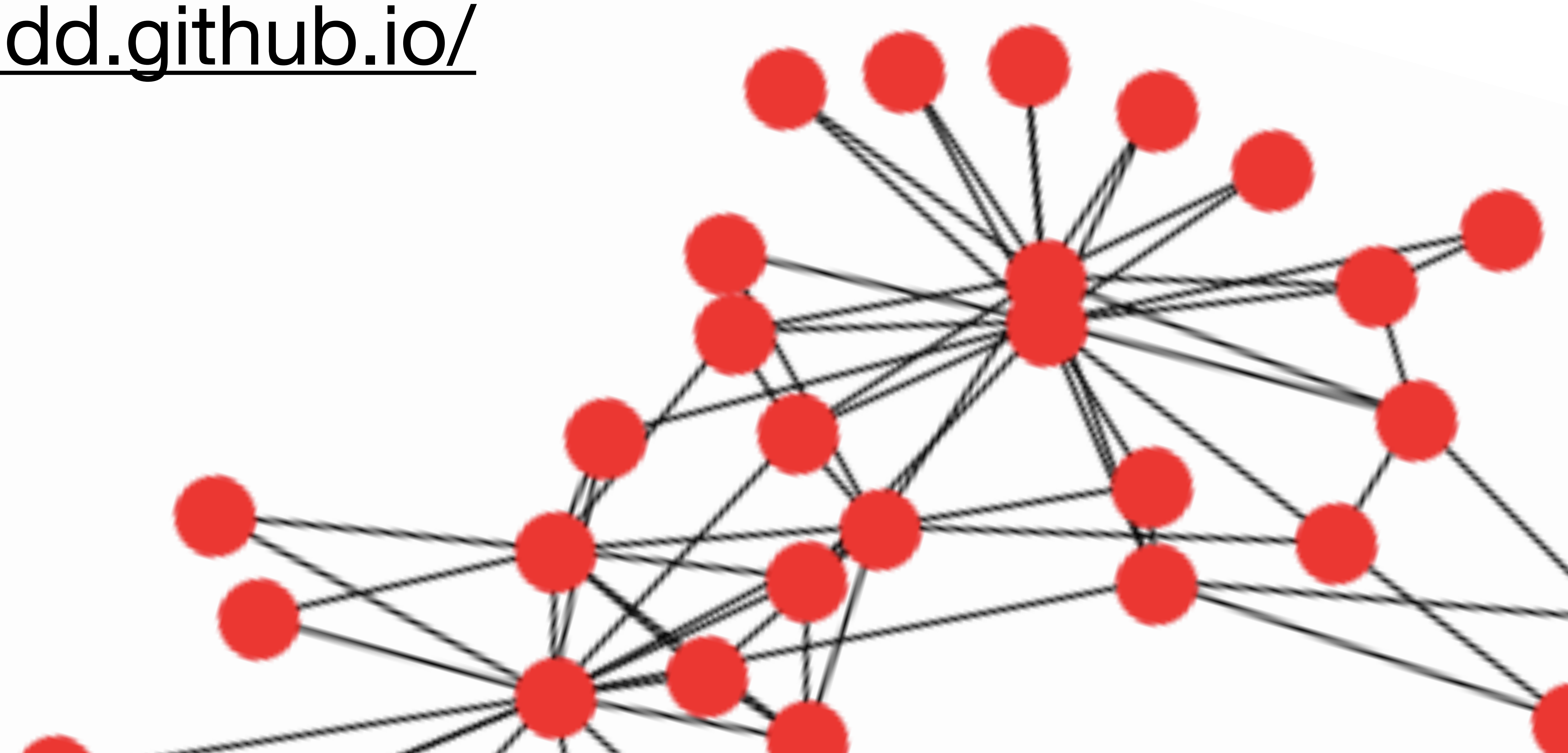
DMSN Tutorial 1: Networks and Random Graphs

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



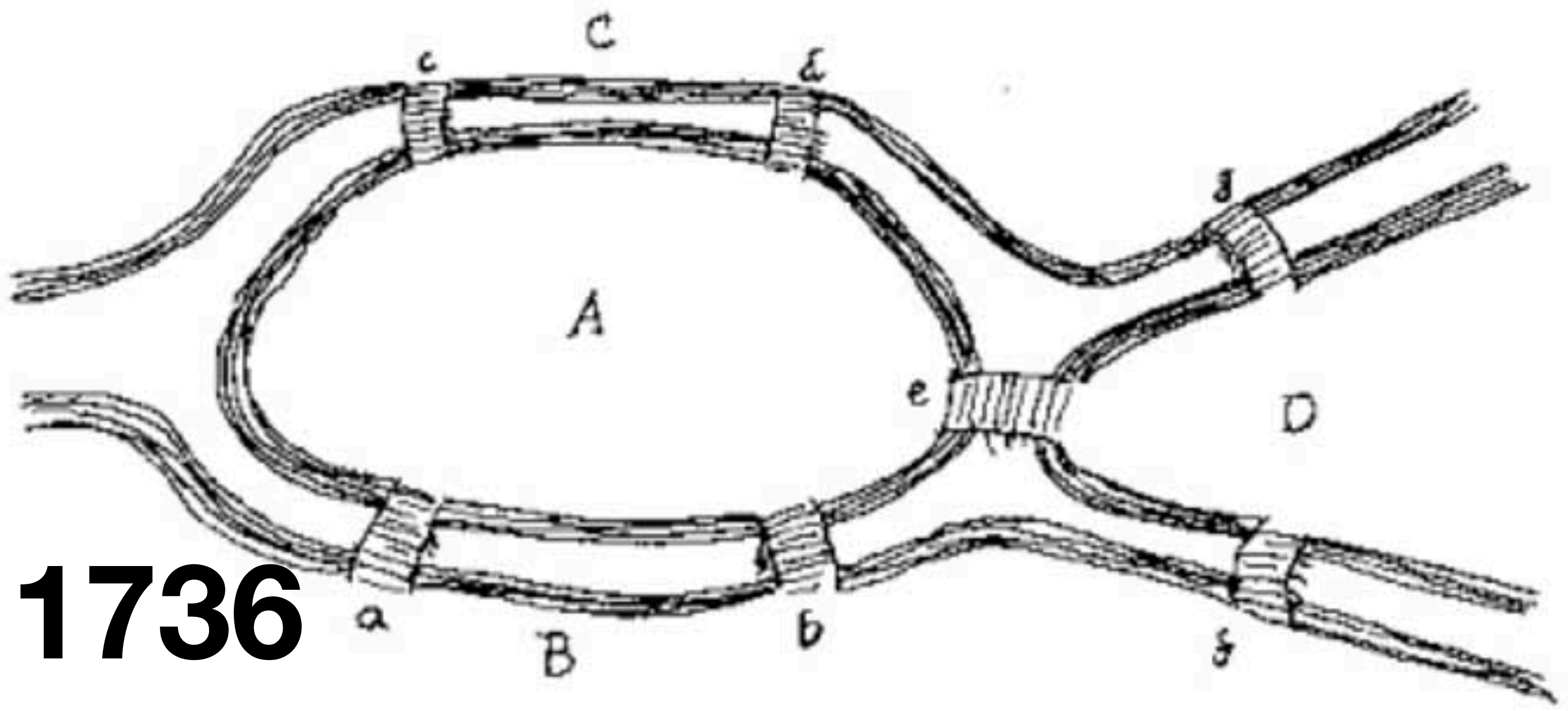
Fully funded 3.5 year industrial PhD in complex networks

- Topic: Equality and social mobility in networks
- Are you interested in studying networks?
- Are you interested in equality?
- QMUL has a 3.5 year PhD fully funded (UK/EU applicants only).
- Funded by Moogsoft, includes fees and stipend.
- Closing date February 15th
- Find out more:
  - Contact supervisor: Richard Clegg
  - Email r.clegg@qmul.ac.uk

In this tutorial:

- **Recap** on concepts and metrics covered in the lecture
- **Get to grips with** the Erdos-Renyi random graph model
- **See** some of the key similarities and differences between random graphs and real networks

A (very) brief history of network science



1736

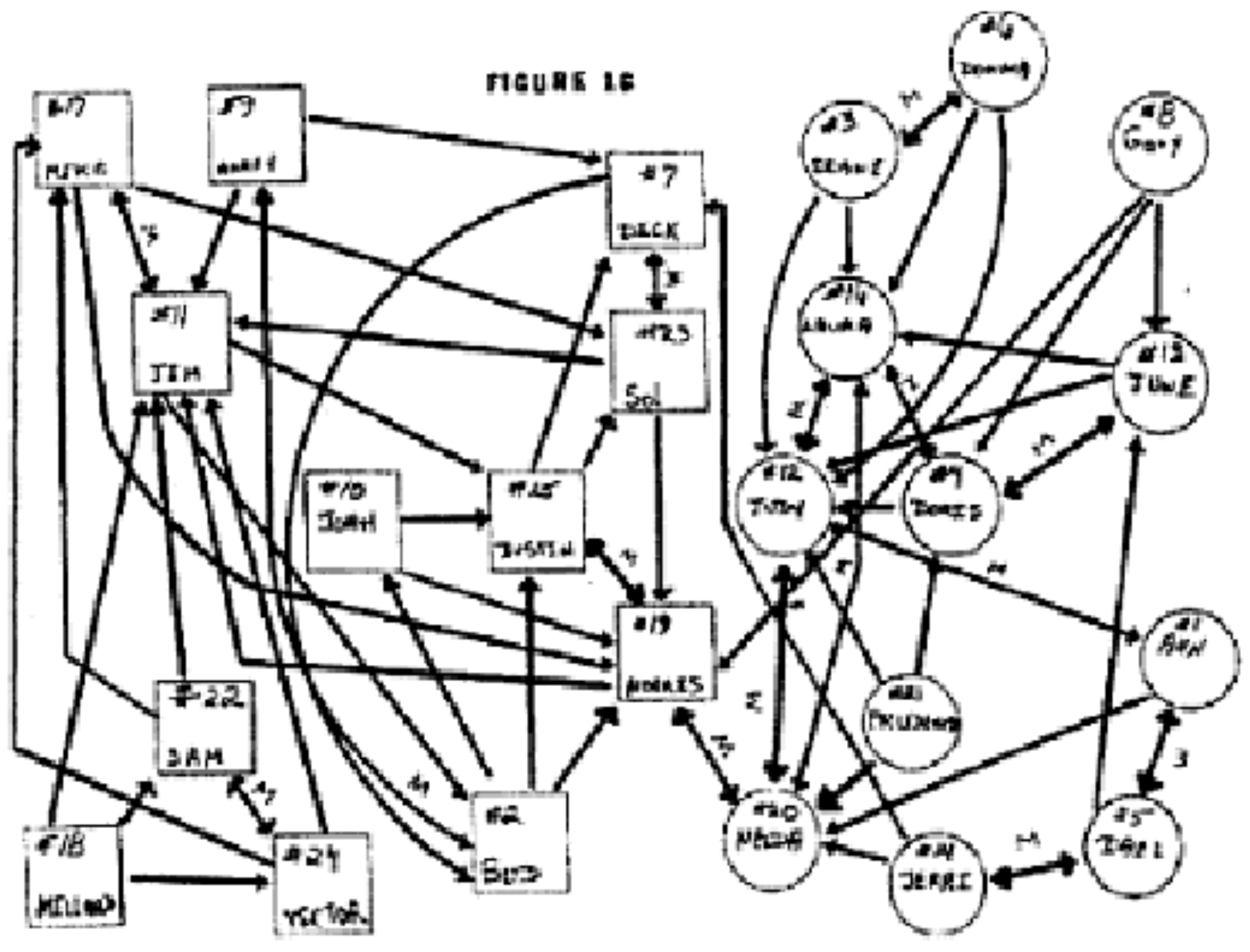
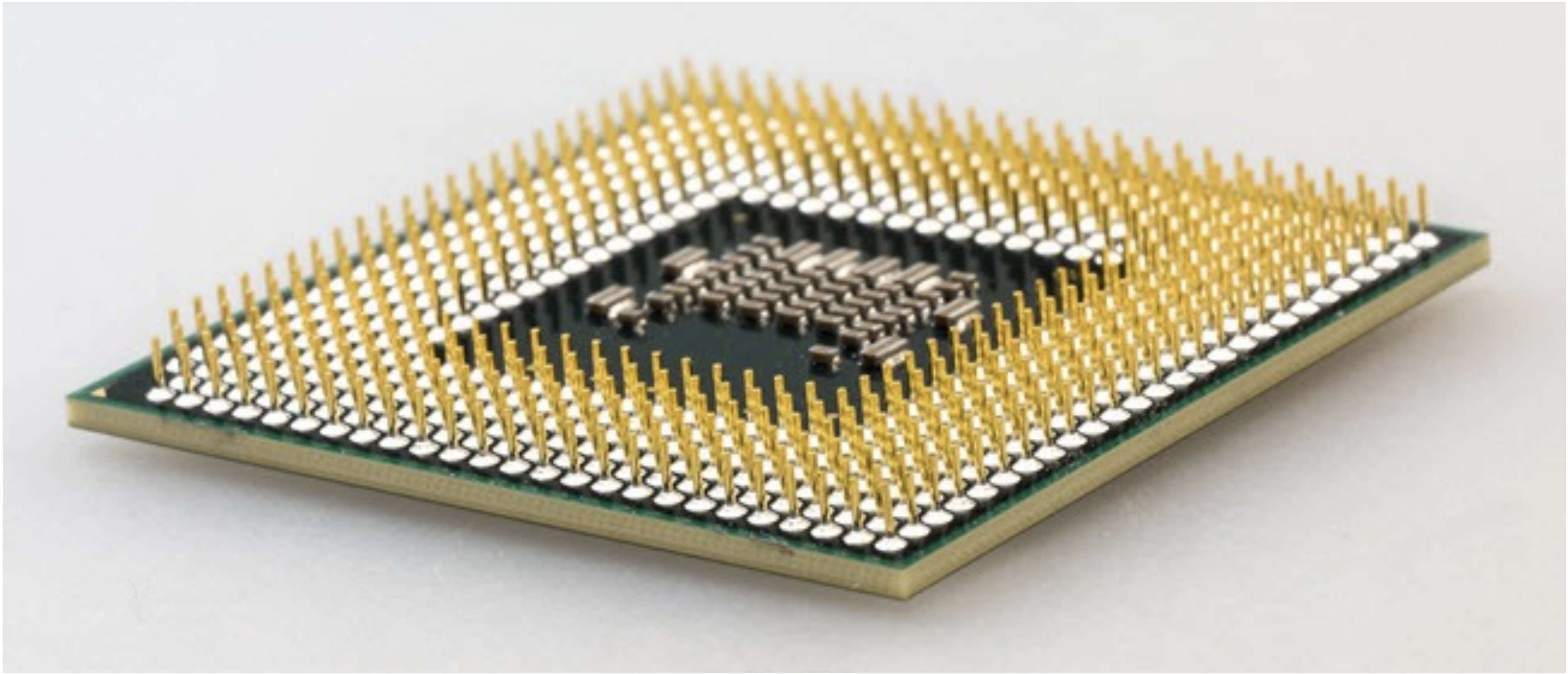


