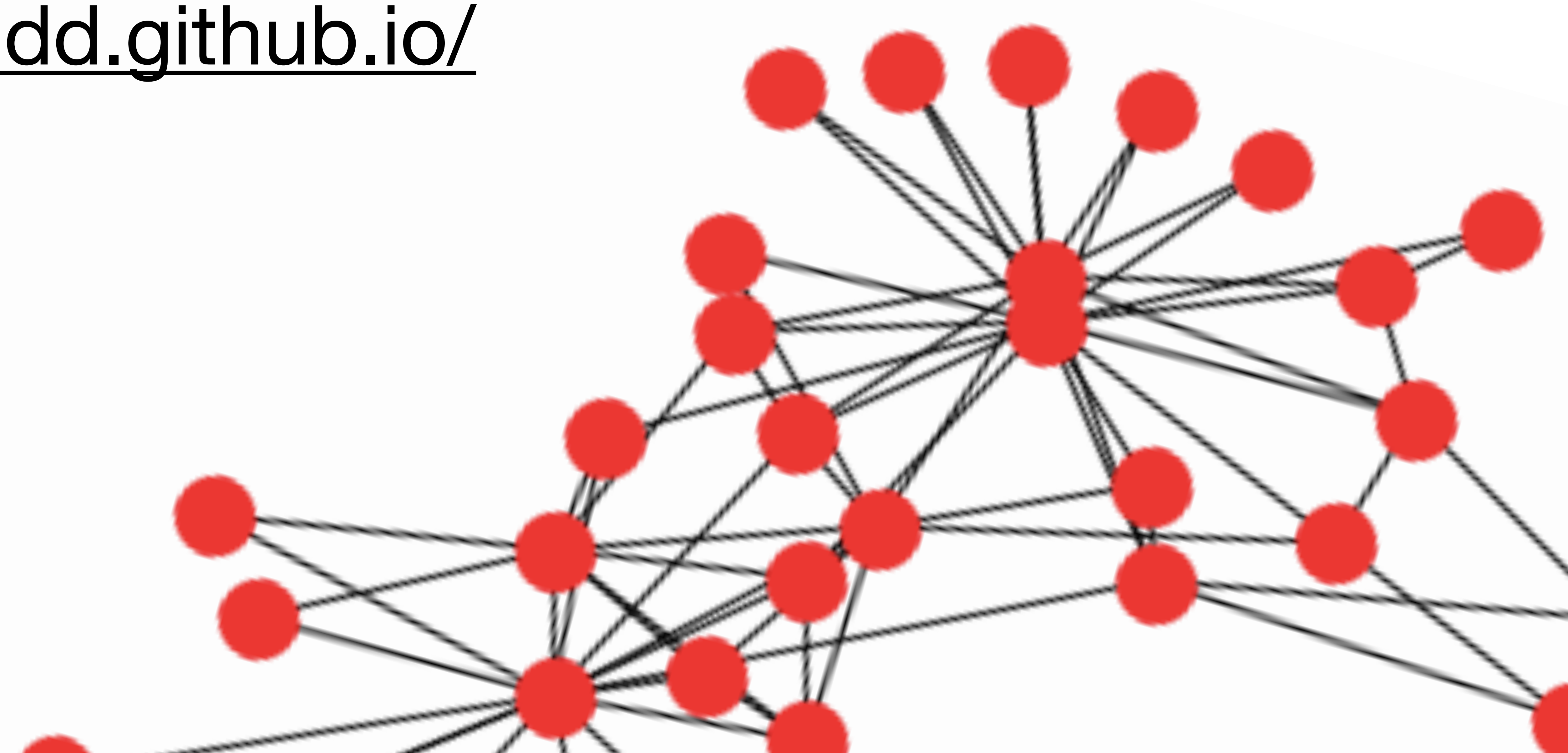
DMSN Tutorial 3: Network Centrality and Applications

Naomi Arnold
https://narnolddd.github.io/

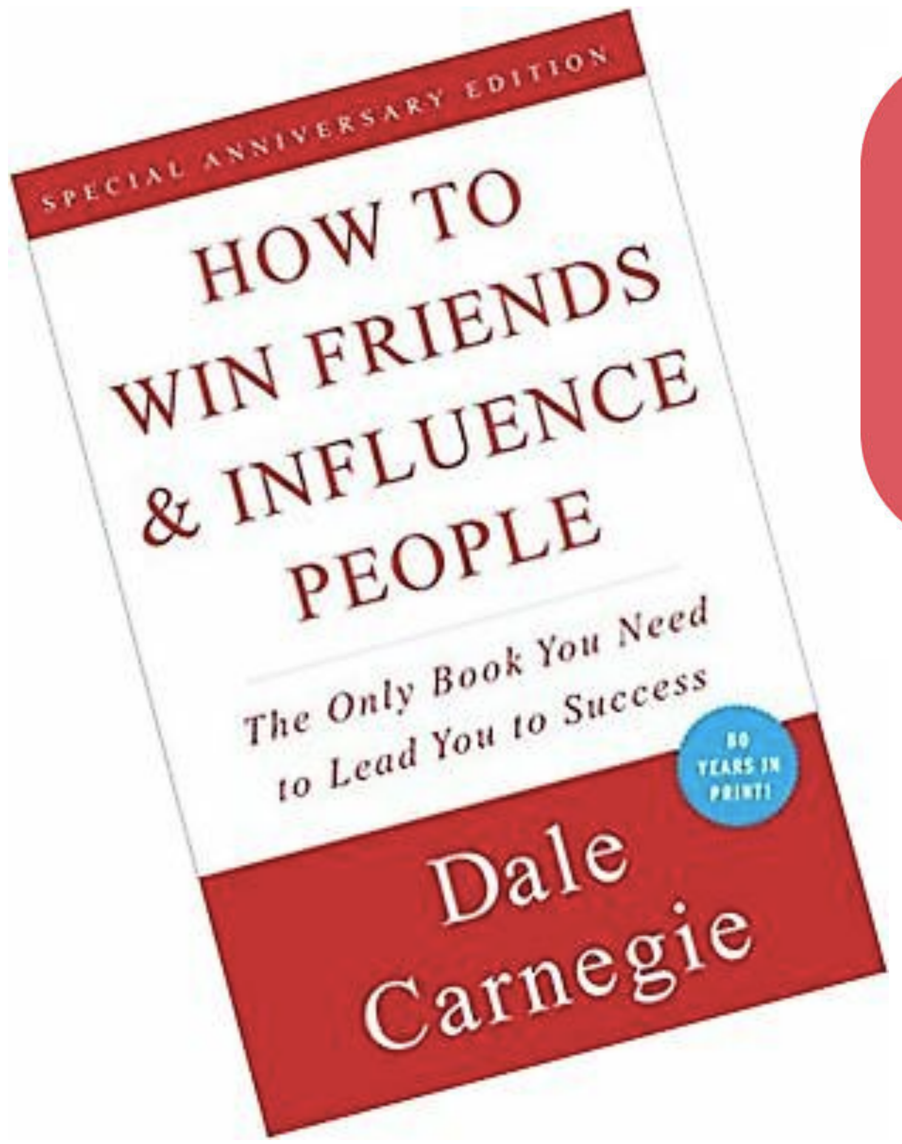


In this tutorial

- **Recap** different centrality measures
- Go over some **centrality calculations**
- Application to the **FIFA 2018** World Cup Final

Who is important in a network?

