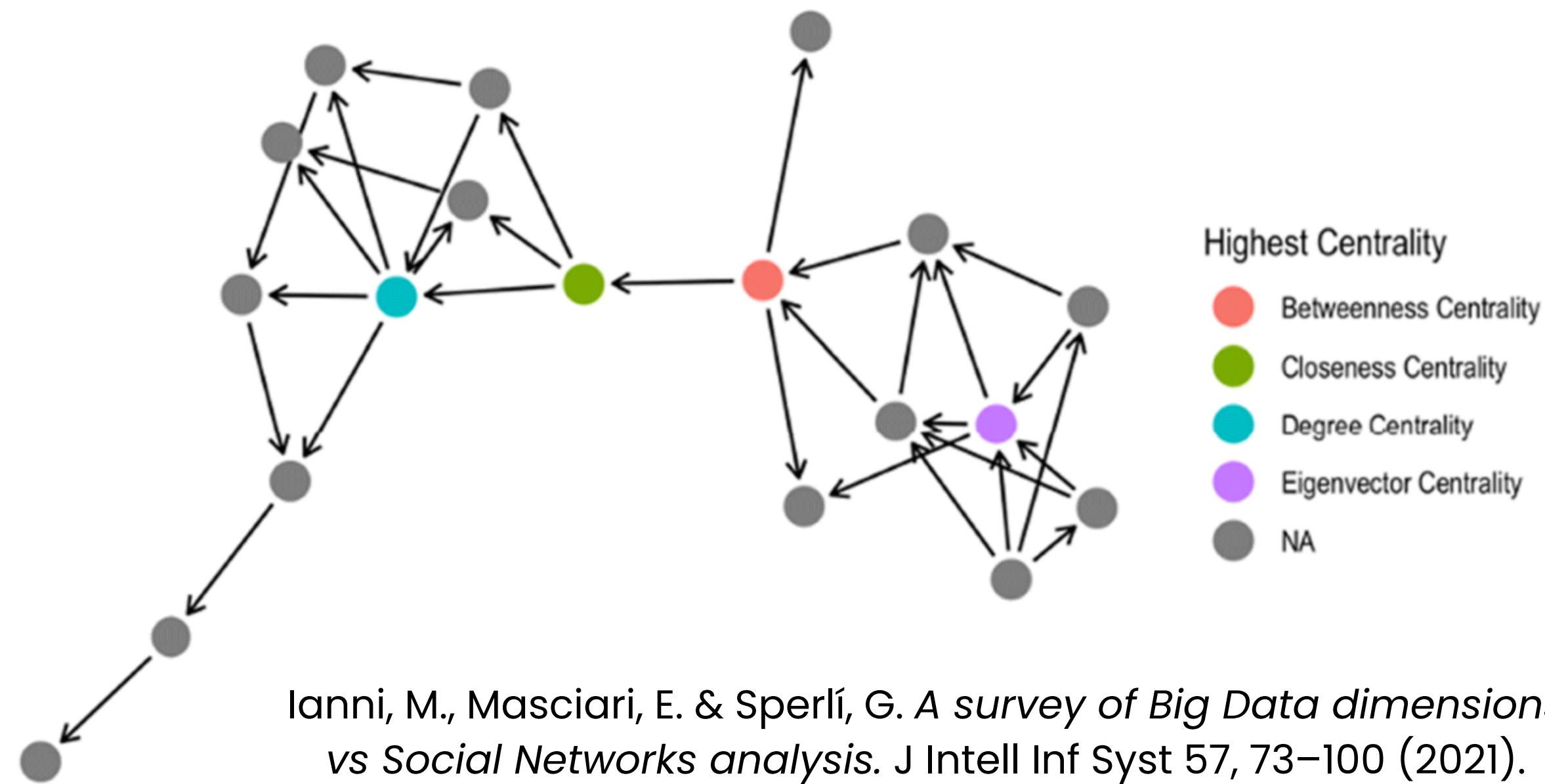
- Nodes with high closeness centrality can reach others quickly, making them influential in spreading information.
- Closeness centrality for a node i is defined as the reciprocal of the average shortest path length from i to all other nodes:

$$C(i) = \frac{N - 1}{\sum_{j \neq i} d(i, j)}$$

Closeness Centrality is particularly useful in social networks, transportation systems, and communication networks to identify strategically positioned nodes.