Figure 3.3 Moreno's 1933 Sociogram

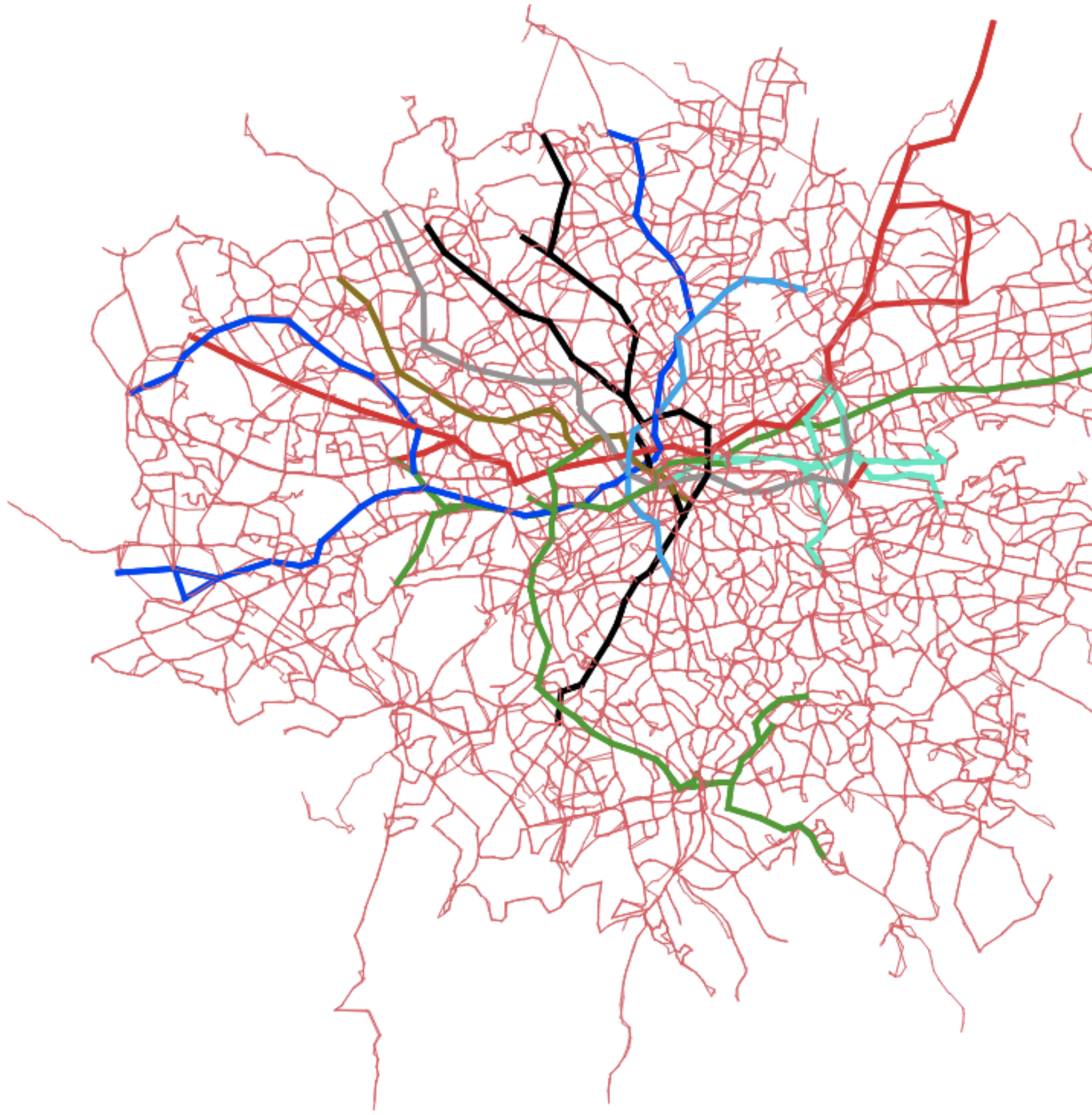
1933



Availability of rich datasets

+

Computing power



If you could draw one edge per second and didn't take breaks, it would take 12,600 years to draw the Facebook graph

Network Science is Interdisciplinary

- **Social sciences:** made first use of ‘sociograms’ as networks, and drive a lot of the motivation for network science
- **Mathematics/Physics:** development of graph theory, models for dynamics on/of networks (often using theory from particle physics!)
- **Computer Science:** developing and implementing algorithms for networks, working with scalability challenges of big data
- **Field specific applications:** epidemiologists studying disease prevention/vaccination, Internet network operators, social network

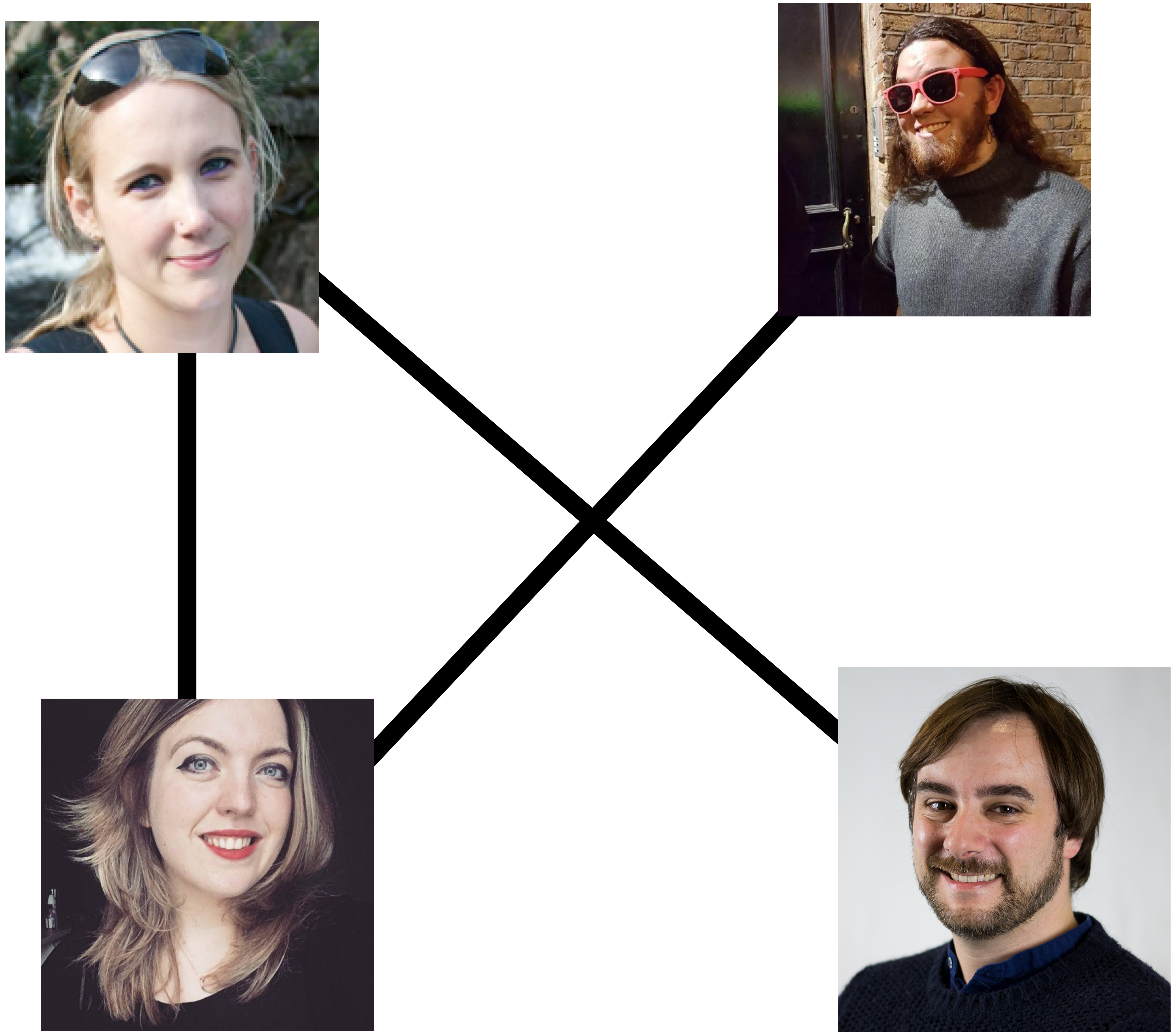
(Undirected) Graph

A **graph** is a tuple **(V,E)** of a set **V** of vertices and **E** of edges

Vertex (node) set: {Laurissa, Ben, Naomi, Mathieu}

Edge (link) set: { (Laurissa, Naomi), (Laurissa, Mathieu), (Naomi, Ben)}

Here, order doesn't matter as graph is **undirected**

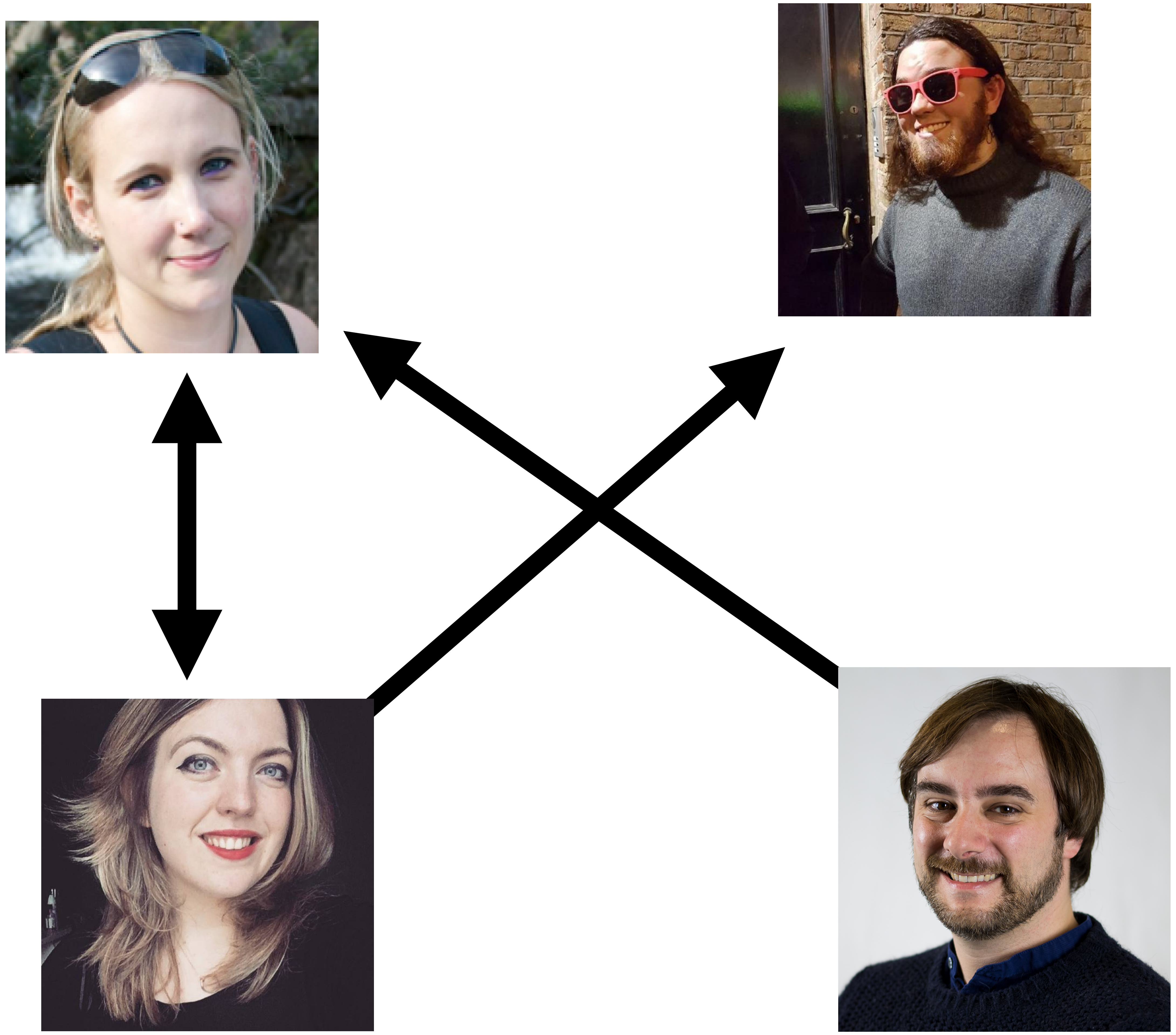


Directed Graph

Vertex (node) set: {Laurissa, Ben, Naomi, Mathieu}

Edge (link) set: { (Laurissa, Naomi), (Naomi, Laurissa) (Mathieu, Laurissa), (Naomi, Ben)}

Here, order **does** matter as graph is **directed**



How do we measure graphs?
How do we compare them?