It depends what you mean by important…

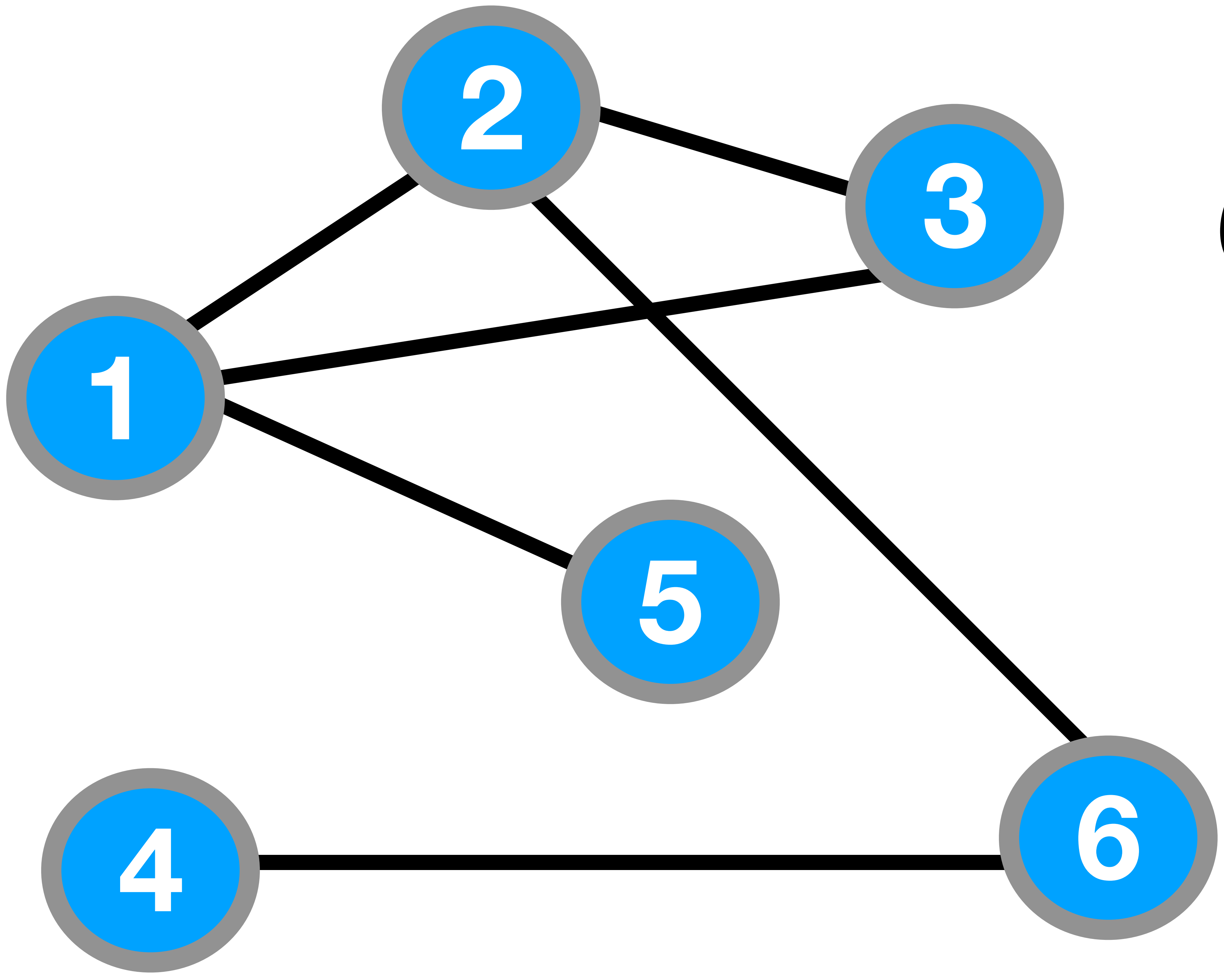


Who has **influence?**
(authority, popularity, …)

Who **facilitates connectivity**/‘glues’ the network together?

(In-)Degree Centrality

Number of connections (followers, friends, citations, …)



Degree of node i

$$C(i) = \frac{\boxed{k_i}}{\boxed{N-1}}$$

Number of **other** nodes in the network, to **normalise** (only really necessary if you're comparing more than one network)

$$C(1) = C(2) = 3/5$$

Freeman Network Centralisation

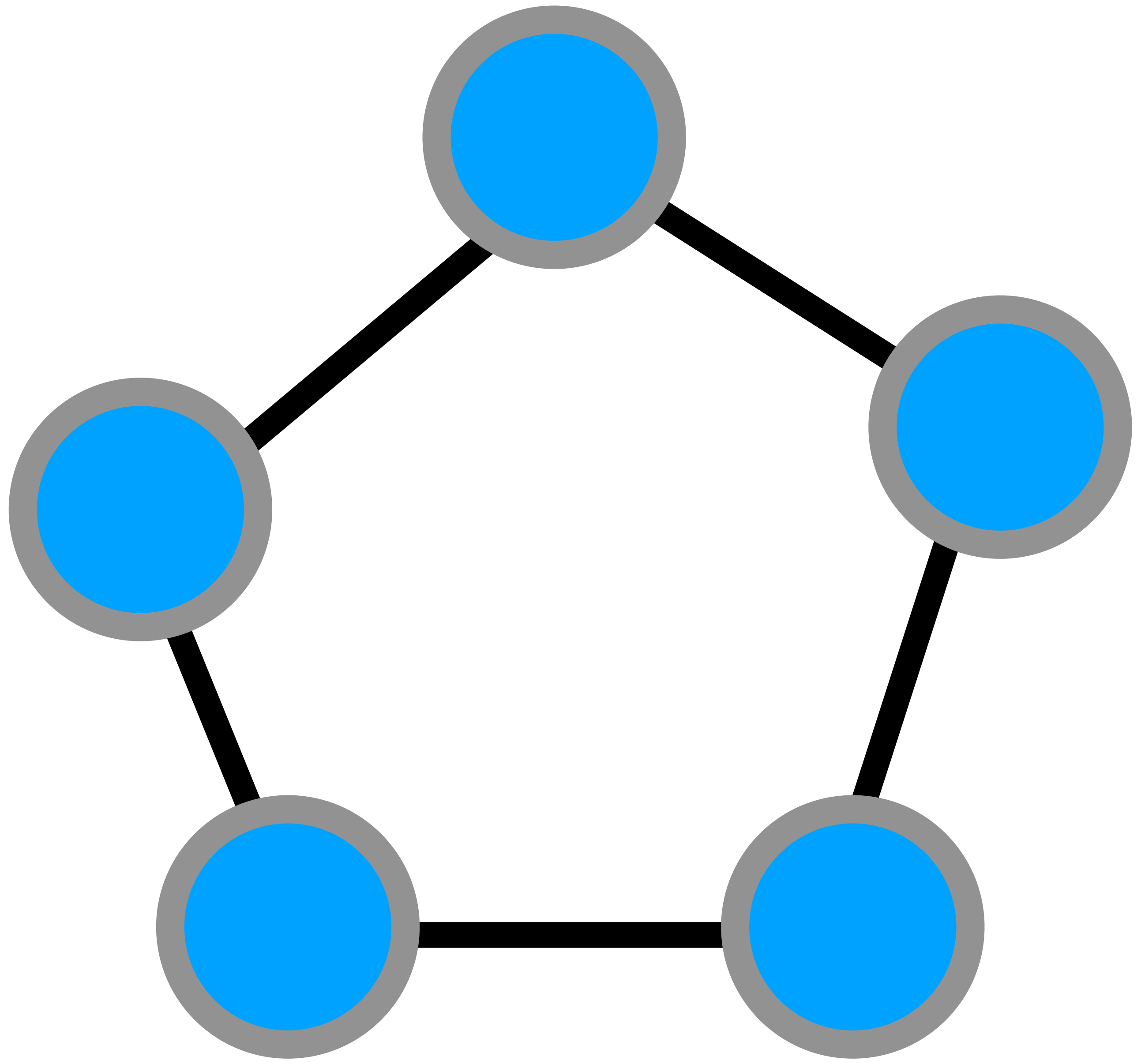
How far from **equal** is a network's degree distribution?