Comparison of Centrality Measures

The example below shows the top node for the different centrality measures we discussed. In many cases there is a strong correlation between them.





Analyzing Criminal Networks

Criminal Organizations as Complex Networks

Criminal organizations often operate as covert, decentralized networks to avoid detection

- network science can help disrupt their activities by targeting key individuals
- criminals are interpreted as nodes and their interactions as edges.
- network measures can be used to identify key players and roles

There are also several challenges, for instance

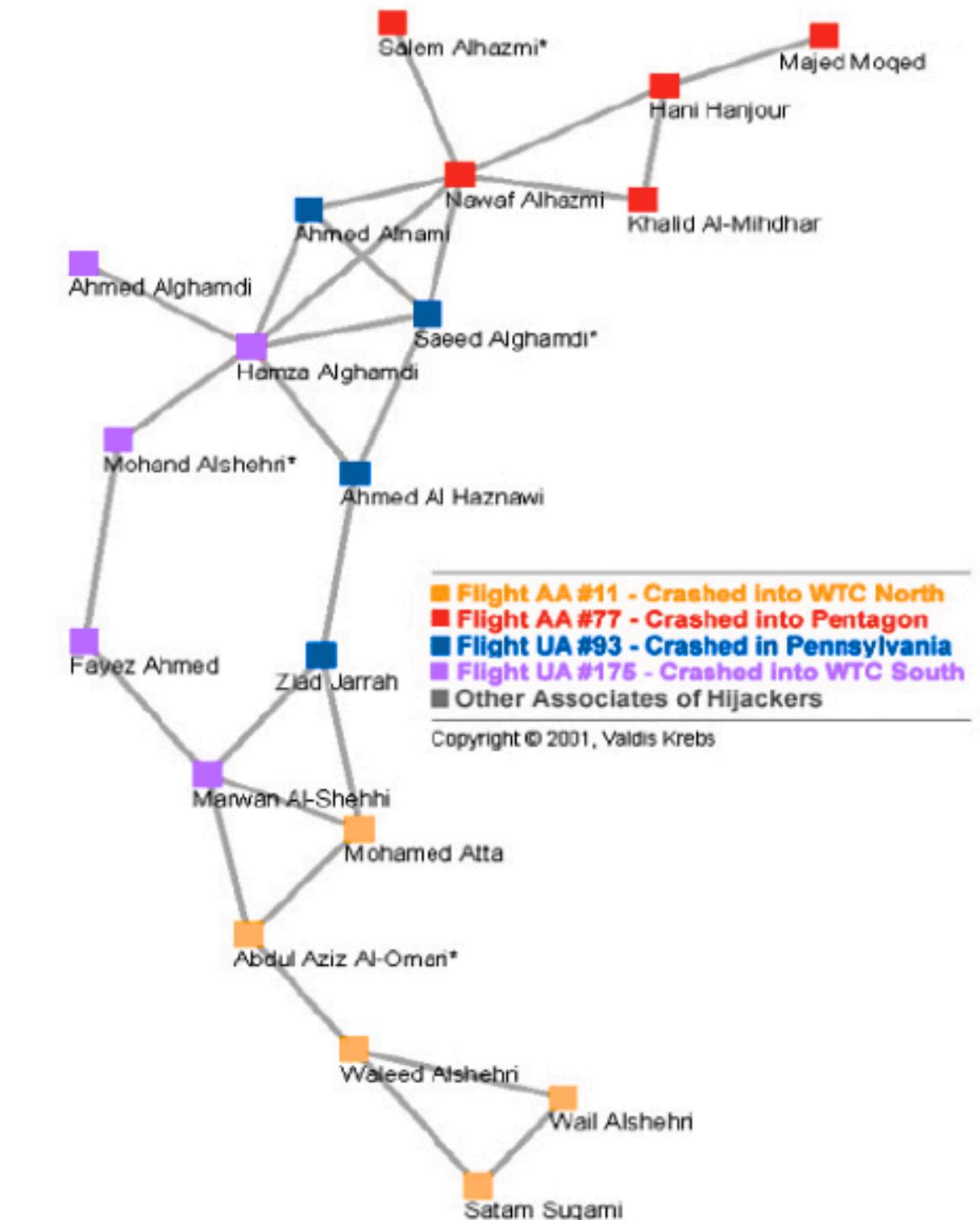
- 1. Incomplete Data:** Hidden nodes/connections.
- 2. Dynamic Nature:** Networks adapt over time.