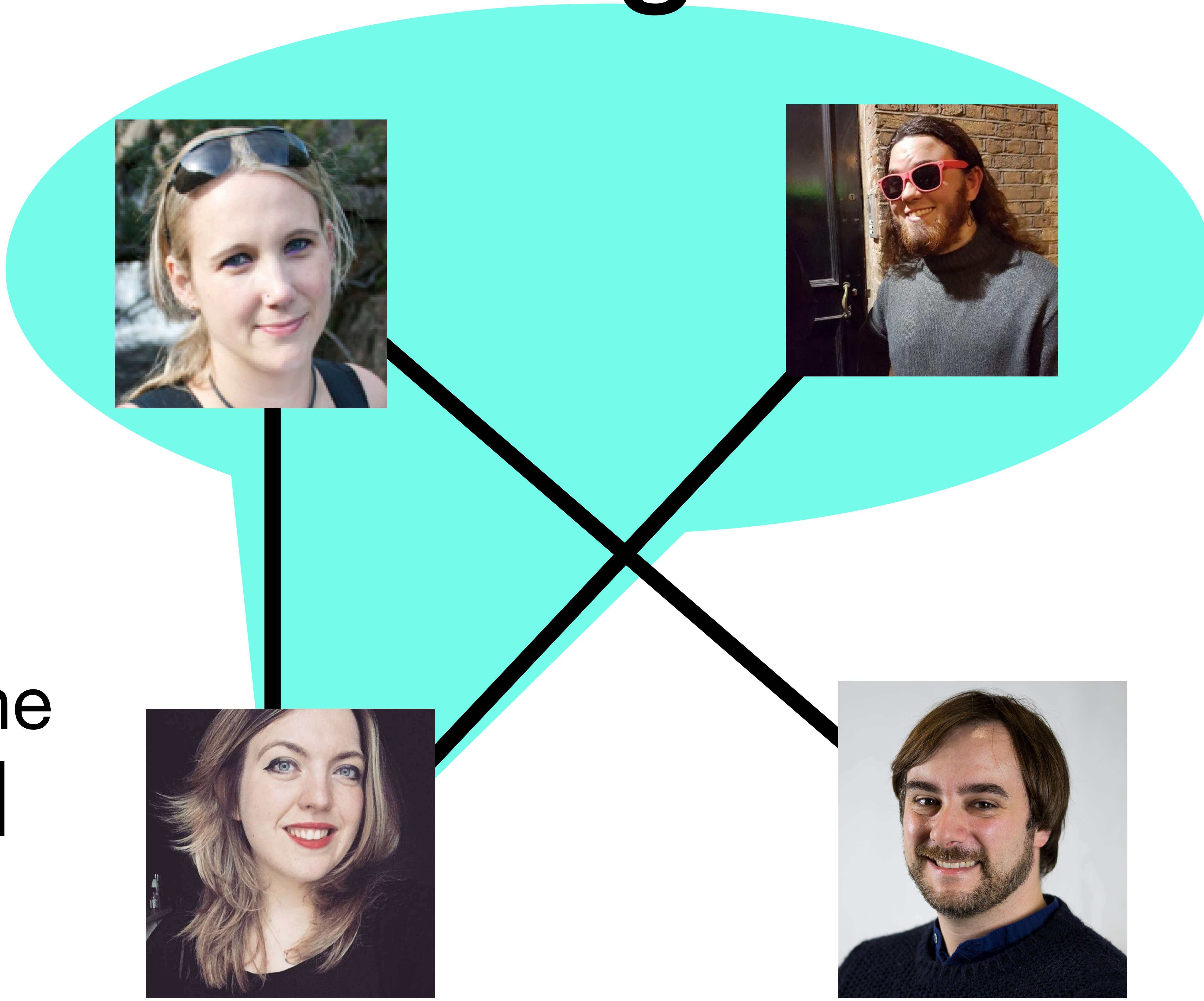
Neighbourhood and Degree

The **neighbourhood** N(v) of a vertex v is the set of vertices adjacent to V

e.g. N(Naomi) = {Laurissa, Ben}

The **degree** k(v) of a vertex v is the size of the neighbourhood: |N(v)|

e.g. k(Naomi) =2



Degree Sequence/Average Degree

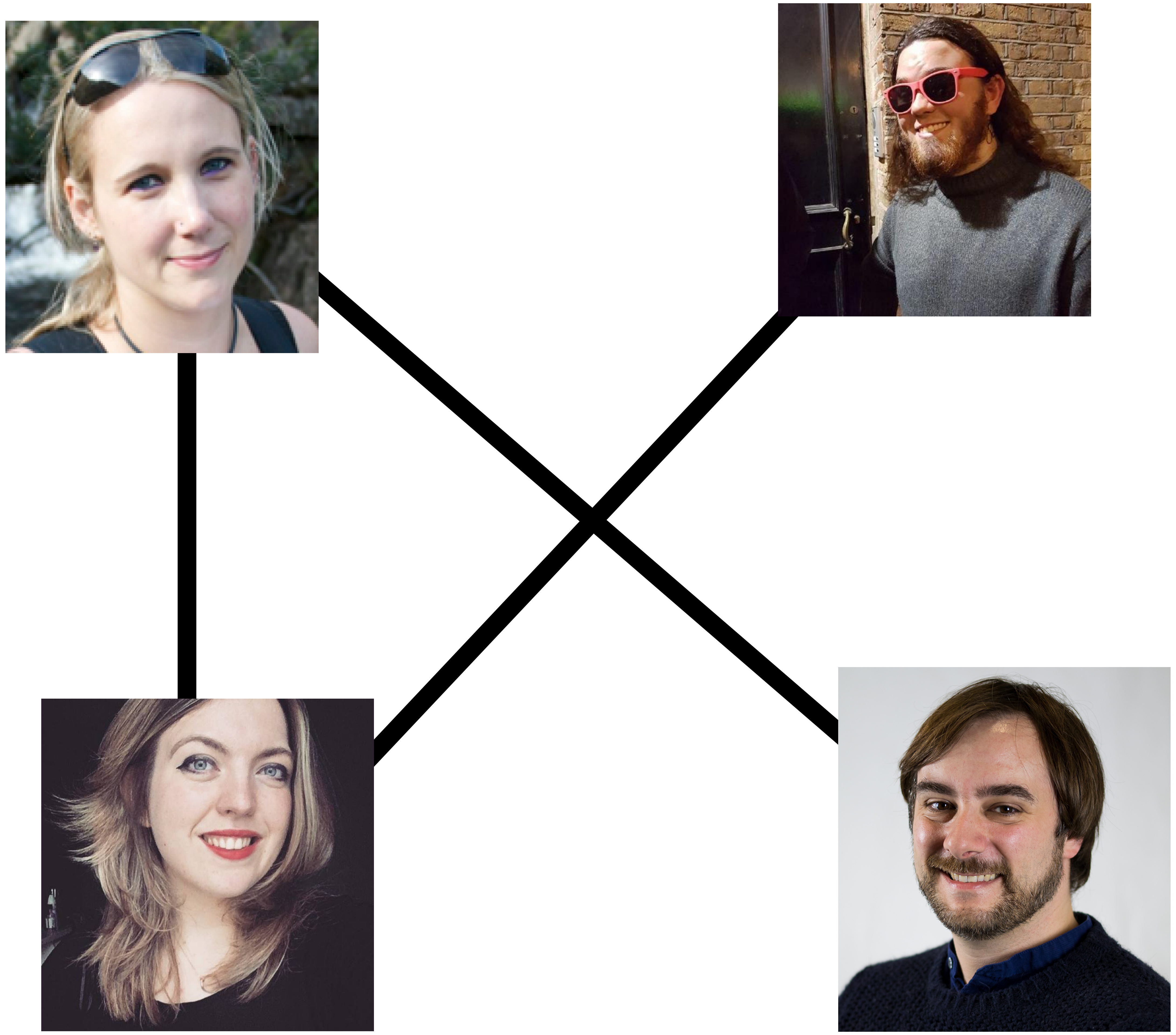
The **degree sequence** of a graph is the list of the vertex degrees for that graph (in decreasing order)

e.g. 2, 2, 1, 1

The **average degree** of a graph **<k>** is the mean of the node degrees

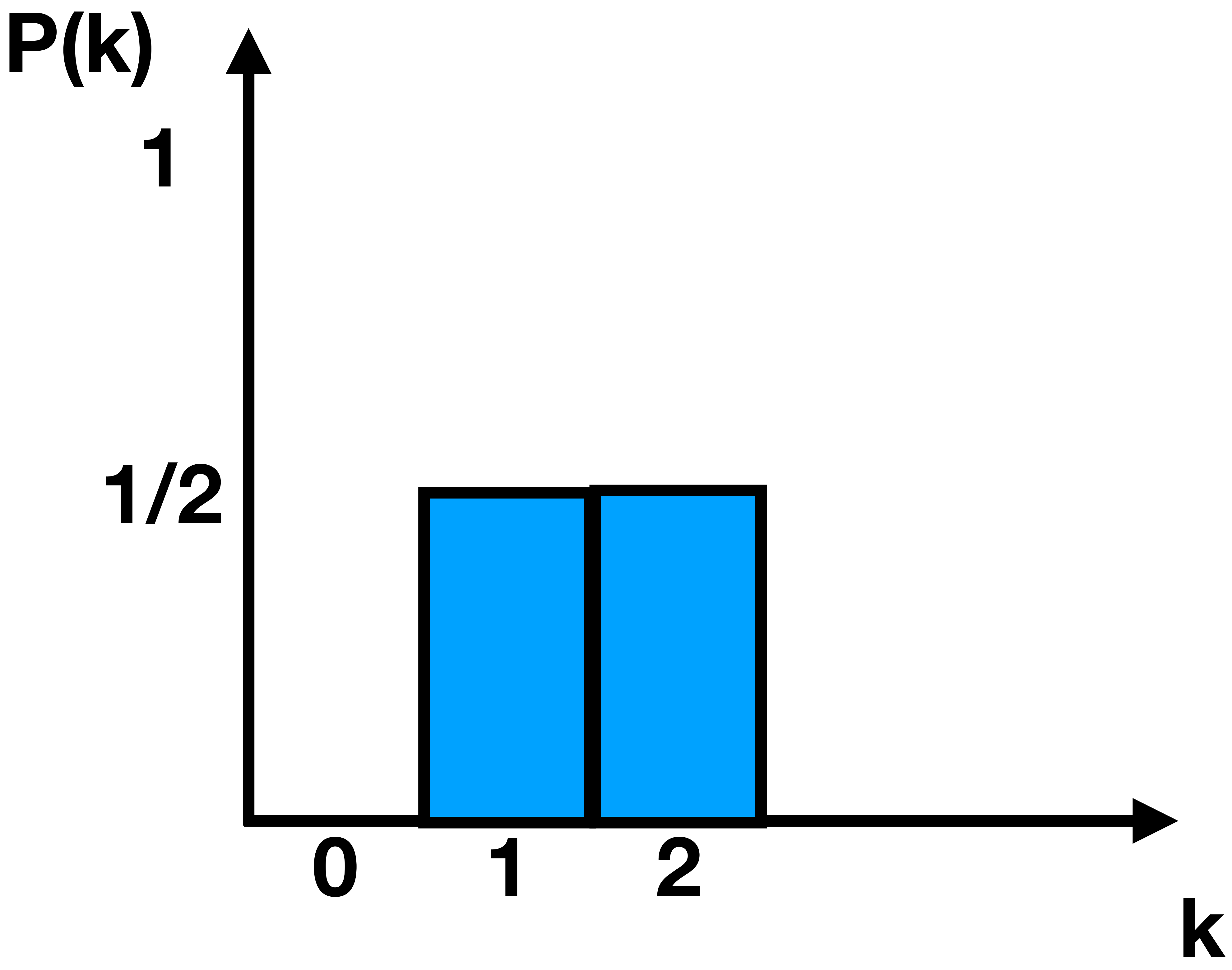
e.g. <k> = (2 + 2 + 1 + 1)/4 = 1.5

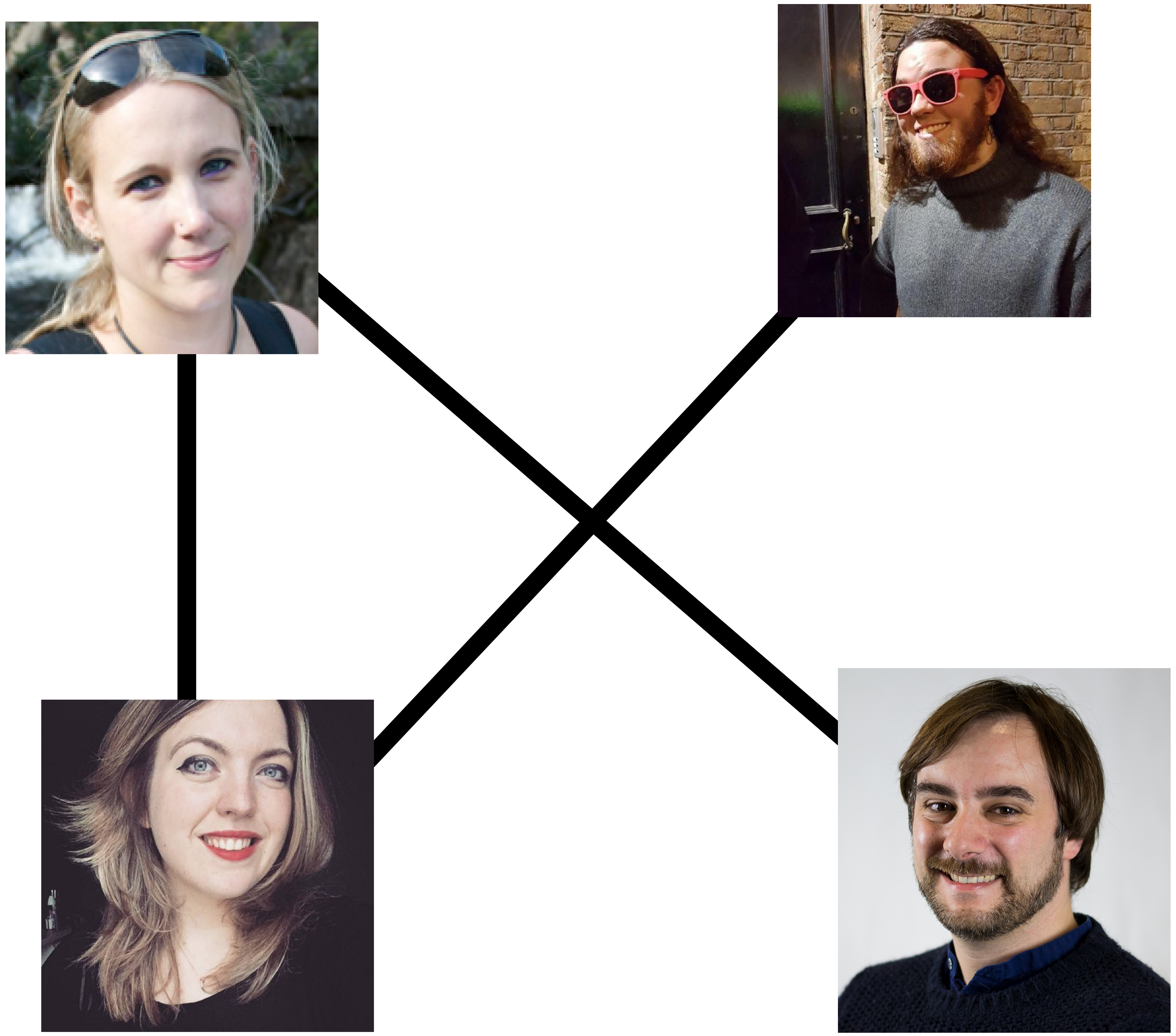
(also equal to 2*|edges|/|nodes| … why?)