Compare the degree of each node with the largest degree

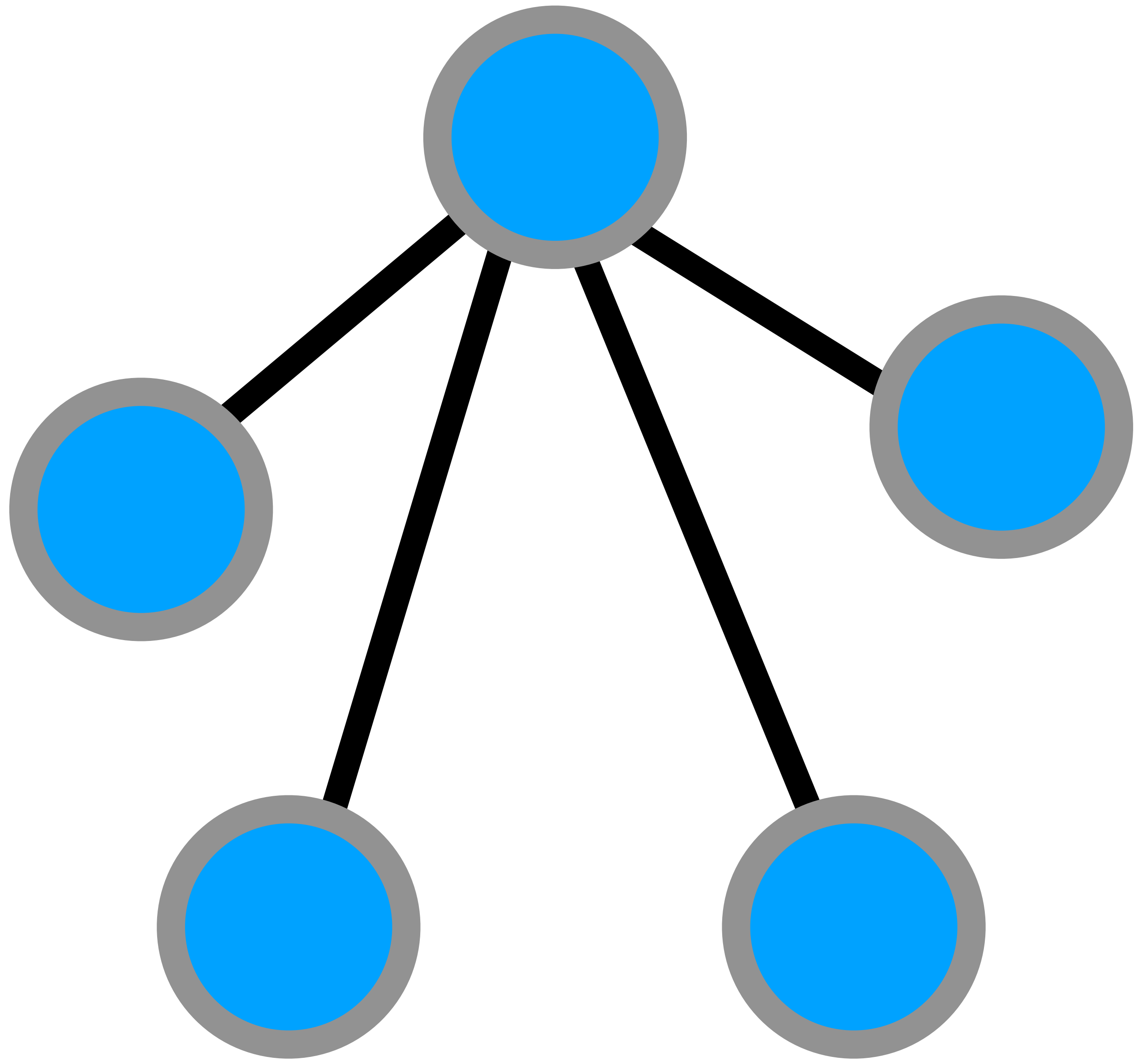
$$C_F(G) = \frac{\sum_{i=1}^{N} \boxed{k_{\max} - k_i}}{\boxed{(N-1)(N-2)}}$$

The largest possible value the top could be in a network of **N** nodes

Freeman Network Centralisation



$$C_F = 0$$

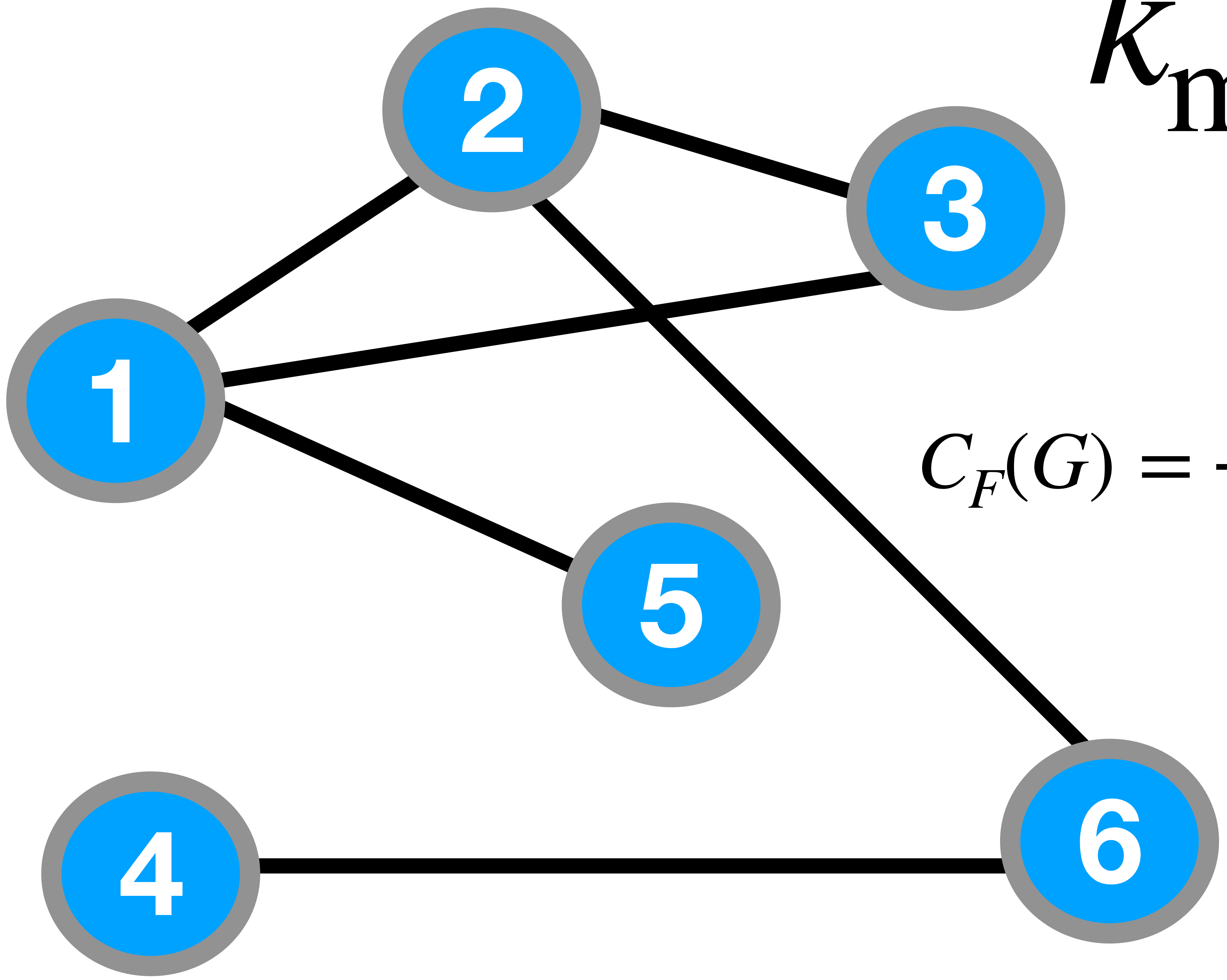


$$C_F = 1$$

Nodes have equal share of the links.