9/11 Terrorist Cell

9/11 terrorist attacks were performed by a total of 19 hijackers

- the figure shows the network of prior trusted contacts (living and learning together)
- it is characterized by a very low density
 - the operation is robust to arrests
 - however communication is less efficient

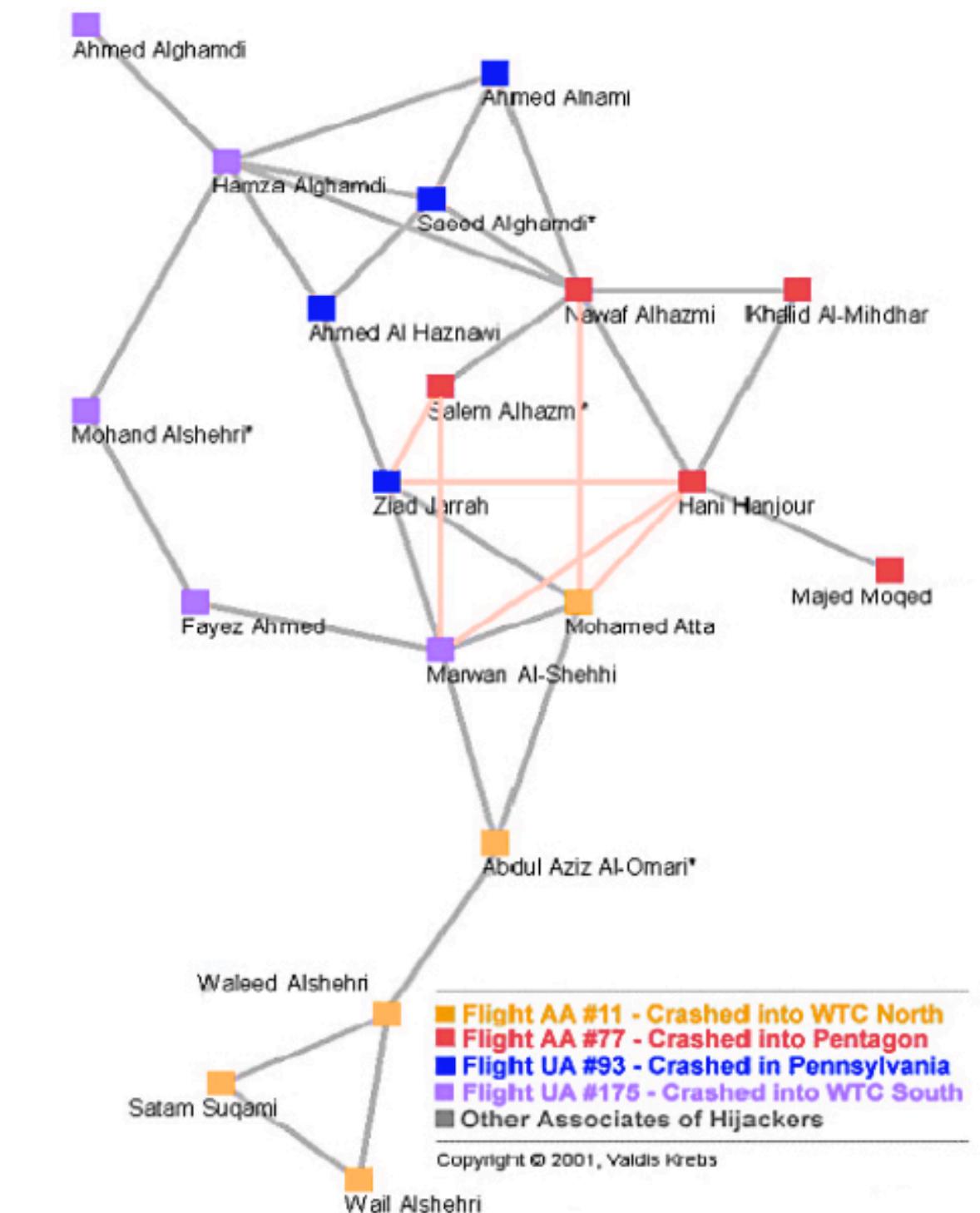


Krebs, Valdis E. "Mapping networks of terrorist cells." *Connections* 24.3 (2002): 43-52.