Degree distribution

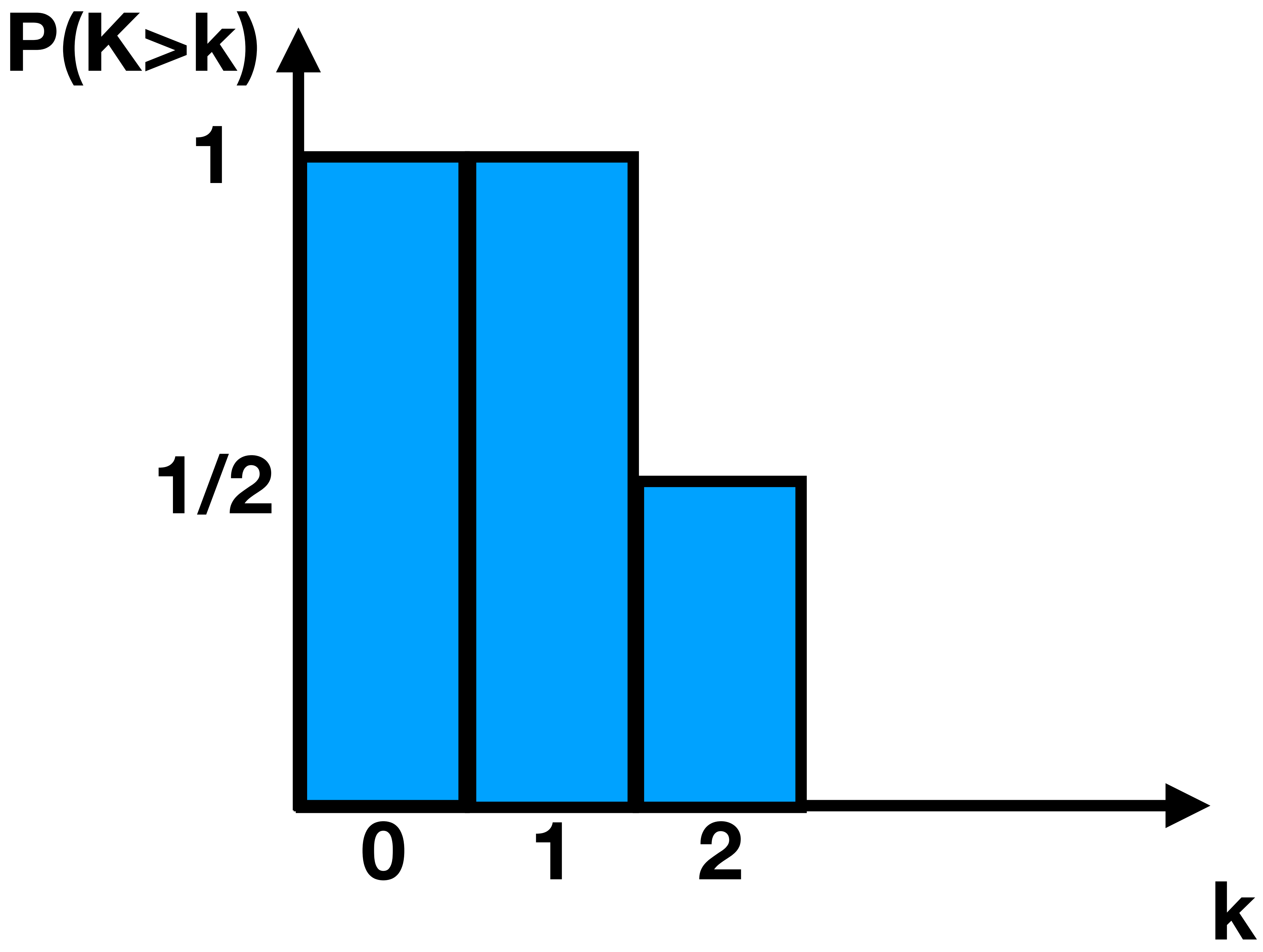
The degree distribution **P(k)** is the proportion of nodes with degree **equal to k**

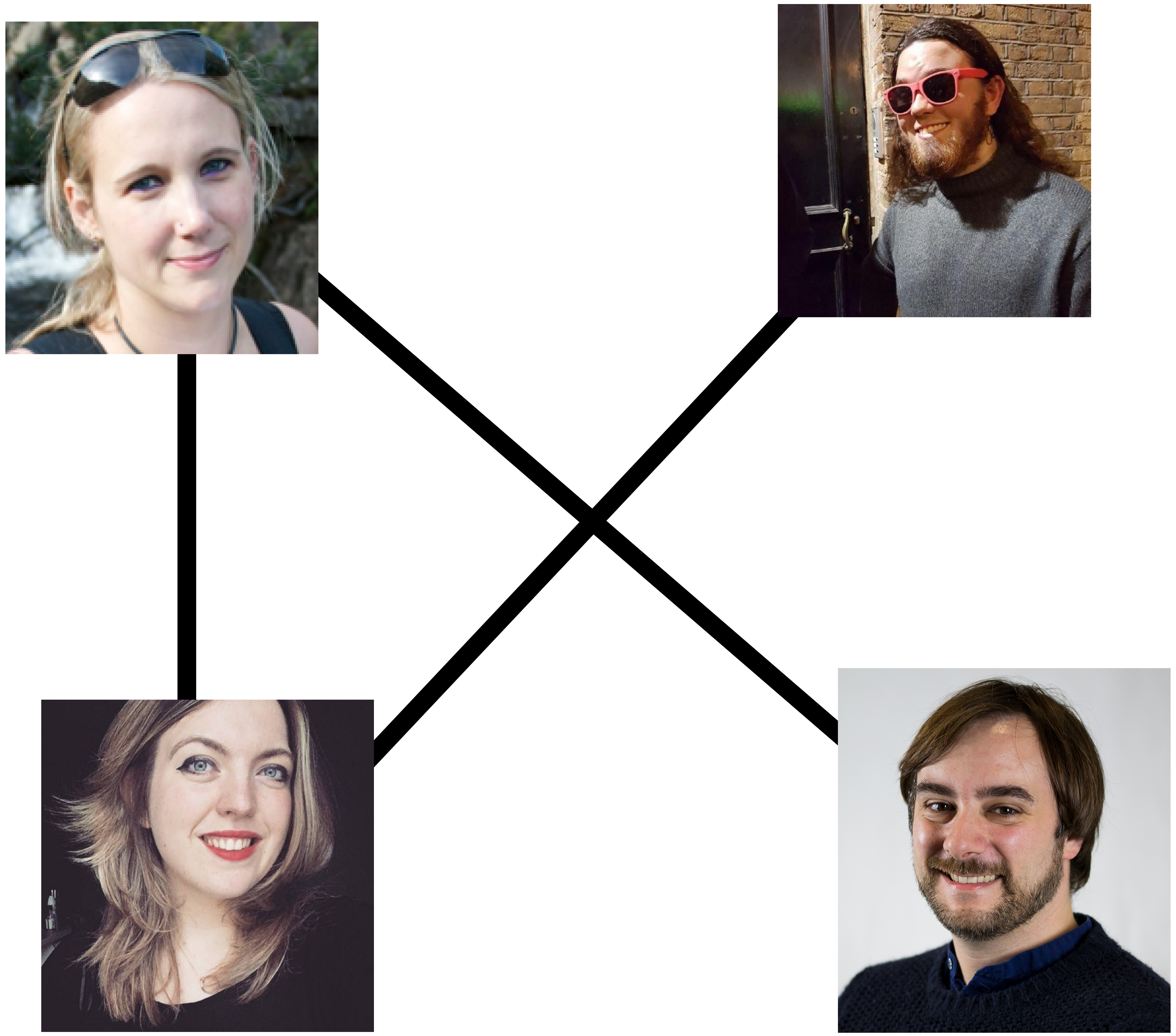




Degree distribution

… but it's common to look at the proportion of nodes with degree **greater than or equal to k**





Clustering Coefficient

Proportion of possible interconnections between neighbours

Node clustering coefficient C(v)

$$C(v) = \frac{|\{(u, w) \mid u, w \in N(v)\}|}{\frac{1}{2}k(v)(k(v) - 1)}$$

Pairs of neighbours of v that are connected

Possible pairs of v's neighbours, "k(v) choose 2"

Special case: if k(v) = 1 or 0, C(v) = 0

Clustering Coefficient

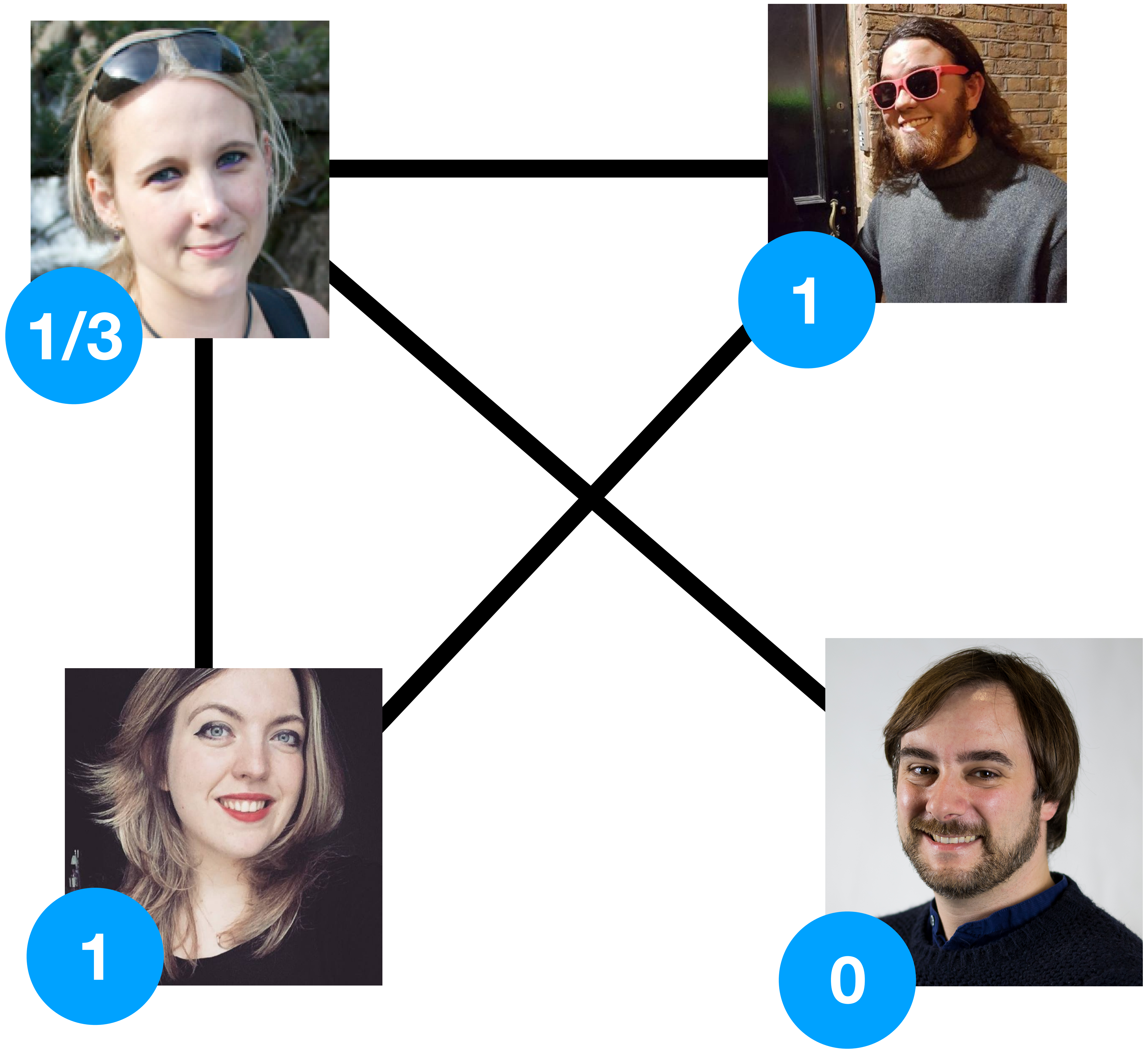
What is Laurissa's clustering coefficient?