One node has all the links

Freeman Network Centralisation



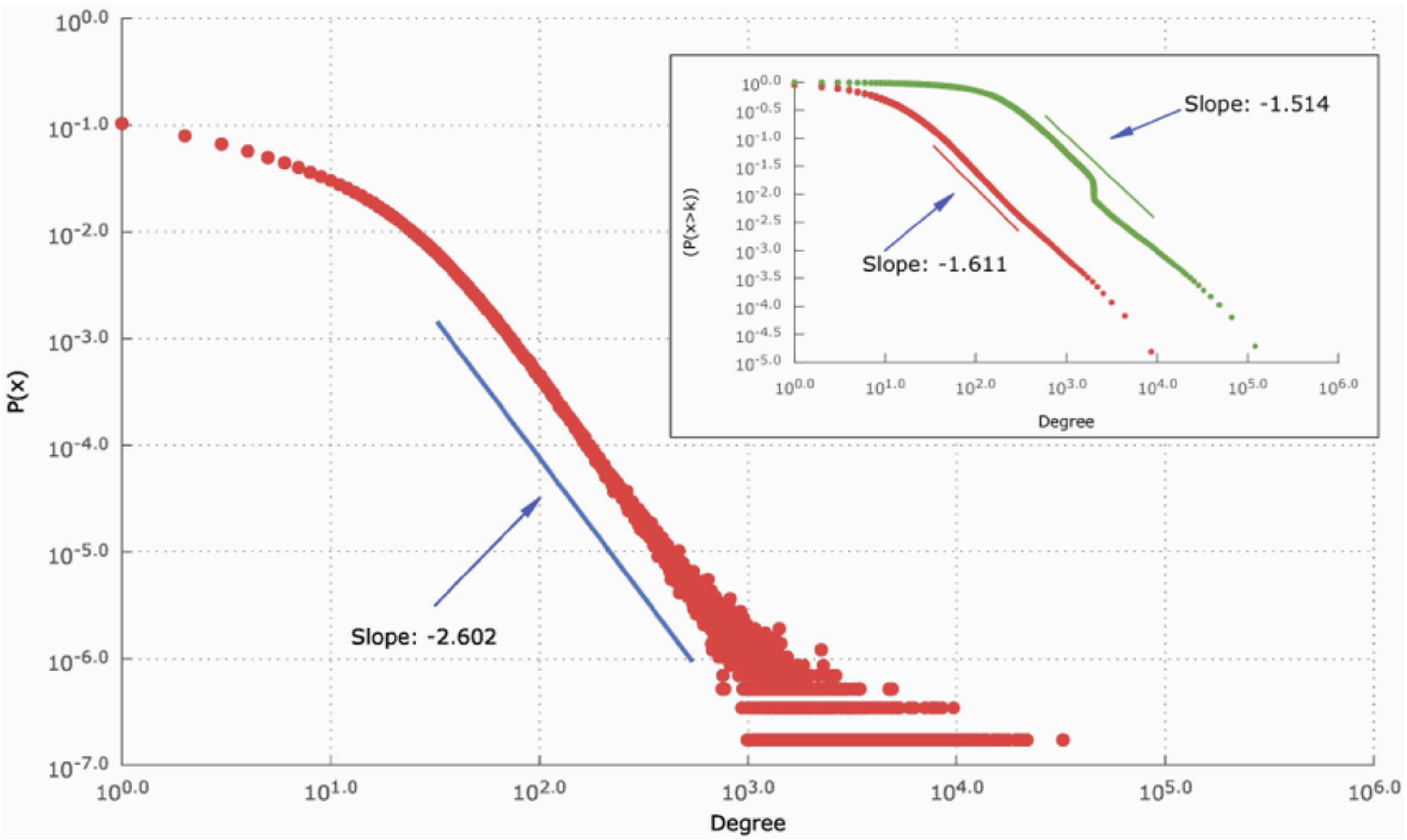
$$k_{\max} = 3$$

$$C_F(G) = \frac{\sum_{i=1}^{N} k_{\max} - k_i}{(N-1)(N-2)}$$

i	k_i	$k_{\max} - k_i$
1	3	0
2	3	0
3	2	1
4	1	2
5	1	2
6	2	1

$$C_F(G) = \frac{0+0+1+2+2+1}{5 \times 4} = \frac{3}{10}$$

(In-)Degree Centrality: Remarks



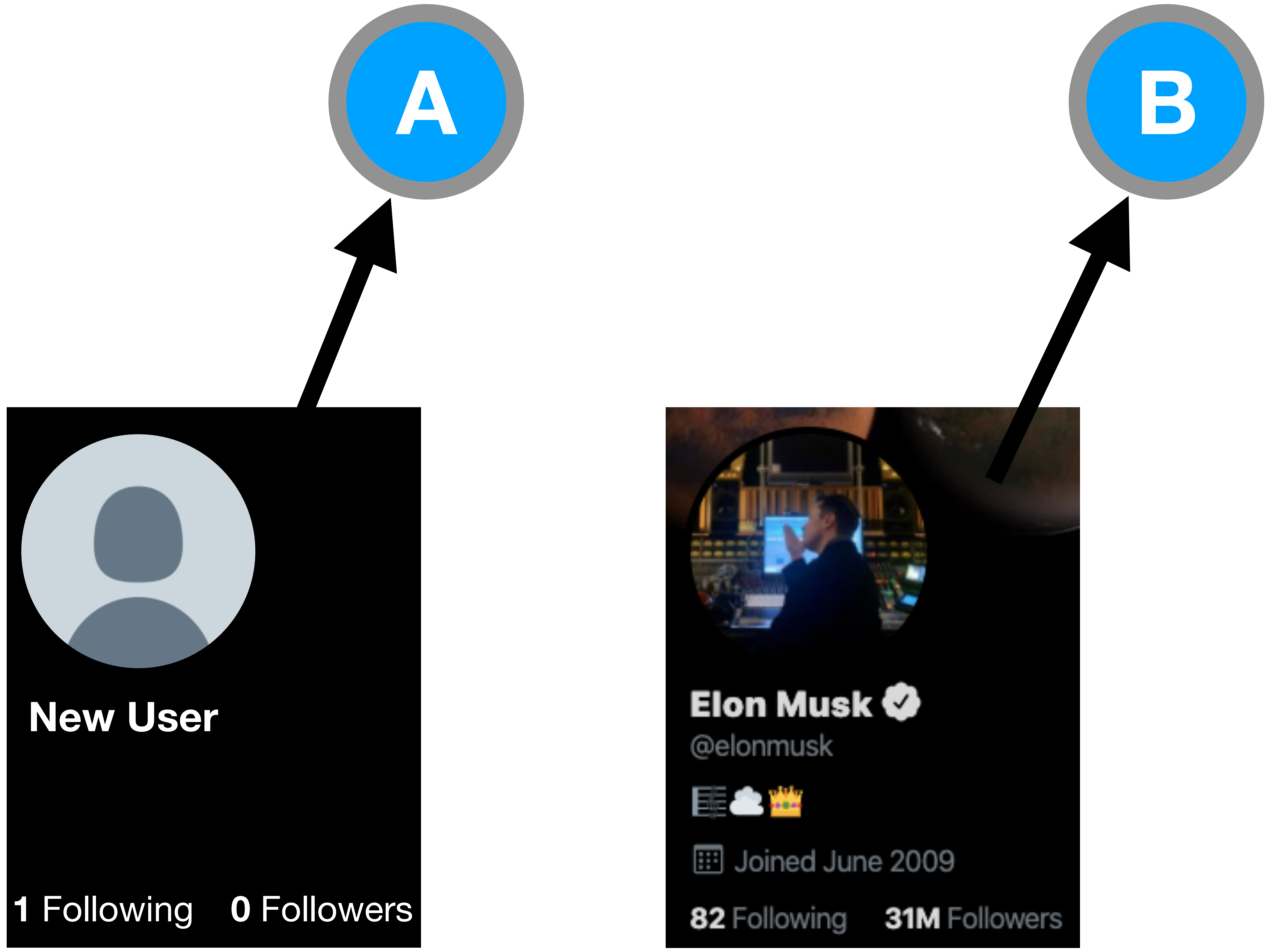
Twitter Follower Distribution

[Massive Social Network Analysis: Mining Twitter for Social Good, David Ediger et al (2010)]

Network (in-)degree distributions are often **heavy tailed**, with a **small number** of nodes having a **huge degree** but most having a **small degree**

Degree centrality **good for ranking top nodes** with extreme values, but **not very good** for ranking nodes in the **middle/low end**

(In-)Degree Centrality: Remarks



A and B have **SAME** degree centrality

Easy metric to **manipulate** in online social networks: buying followers, using bots etc

Gives **every link equal weight** — is this meaningful?