9/11 Terrorist Cell

9/11 terrorist attacks were performed by a total of 19 hijackers

- the figure shows the network of prior trusted contacts (living and learning together)
- it is characterized by a very low density
 - the operation is robust to arrests
 - however communication is less efficient
- meetings were held to connect distant parts of the network and coordinate tasks
- these meetings added shortcuts to the network (shown in pink)

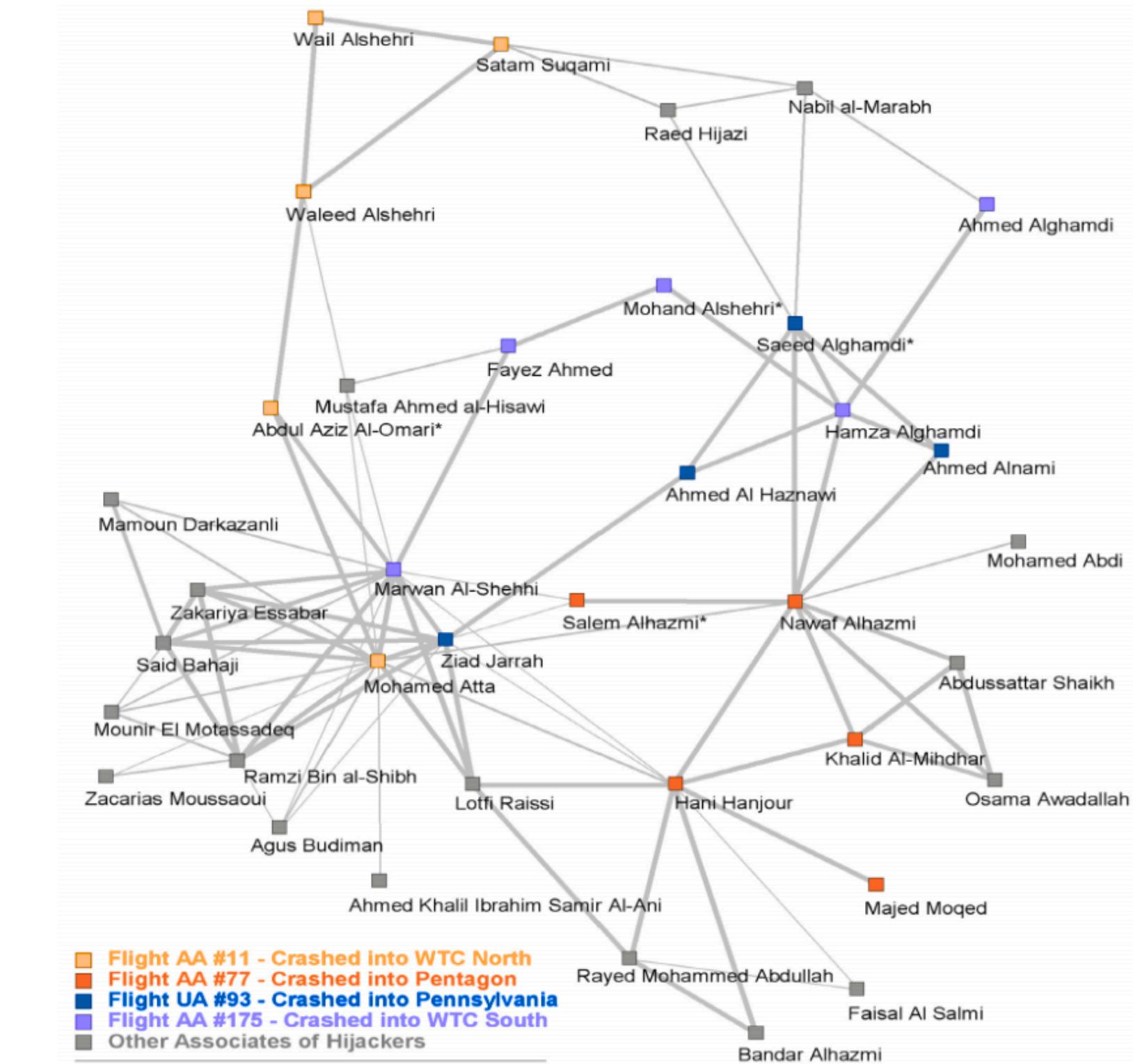


Krebs, Valdis E. "Mapping networks of terrorist cells." *Connections* 24.3 (2002): 43-52.

Hijacker's Network Neighborhood