Denominator: Laurissa's degree is 3, so 0.5*3*2 = <u>3</u>

Numerator: Only <u>one</u> pair of Laurissa's neighbours are connected (Naomi, Ben)

So C(Laurissa) = <u>1/3</u>

Average clustering C(G) = 7/12



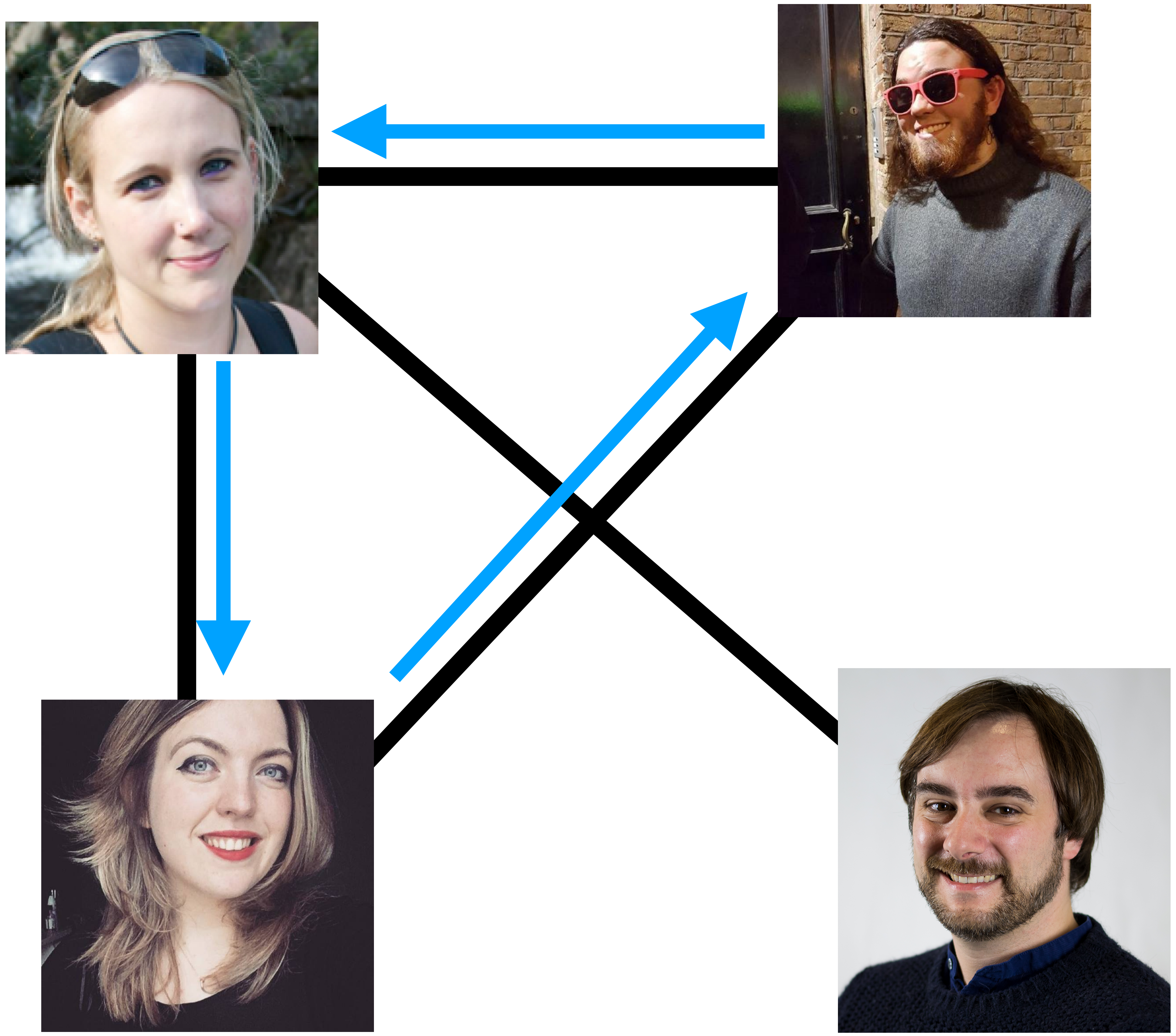
Paths and Cycles

A **path** is a sequence of nodes where each consecutive pair of nodes is linked by an edge

Ben, Laurissa, Naomi

A **cycle** is a path where the start node is also the end node

Ben, Laurissa, Naomi, Ben



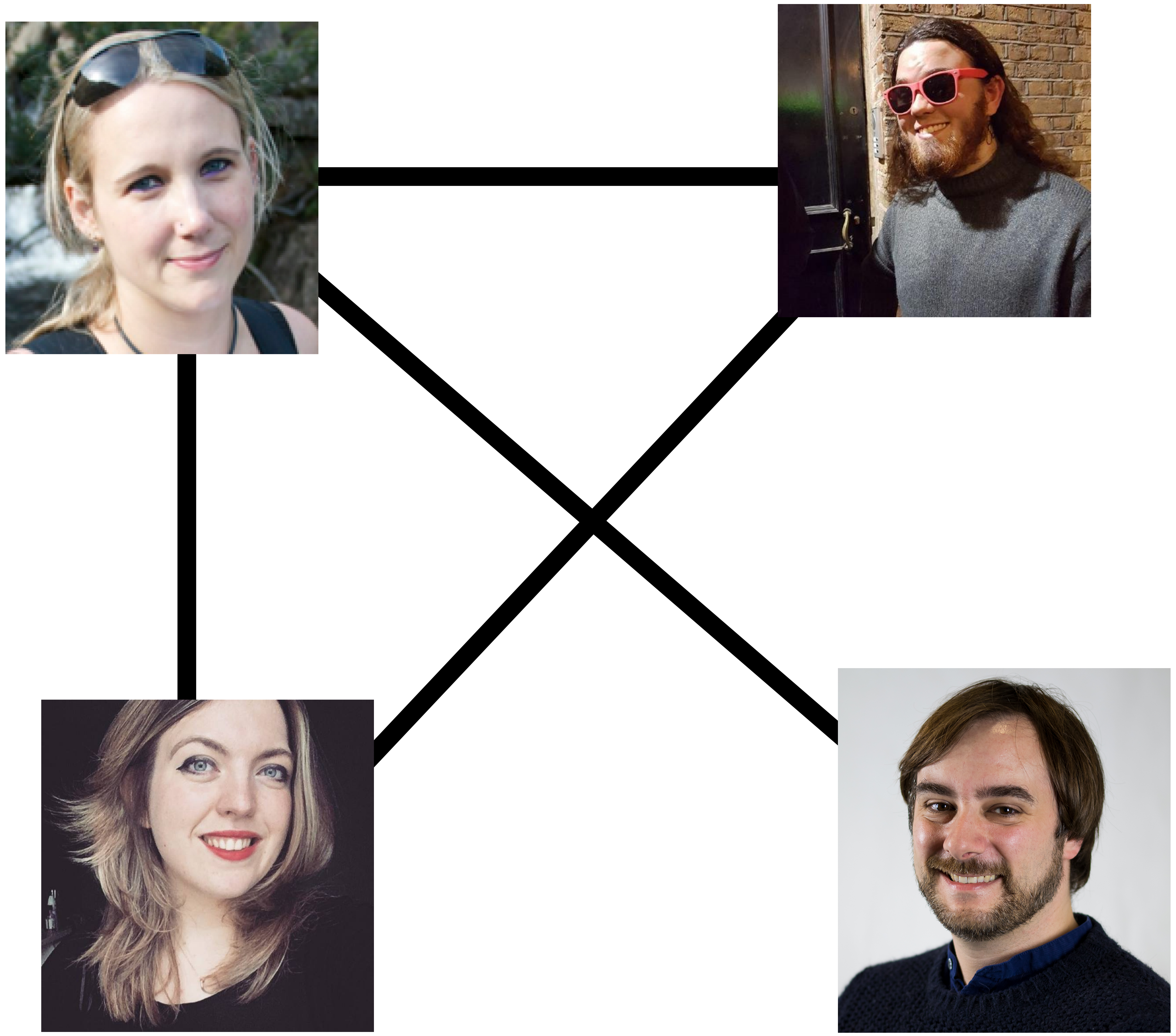
Paths and Cycles

The **distance** d(u,v) between two nodes is the length of the shortest path connecting them

d(Ben, Mathieu) = 2

The **diameter** of a graph is the largest distance between a pair of nodes in the graph