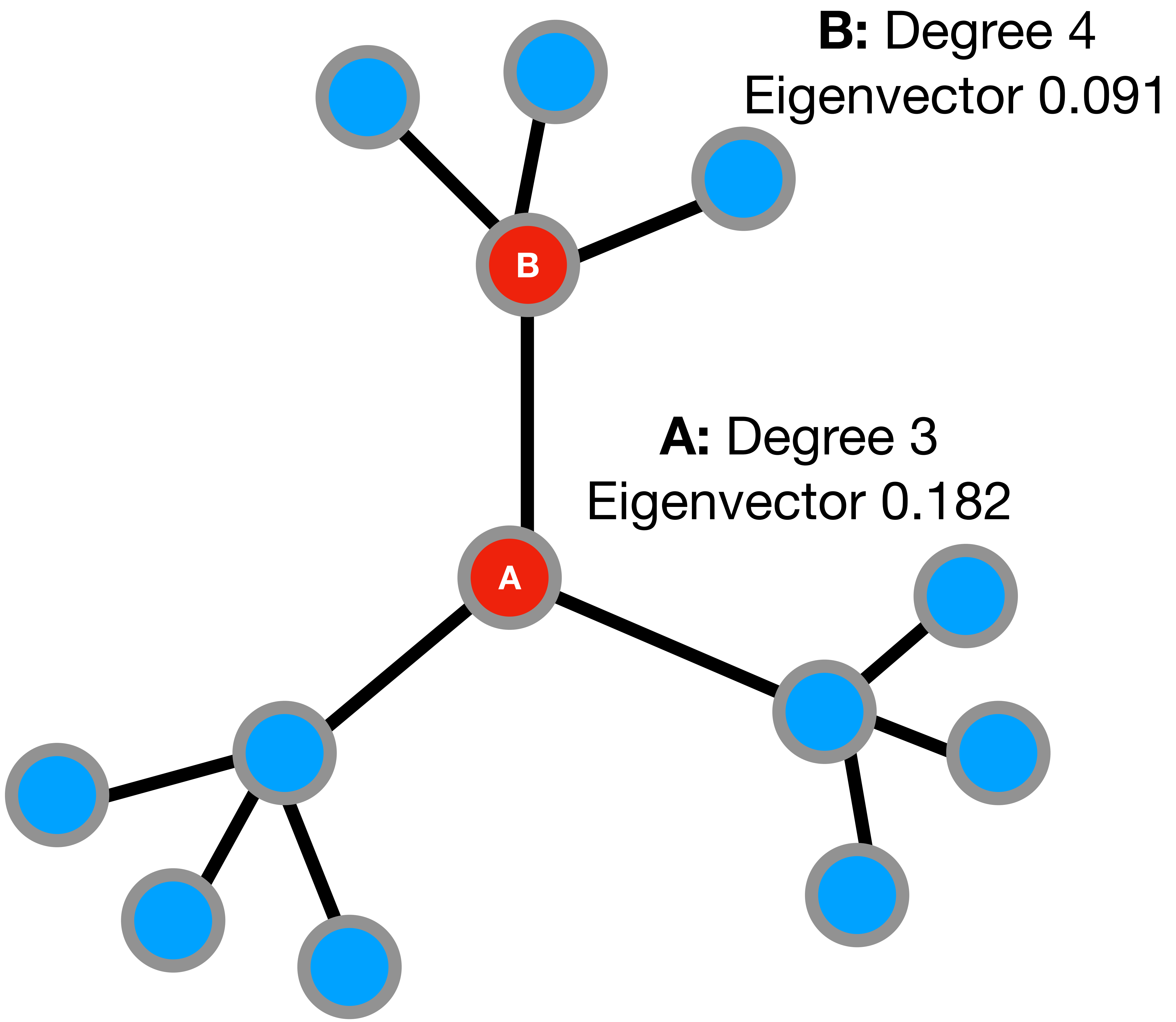
Eigenvector Centrality

Eigenvector centrality addresses **BOTH** of these issues

Each node’s centrality is **proportional** to the **sum** of it’s neighbours centrality

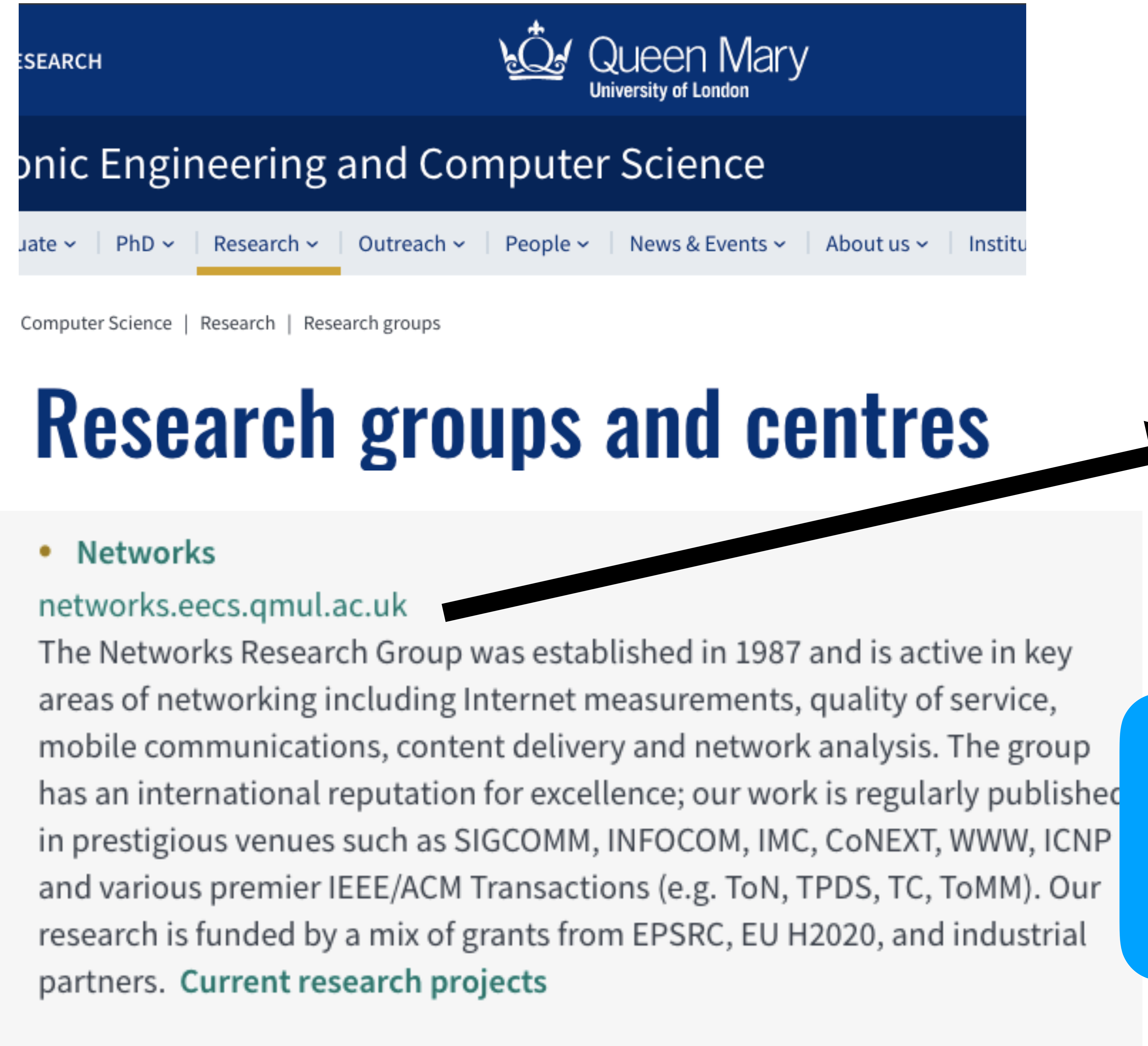
Node **B** has higher degree centrality **(4 vs 3)**

But **A** has higher eigenvector centrality



Google PageRank

Challenge: how to rank search engine results well, even in the middle?