From the trusted contacts network it seems that the hijackers had very little contacts and didn't know each other

- a different picture emerges looking at the prior contacts network
- these ties were forged in school, through kinship, and training/fighting
- this network was “dormant” in the USA but it ensured robustness to arrests



Copyright ©, Valdis Krebs

Centrality Measures

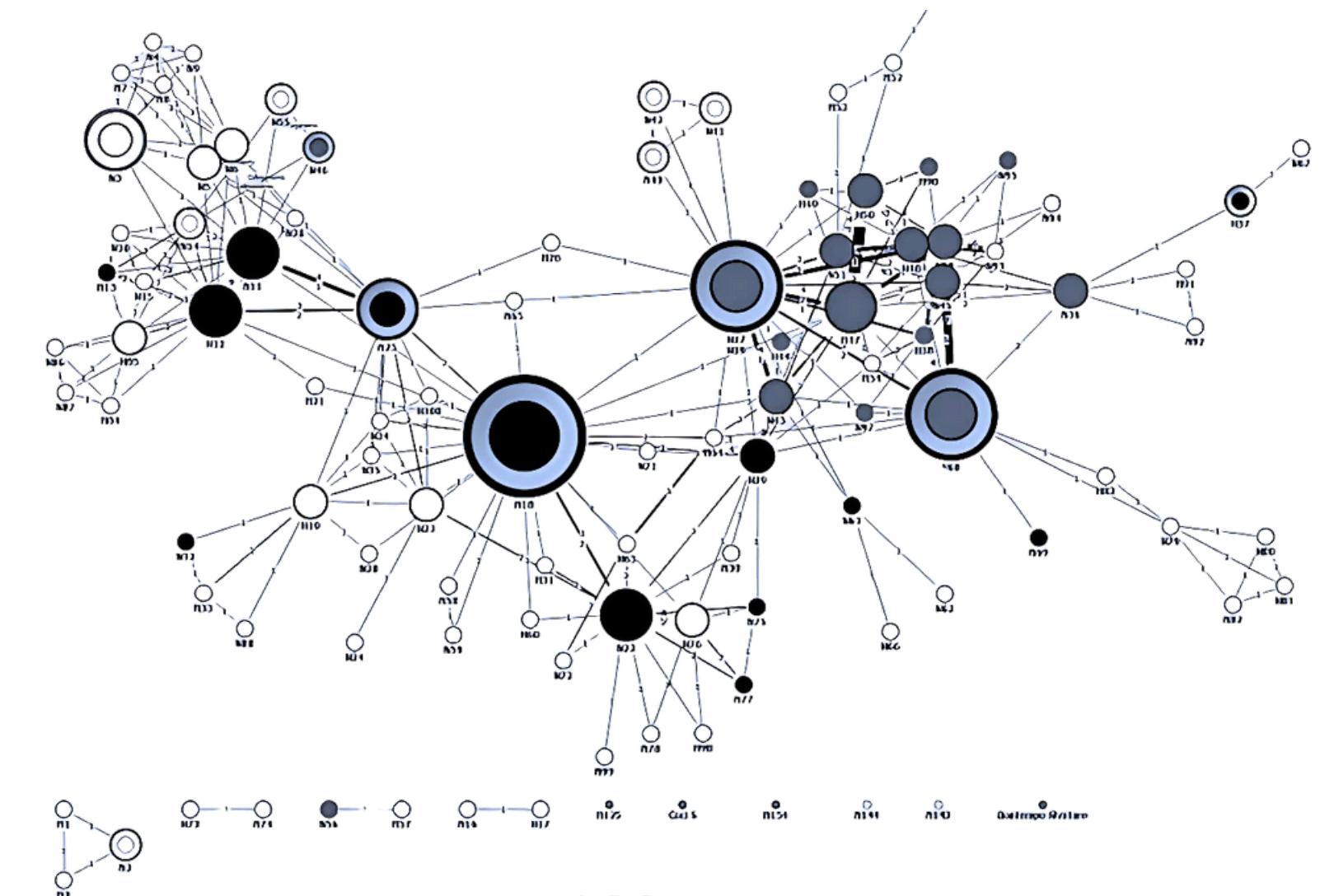
By applying various centrality measures to the Hijacker's Network Neighborhood Mohamed Atta appears as the most relevant node. This is consistent with the investigations that determined he was the leader of the operation