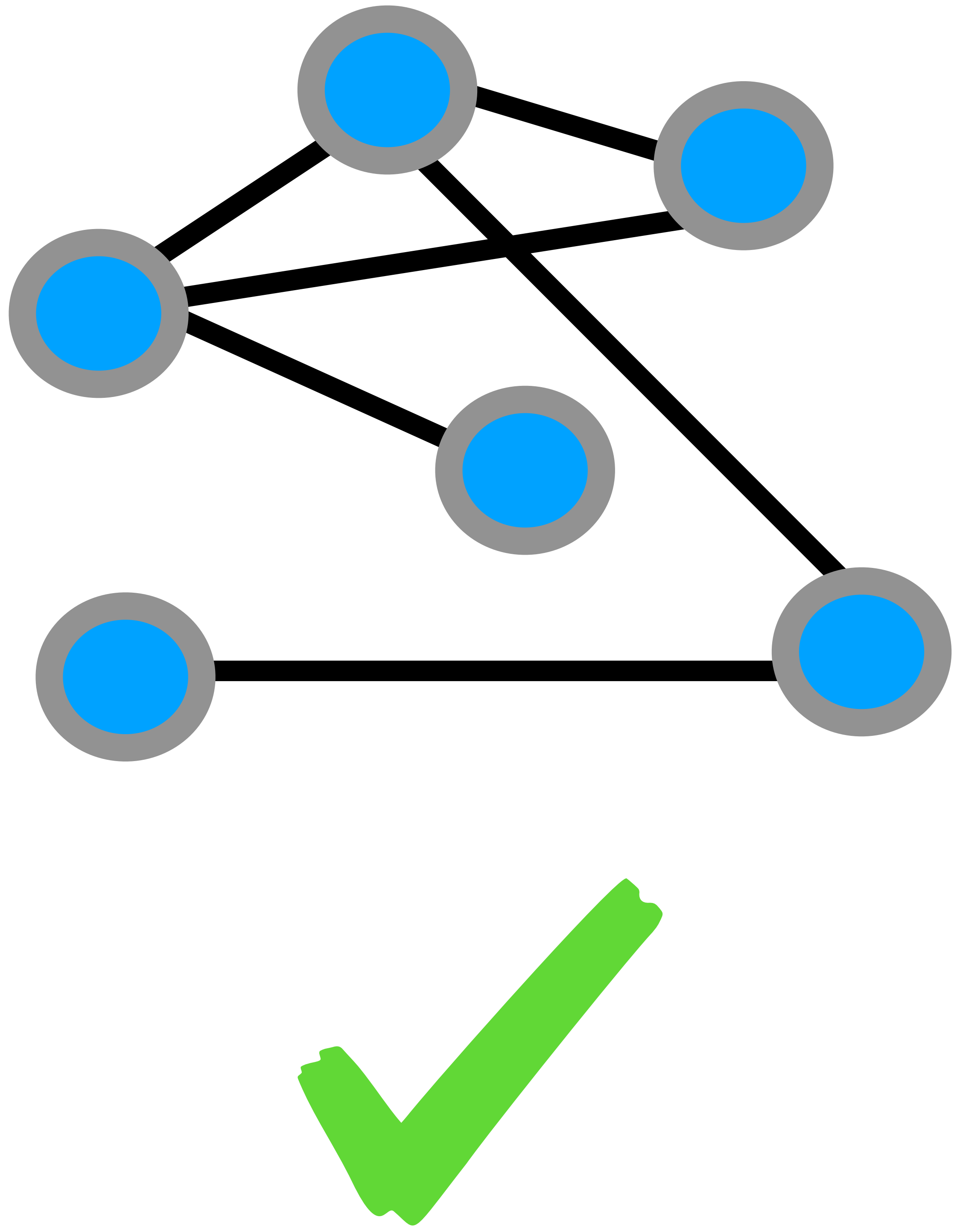
d(G) = 2

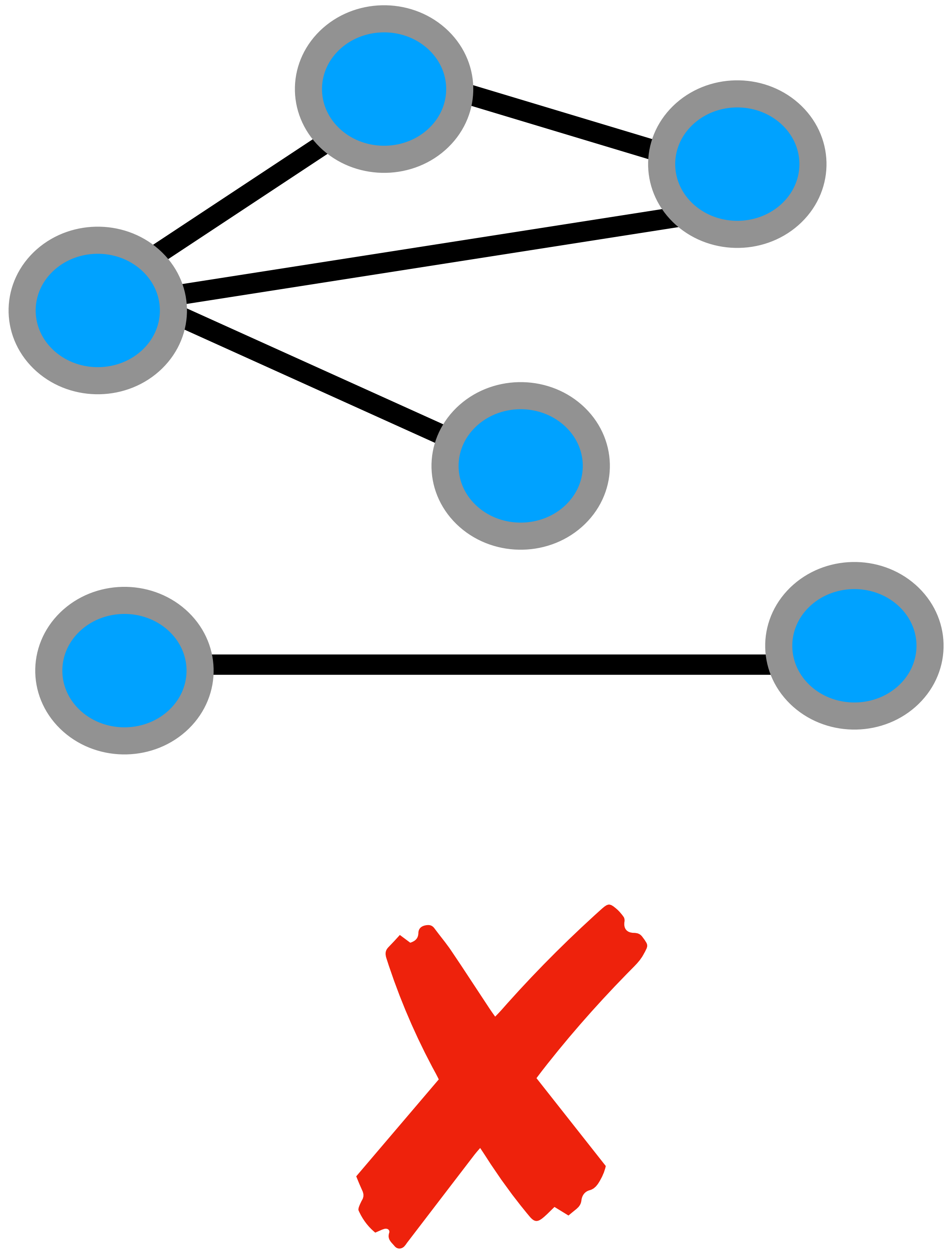


Often more meaningful to look at average path length

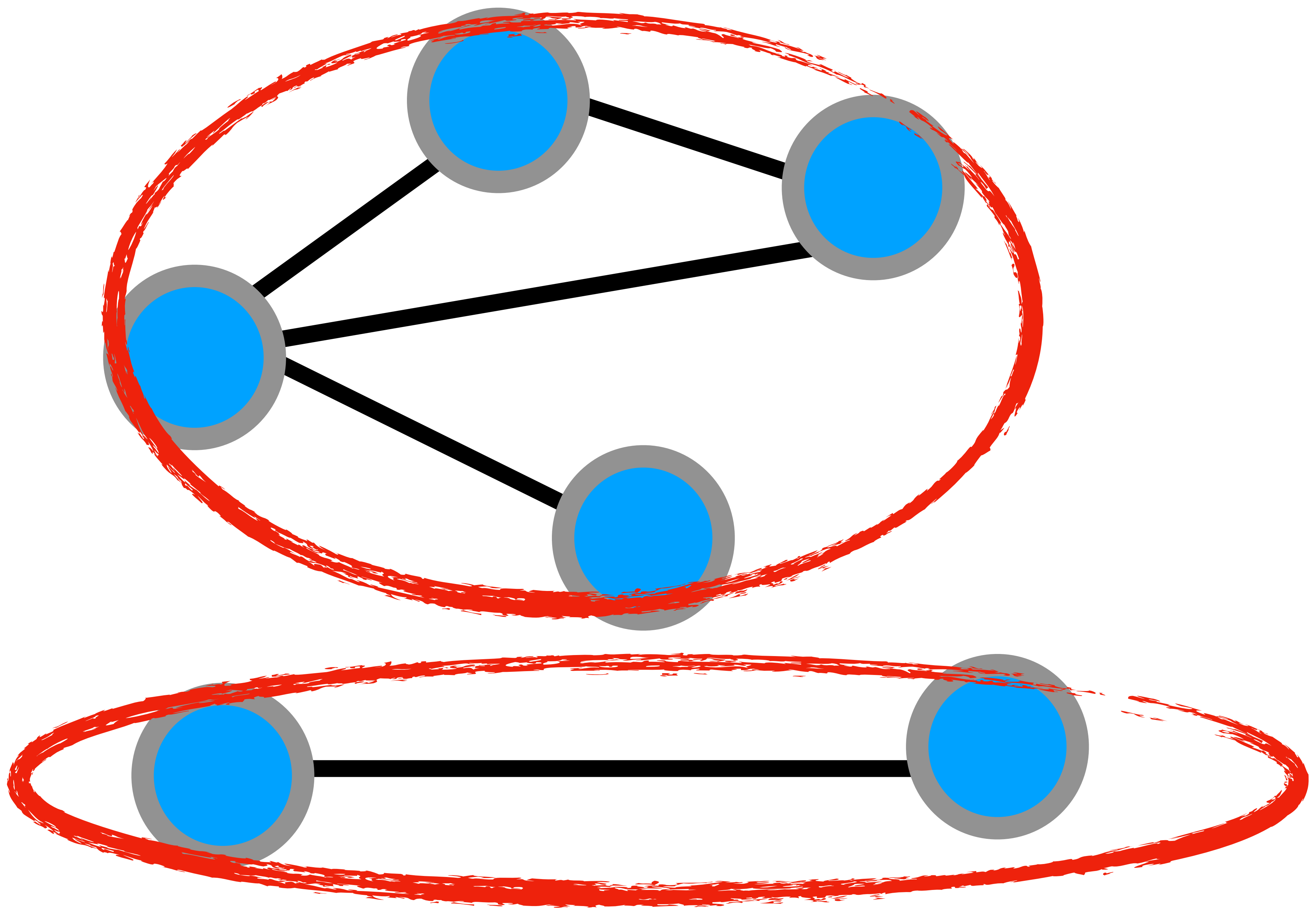
Connected Graph

A graph is **connected** if there is a path between every pair of vertices





Connected Components



A **connected component** of a graph G is a subgraph in which:

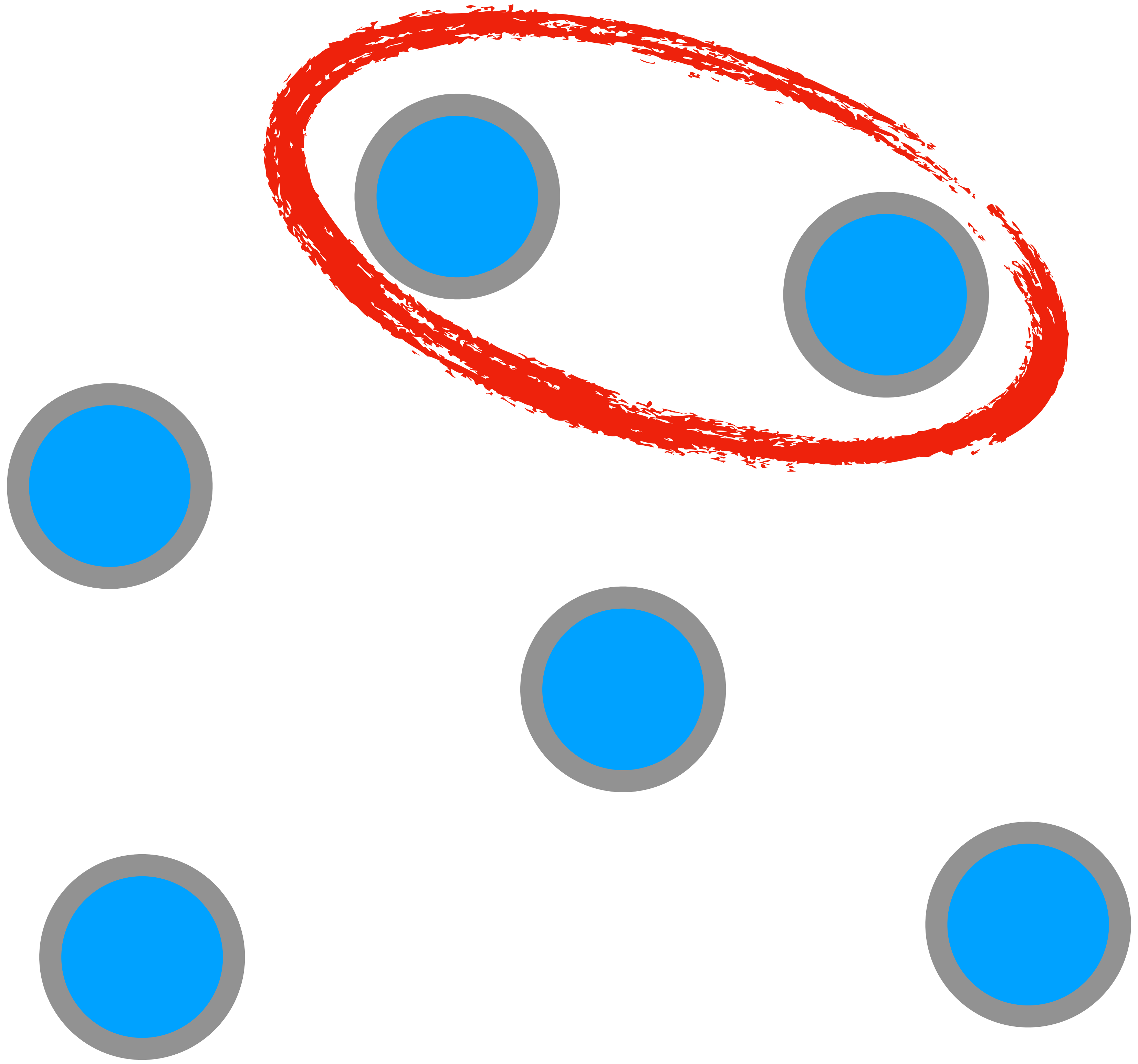
1. Any two vertices are **connected** by paths
2. There are **no edges** to other vertices in G.

Questions?

Erdos-Renyi Random Graph Model

- Want to model real networks, have some **baseline** to compare
- "Is the value of this network metric unusual?" Want a **null model**
- What is the **very simplest** model formulation we can look at?

Erdos-Renyi G(n,p) Model



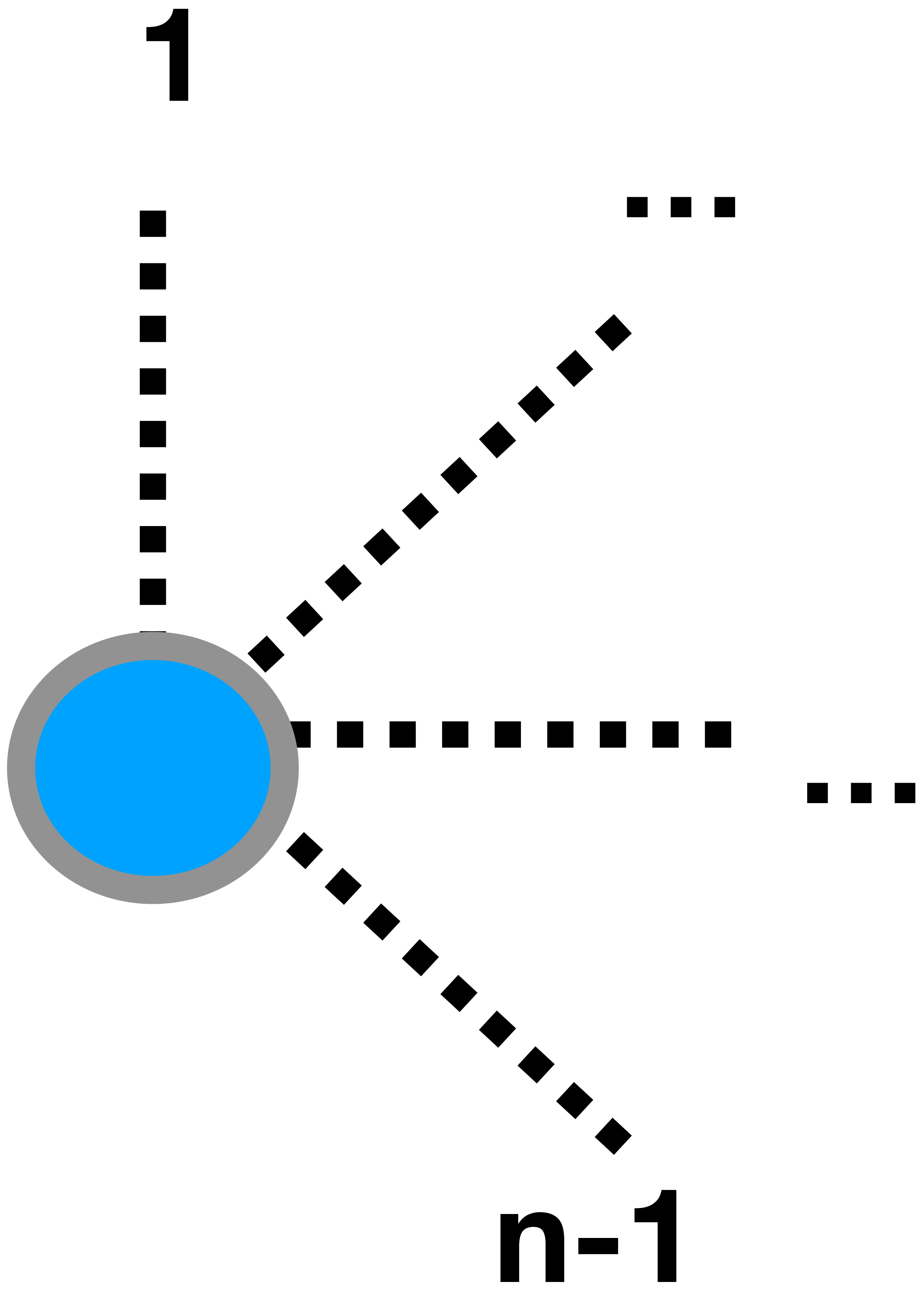
1. Start with an empty graph of **n** nodes

2. Acquire a biased coin with head probability **p**

3. For each pair of nodes, do a coin toss. If heads, draw an edge between them. If not, move on.

Erdos-Renyi G(n,p) model

Average degree of ER networks



For each node, there are n-1 others in the graph it could connect to.