ESEARCH

Queen Mary
University of London

onic Engineering and Computer Science

uate | PhD | Research | Outreach | People | News & Events | About us | Institu

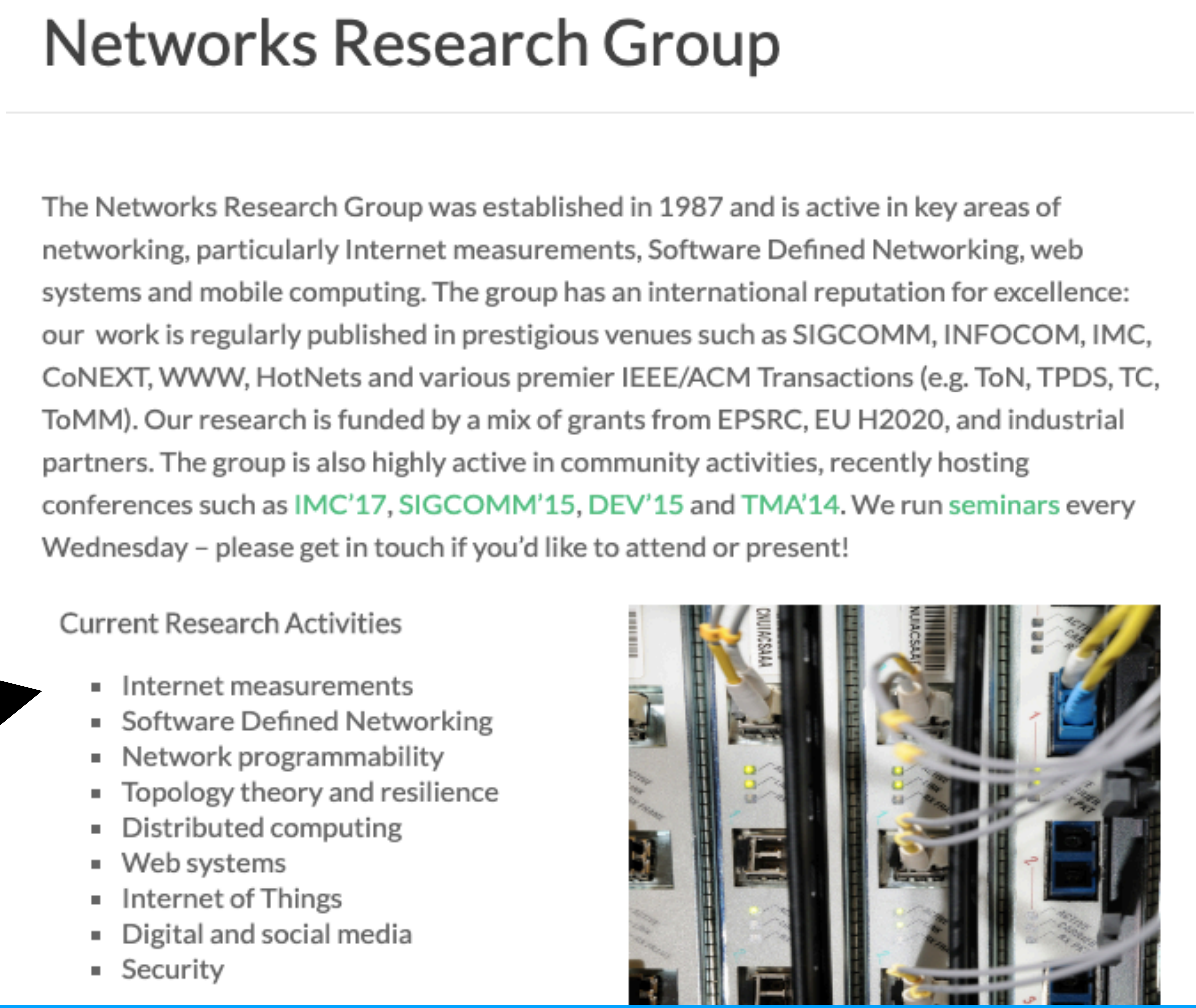
Computer Science | Research | Research groups

Research groups and centres

- Networks

networks.eecs.qmul.ac.uk

The Networks Research Group was established in 1987 and is active in key areas of networking including Internet measurements, quality of service, mobile communications, content delivery and network analysis. The group has an international reputation for excellence; our work is regularly published in prestigious venues such as SIGCOMM, INFOCOM, IMC, CoNEXT, WWW, ICNP and various premier IEEE/ACM Transactions (e.g. ToN, TPDS, TC, ToMM). Our research is funded by a mix of grants from EPSRC, EU H2020, and industrial partners. Current research projects



Networks Research Group

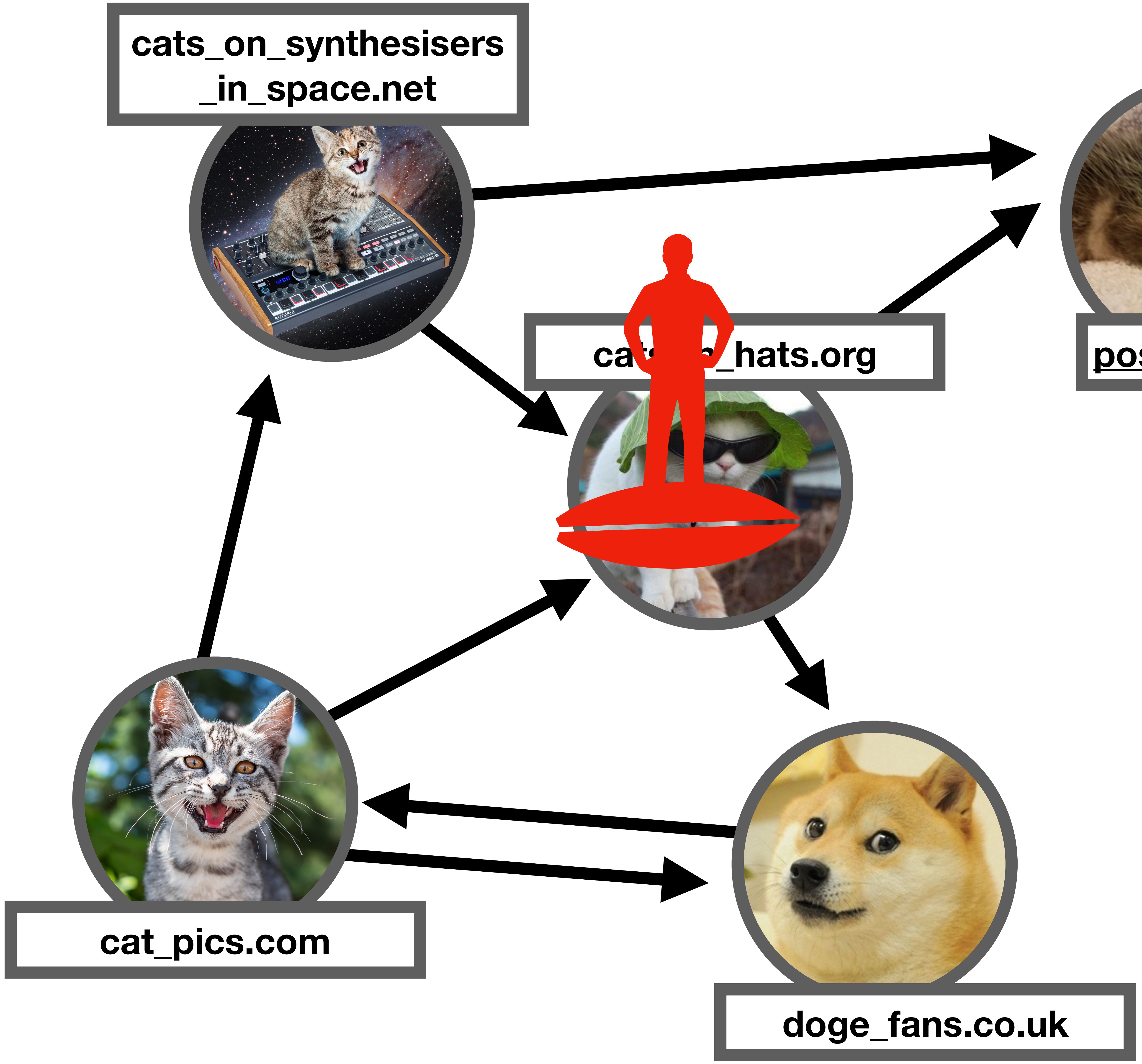
The Networks Research Group was established in 1987 and is active in key areas of networking, particularly Internet measurements, Software Defined Networking, web systems and mobile computing. The group has an international reputation for excellence: our work is regularly published in prestigious venues such as SIGCOMM, INFOCOM, IMC, CoNEXT, WWW, HotNets and various premier IEEE/ACM Transactions (e.g. ToN, TPDS, TC, ToMM). Our research is funded by a mix of grants from EPSRC, EU H2020, and industrial partners. The group is also highly active in community activities, recently hosting conferences such as IMC'17, SIGCOMM'15, DEV'15 and TMA'14. We run seminars every Wednesday – please get in touch if you'd like to attend or present!

Current Research Activities

- Internet measurements
- Software Defined Networking
- Network programmability
- Topology theory and resilience
- Distributed computing
- Web systems
- Internet of Things
- Digital and social media
- Security

View the Web as a directed graph with web pages as nodes and hyperlinks as links

PageRank idea: the random surfer



“Random surfer” navigates the web by clicking on hyperlinks.

With probability **p** they start over and pick a new webpage to start again

Centrality of node i: what is the probability of finding the surfer at node i if you check after a long time?

Properties of Eigenvector/PageRank

- More difficult metric to **‘cheat’** than degree
- Takes into account full **network structure** and can help **distinguish** nodes in the ‘middle’
- Still **not too difficult** for a computer to calculate (can be done in distributed way)

Questions so far?

Path-based metrics

Which nodes provide the most important **connectivity** or **reachability**?