	Degrees		Betweenness		Closeness
0.417	Mohamed Atta	0.334	Nawaf Alhazmi	0.571	Mohamed Atta
0.389	Marwan Al-Shehhi	0.318	Mohamed Atta	0.537	Nawaf Alhazmi
0.278	Hani Hanjour	0.227	Hani Hanjour	0.507	Hani Hanjour
0.278	Nawaf Alhazmi	0.158	Marwan Al-Shehhi	0.500	Marwan Al-Shehhi
0.278	Ziad Jarrah	0.116	Saeed Alghamdi*	0.480	Ziad Jarrah
0.222	Ramzi Bin al-Shibh	0.081	Hamza Alghamdi	0.429	Mustafa al-Hisawi
0.194	Said Bahaji	0.080	Waleed Alshehri	0.429	Salem Alhazmi*
0.167	Hamza Alghamdi	0.076	Ziad Jarrah	0.424	Lotfi Raissi
0.167	Saeed Alghamdi*	0.064	Mustafa al-Hisawi	0.424	Saeed Alghamdi*
0.139	Lotfi Raissi	0.049	Abdul Aziz Al-Omari*	0.419	Abdul Aziz Al-Omari*

Mafia Networks

Mafia organizations are built on relationships and collaboration. Networks help uncover the hidden structures of clans and their operations:

- Nodes are members of mafia clans
- Edges are connections based on meetings attendance
- Weights reflect the number of meetings attended together
- Different colors represent distinct mafia clans, circled dot are bosses



Cavallaro, Lucia, et al. "Graph and network theory for the analysis of criminal networks." DS and IoT (2021): 139-156.

Humans vs Social Capital

Betweenness centrality highlights a dichotomy between human capital and social capitals. Many individuals with high betweenness are just members, but play a more important role than leader in acting as bridges

Position	Node ID	Betweenness centrality	Role
1	18	0.373	Leader
2	47	0.22	Member
3	27	0.159	Leader
4	68	0.126	Member
5	12	0.117	Member
6	25	0.114	Leader
7	29	0.09	Member
8	36	0.072	Member
9	22	0.069	Member
10	11	0.063	Member