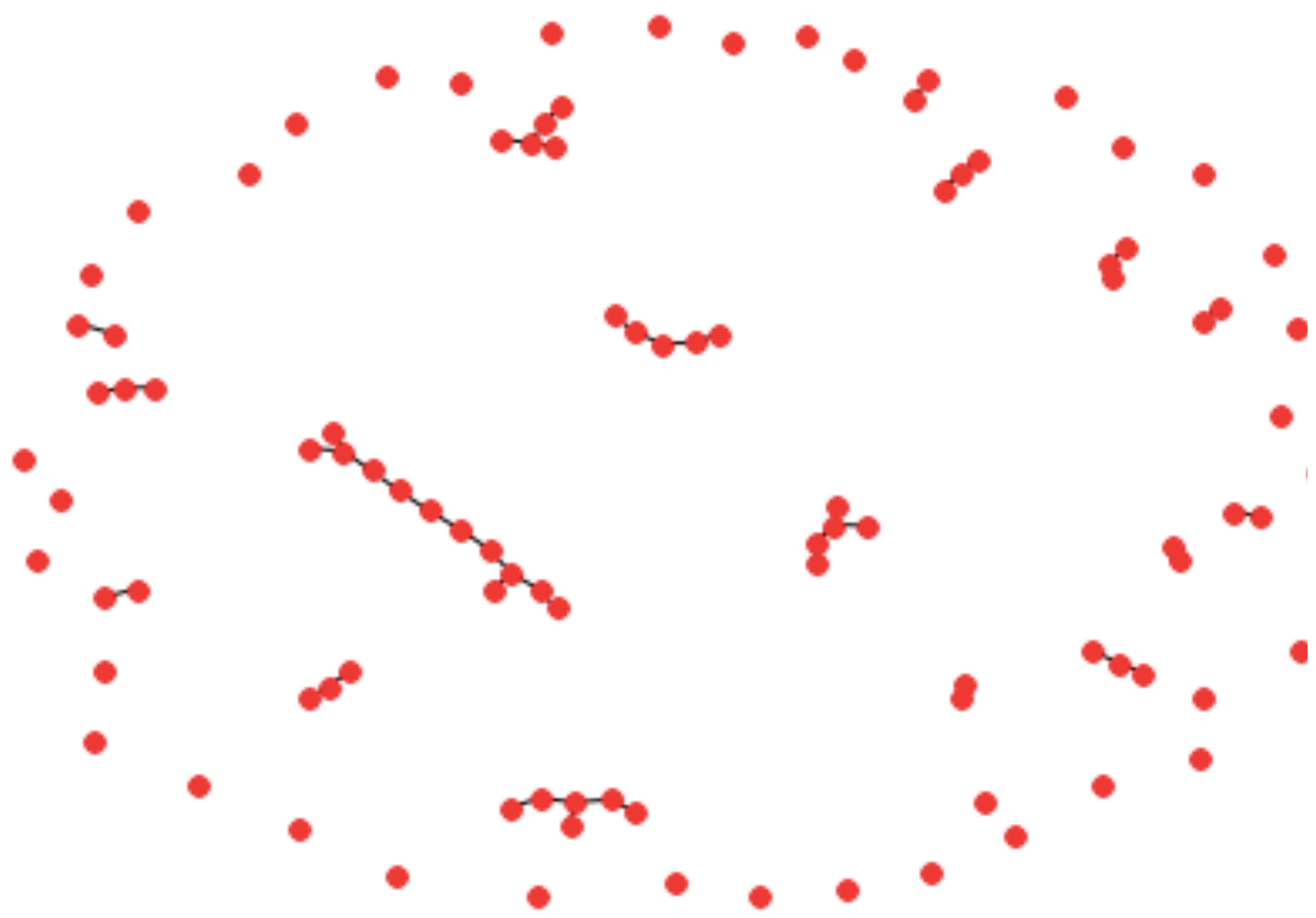
Each of those connections can happen with probability p

(If you were a fan of Probability and Matrices, this is a binomial with **n-1 trials** and **success probability p**)

So average degree is **(n-1)p**, or approximately **np**

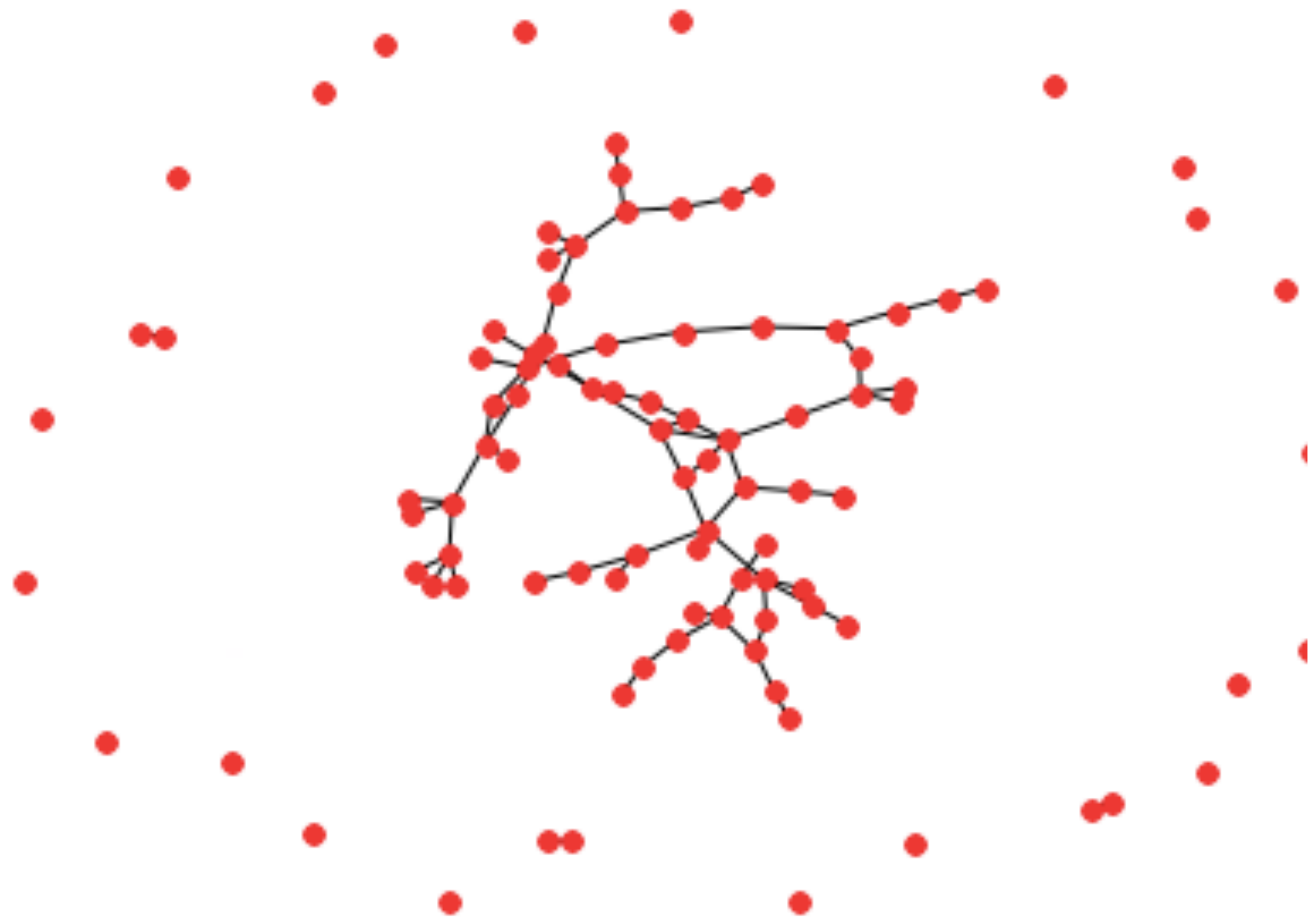
What do ER graphs look like?

$$p < \frac{1}{n}$$



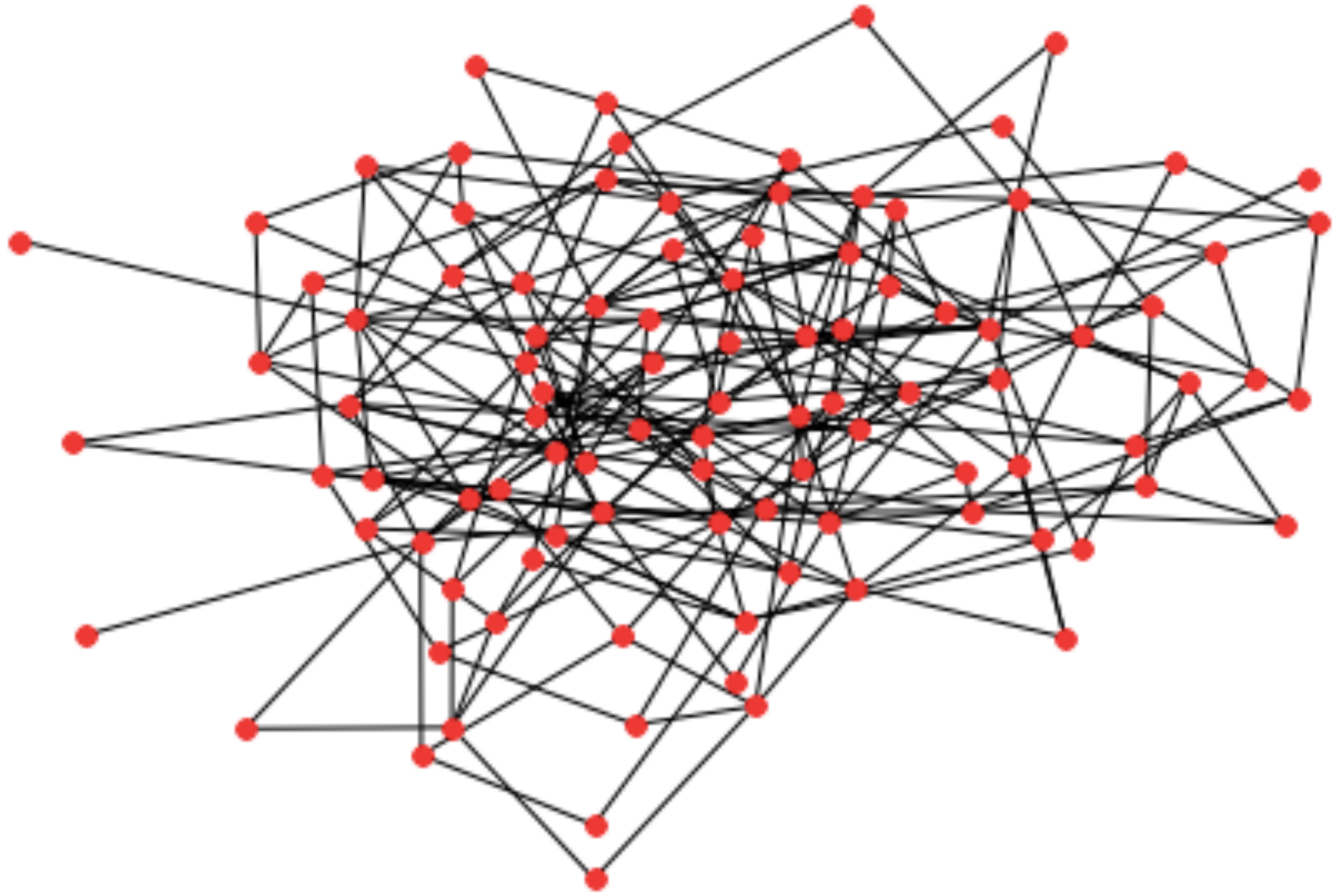
Very disconnected graph, only tiny connected components

$$p = \frac{1}{n} + \epsilon$$



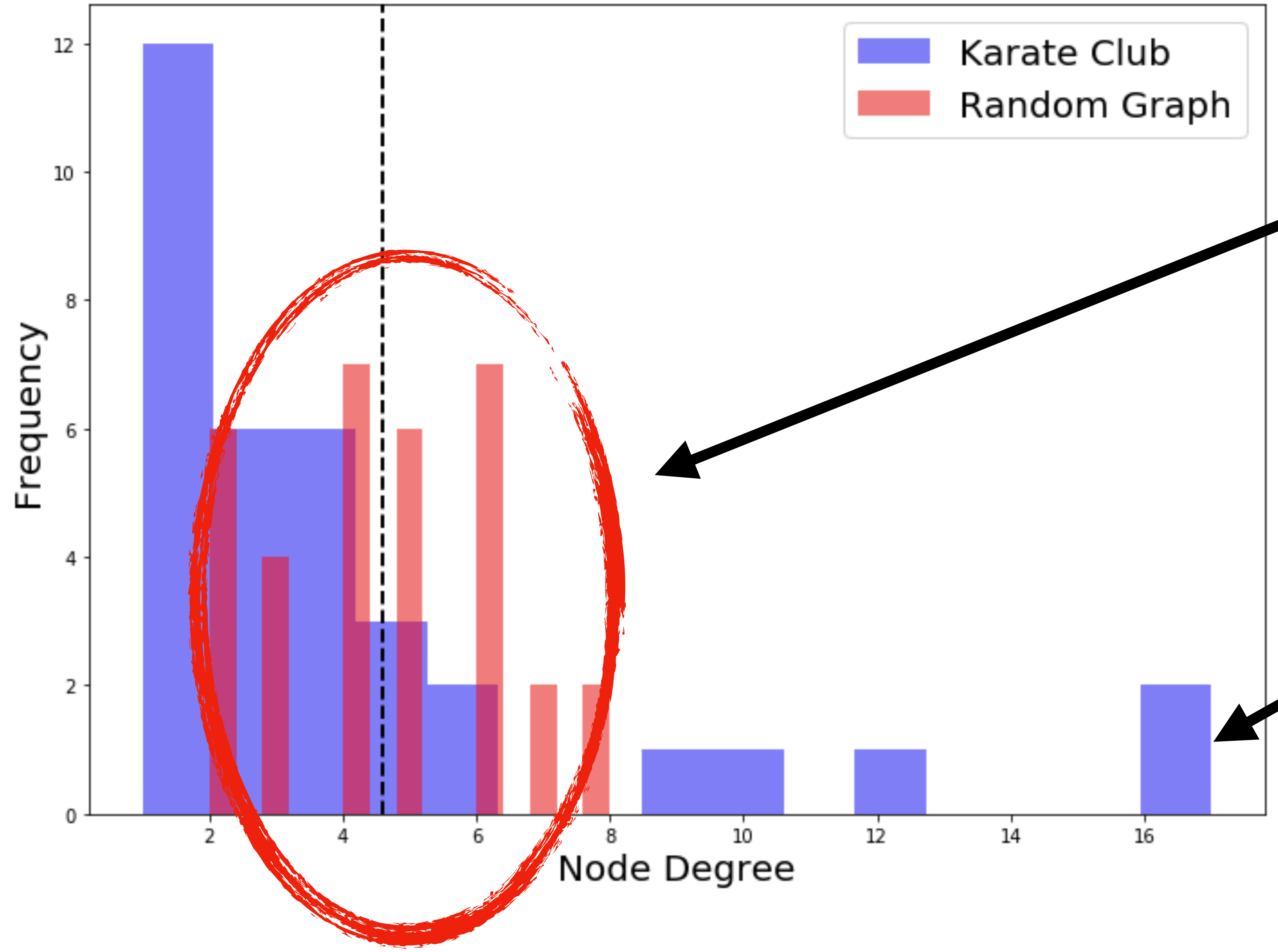
A giant component appears, no/very few cycles