Which nodes, if **removed**, would **damage** the network most?

Less about influence/popularity

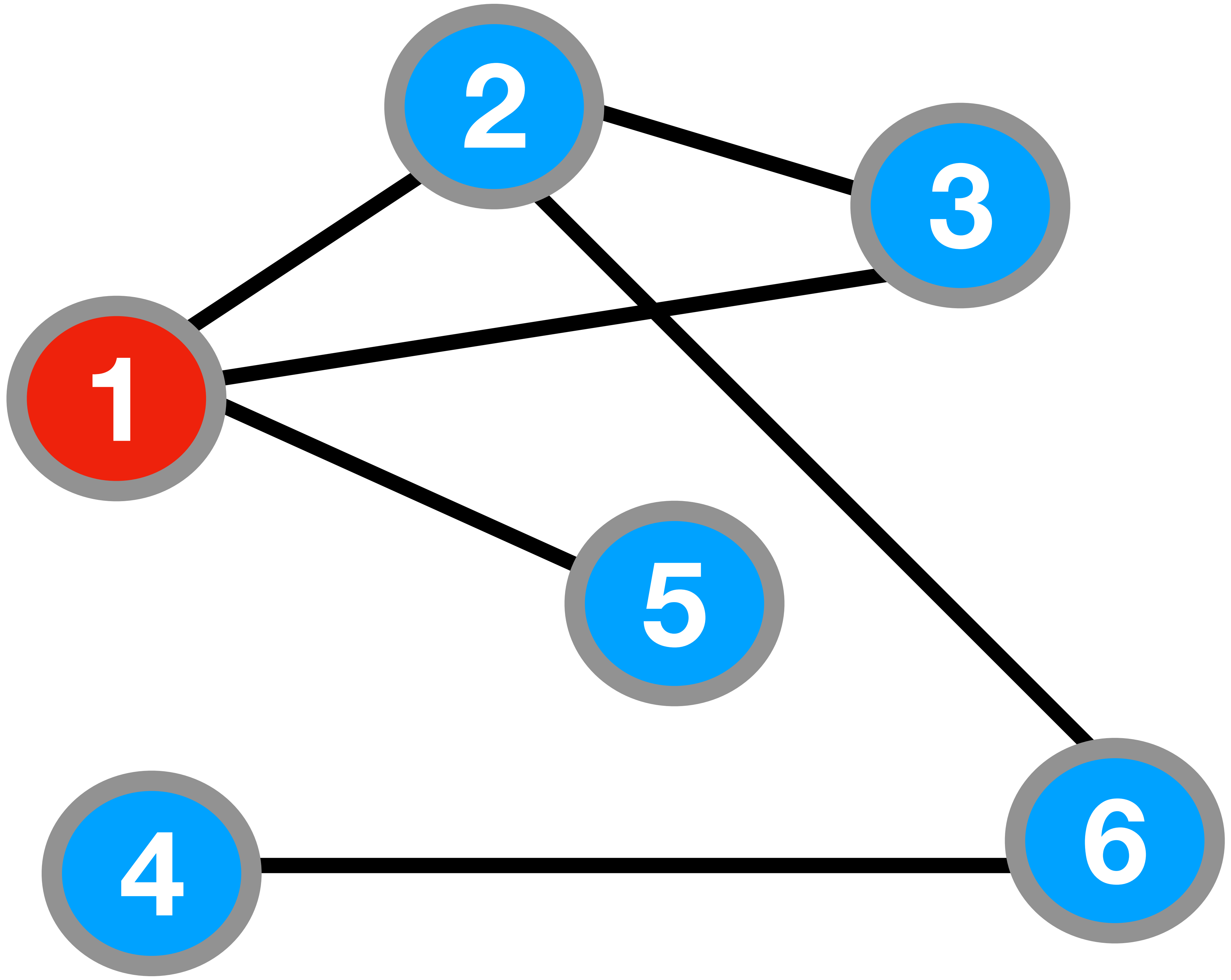
Closeness centrality

Which node is closest to everything else?

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

The **smaller** the distances, the **larger** the centrality value

Closeness centrality: example



j	d(1,j)
2	1
3	1
4	3
5	1
6	2

$$C_c(1) = [1 + 1 + 3 + 1 + 2]^{-1} = \frac{1}{8}$$

Closeness centrality: properties

- Usually has **small range** because of short path lengths
- Can be **unstable** — adding or removing a link can dramatically change who is closest
- Path lengths are **expensive** to compute

Betweenness centrality

Which node(s) are most vital for maintaining connectivity?

Number of s**hortest path** routes from **j to k** that go through **i**

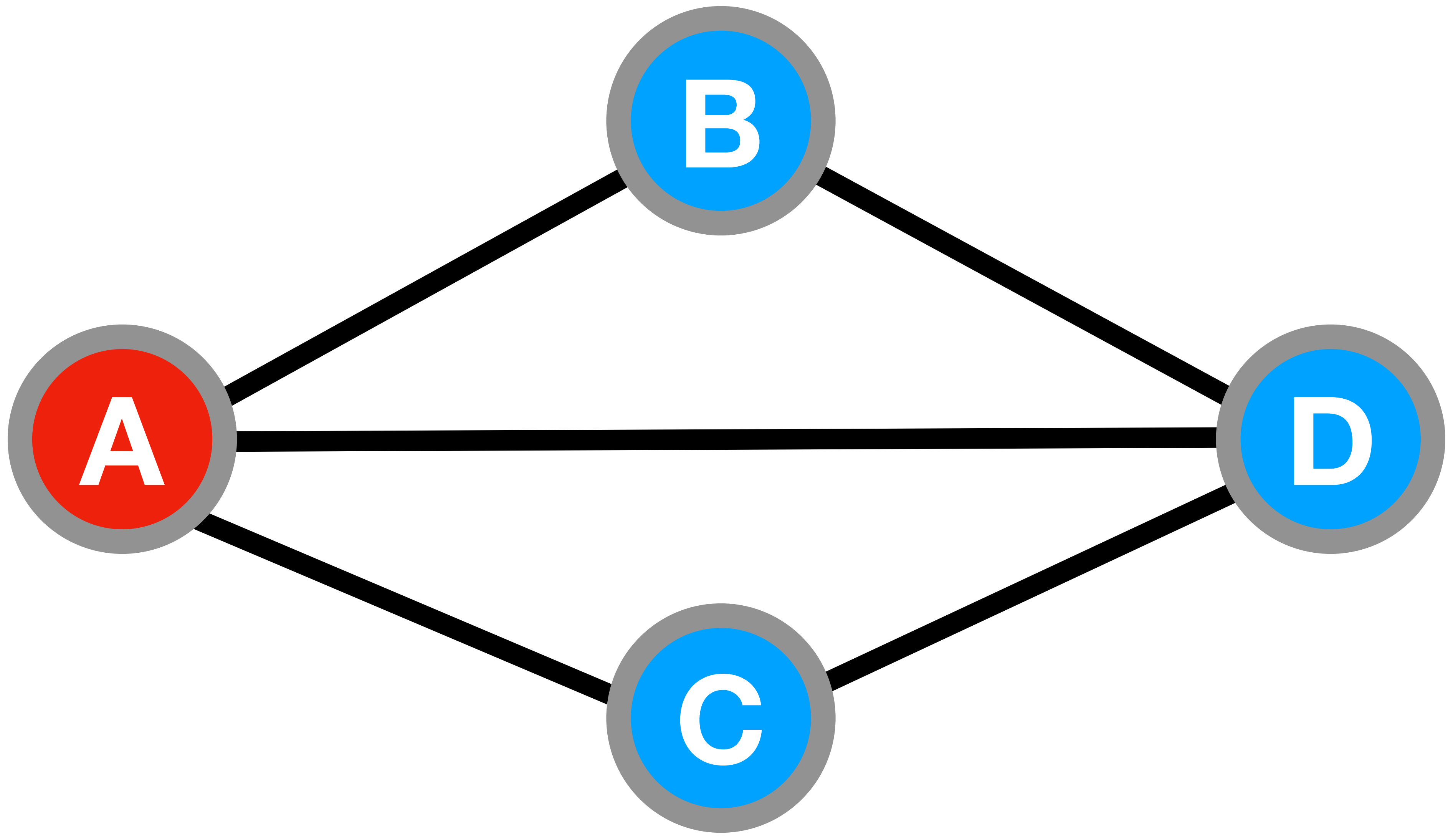
$$C_B(i) = \sum_{j \neq i \neq k} \boxed{g_{jk}(i)} / \boxed{g_{jk}}$$

Total number of shortest paths from j to k

Sum of all paths apart from those starting or ending at i

Betweenness: Example

$$C_B(i) = \sum_{j \neq i \neq k} g_{jk}(i)/g_{jk}$$



Betweenness centrality of A?