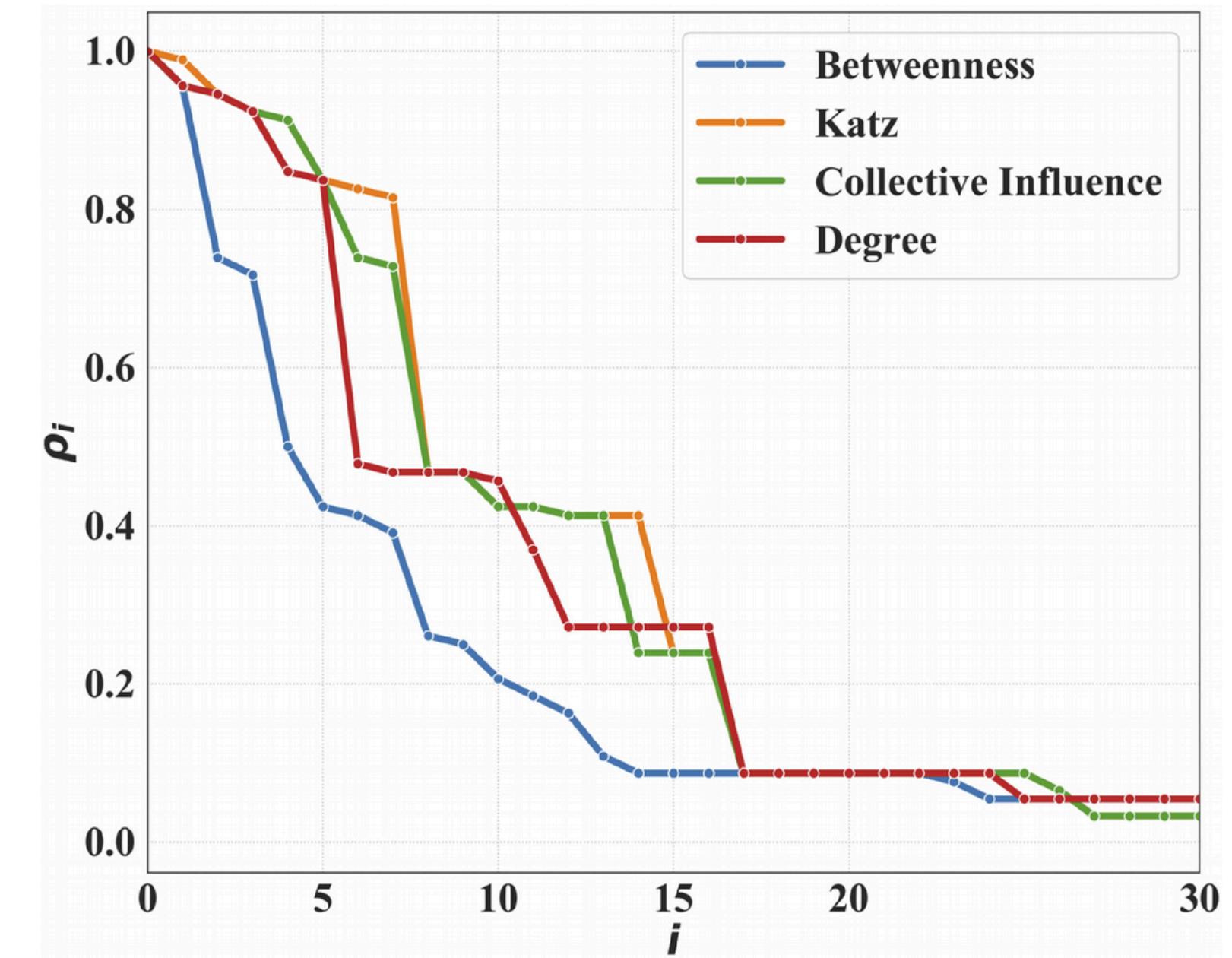
Cavallaro, Lucia, et al. "Disrupting resilient criminal networks through data analysis: The case of Sicilian Mafia." Plos one 15.8 (2020): e0236476.

Guiding Police Operations

Network science can be used to guide police operations and arrests

- centrality measures can be used to identify the most influential criminals
- instead of performing random arrests, the arrest of these individuals should be prioritized
- betweenness centrality is the most efficient measure for this task

Network science can be a valuable tool in persecuting criminal activities



Conclusions

The Quest for Online Search Engines

From the public release of the WWW in 1991, the exponential growth of the number of pages made standard approaches to searches unfeasible

The PageRank

Google introduced the PageRank, which focuses of the role of pages within the network instead of the content of the pages

Centrality Measures

Different tasks require different centrality measures. We considered closeness, betweenness, eigenvector centrality and PageRank centrality

Analyzing Criminal Networks

Criminal organizations can be described as networks and network science can be used to analyze and disrupt them.

Quiz

- How many active pages does the web currently have?
- Do you know the concept of algorithmic bias? How does this apply to the PageRank?
- What does it mean for a node to have high closeness centrality but low degree centrality? Can you think of a real-world example?
- Which centrality measure might highlight inefficiencies in a company's communication network?
- Which other illegal activities could be investigated using network?
- What are the limits and challenges of network science in this area?