$$p > \frac{\log(n)}{n}$$



Whole graph is connected, some cycles present

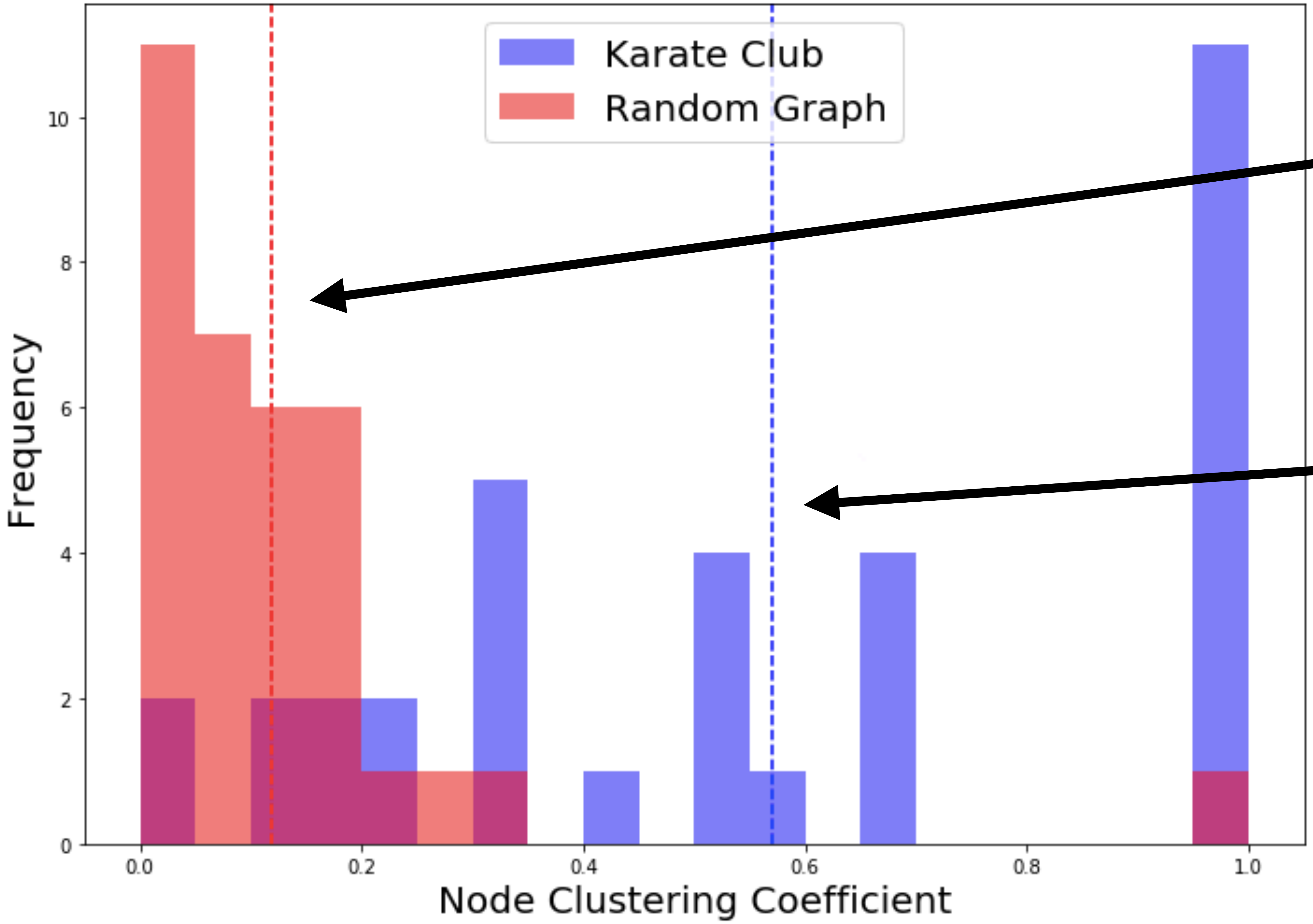
Random Graphs vs Real Networks



Random: node degrees all clustered round the average value

Real: small number of high degree nodes, large number of low degree nodes

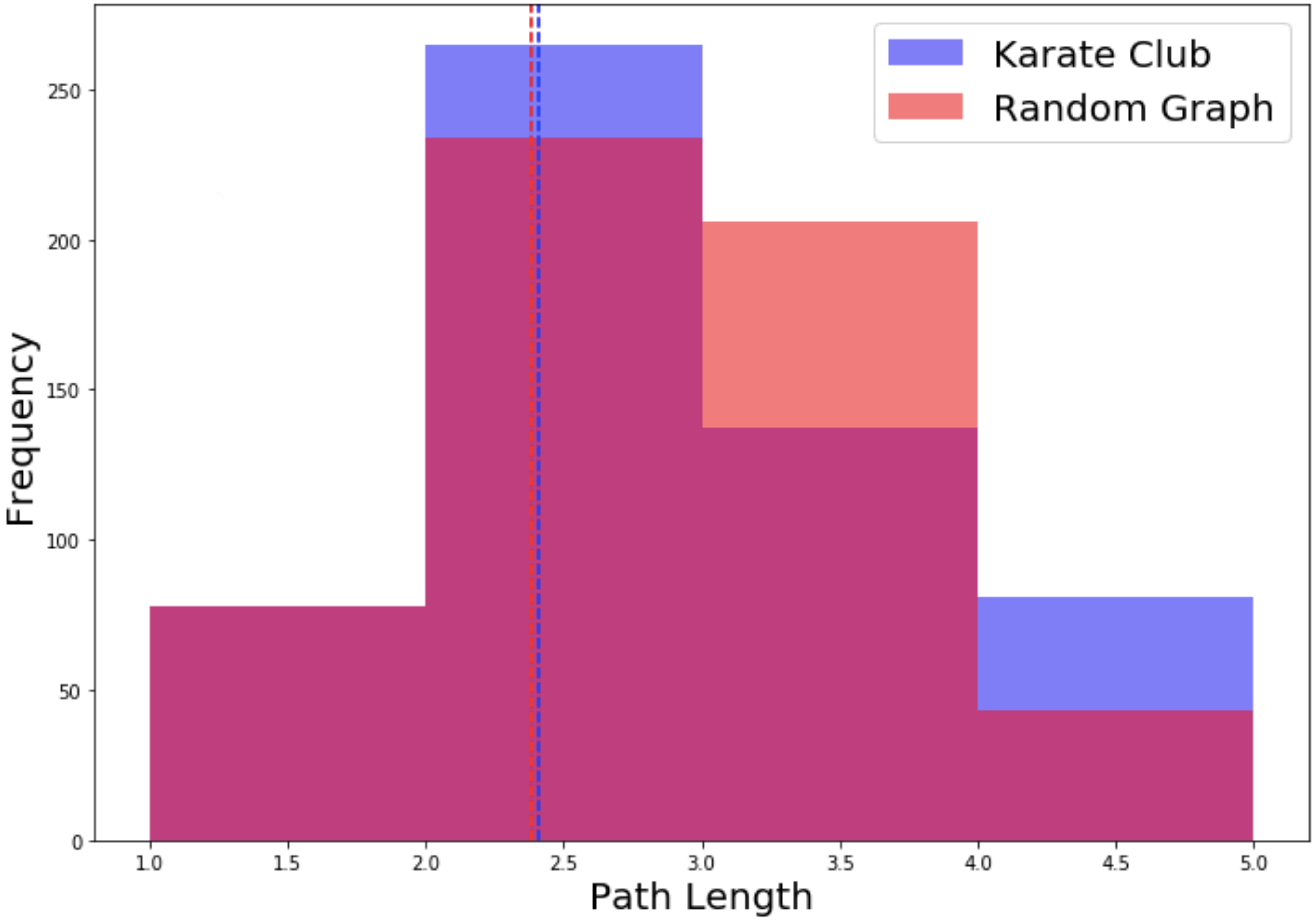
Random Graphs vs Real Networks



Random: very low average clustering coefficent

Real: much higher average clustering coefficient, with some nodes having very high values

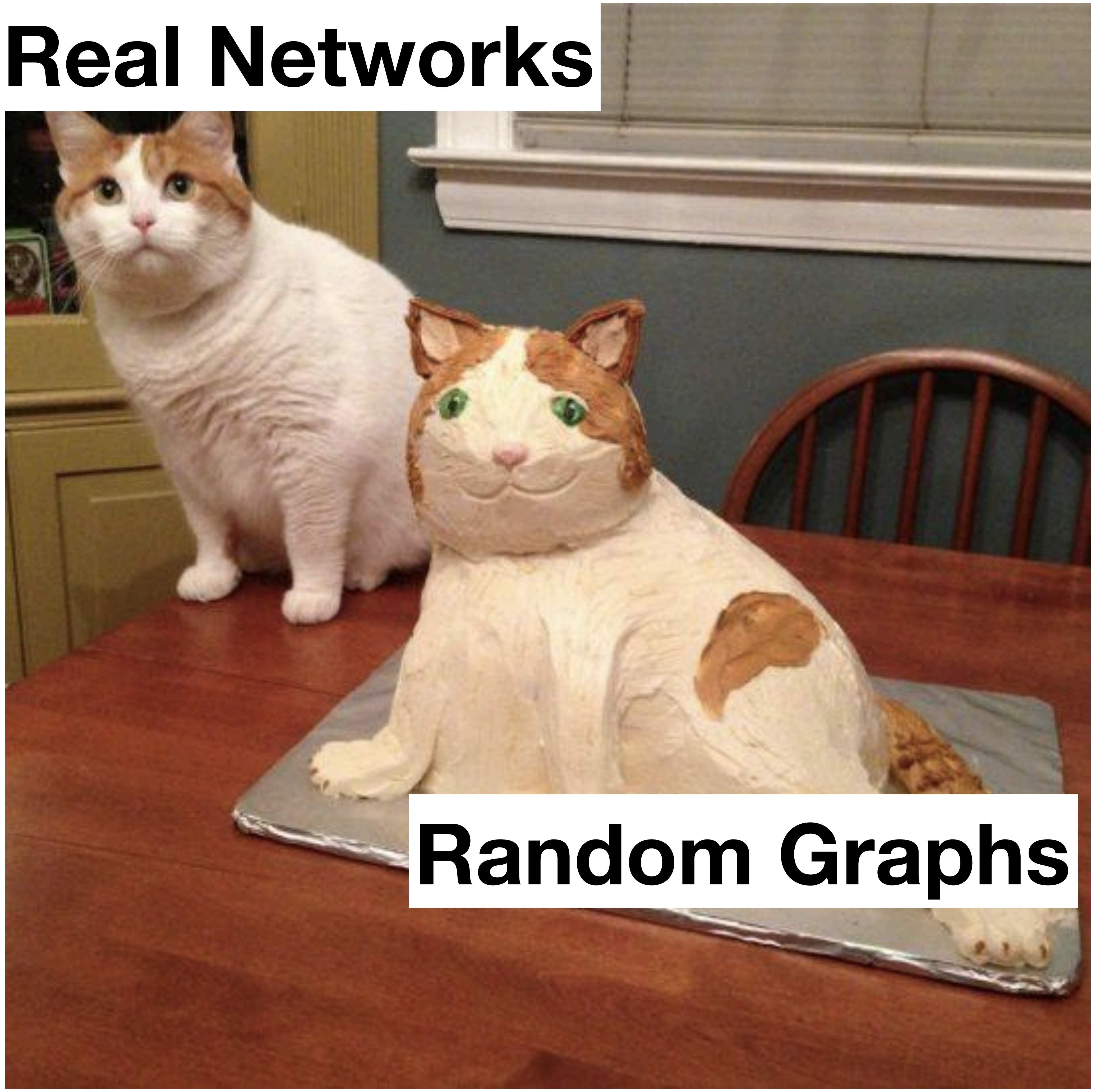
Random Graphs vs Real Networks



Fairly spot on with **almost the same** average path length for each!

Summary: Random Graphs vs Real Networks

	Real Social Networks	Random Graphs	?
Degree Distribution	Heavy Tailed (most nodes have low degree, small few with high degree)	Light tailed (all nodes have close to the average degree)	?
Clustering Coefficient	High	Low	?
Path Lengths	Low	Low	?
?	?	?	?



Thank you for listening! What are your questions?