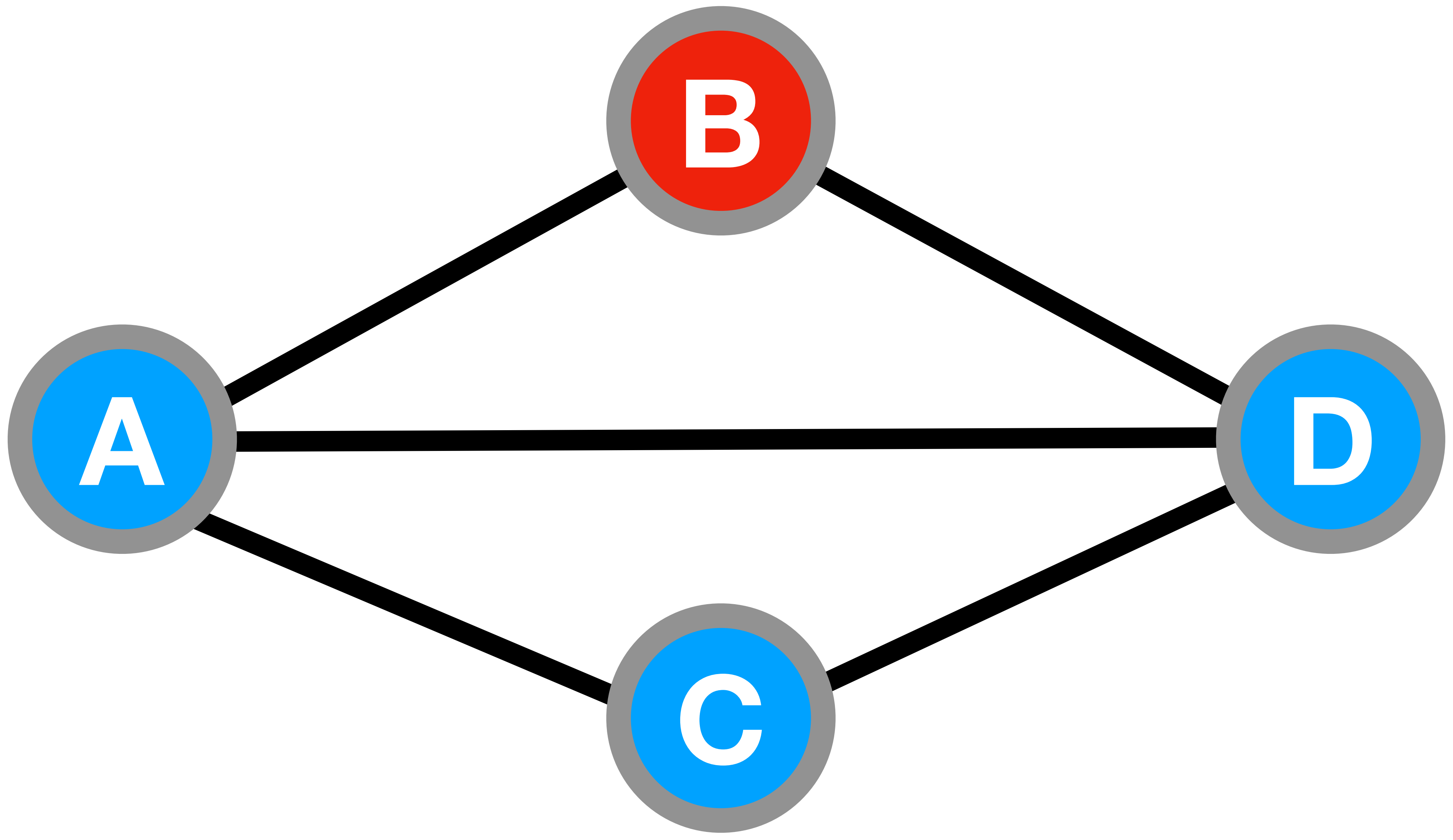
$$g_{BC}(A) = 1$$ (B-> A -> C)

$$g_{BC} = 2$$ (B-> A -> C) (B-> D -> C)

$$g_{DC}(A) = 0, \; g_{BD}(A) = 0$$

$$C_B(A) = 1/2$$

Betweenness: Example



$$C_B(i) = \sum_{j \neq i \neq k} g_{jk}(i)/g_{jk}$$

Betweenness centrality of B?

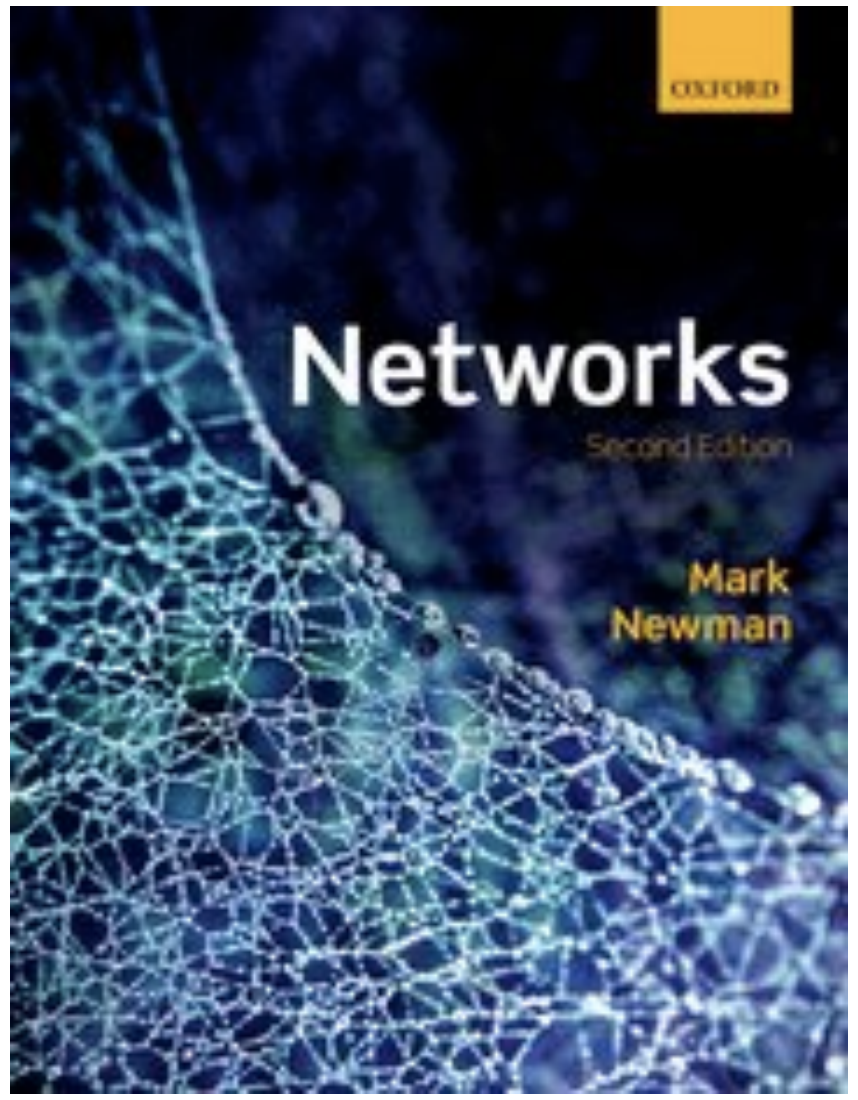
No shortest paths between A, C and D that go through B, so B has betweenness 0.

Betweenness centrality: properties

- Can help identify **vital nodes** for maintaining connectivity
- Used in **Internet monitoring** where (care most about packets getting to destination)
- Finding shortest paths is **expensive** to compute

The more detailed the network, the more complex centrality measures you can devise!

Detailed e.g. directed, weighted, …



See **Networks: an Introduction** by **Mark Newman** for a deeper dive into different centrality measures

FIFA Data Analysis: Jupyter Notebook

[A public data set of spatio-temporal match events in soccer competitions, Luca Pappalardo et al 2018, Nature]

https://www.nature.com/articles/s41597-019-0247-7#Sec9