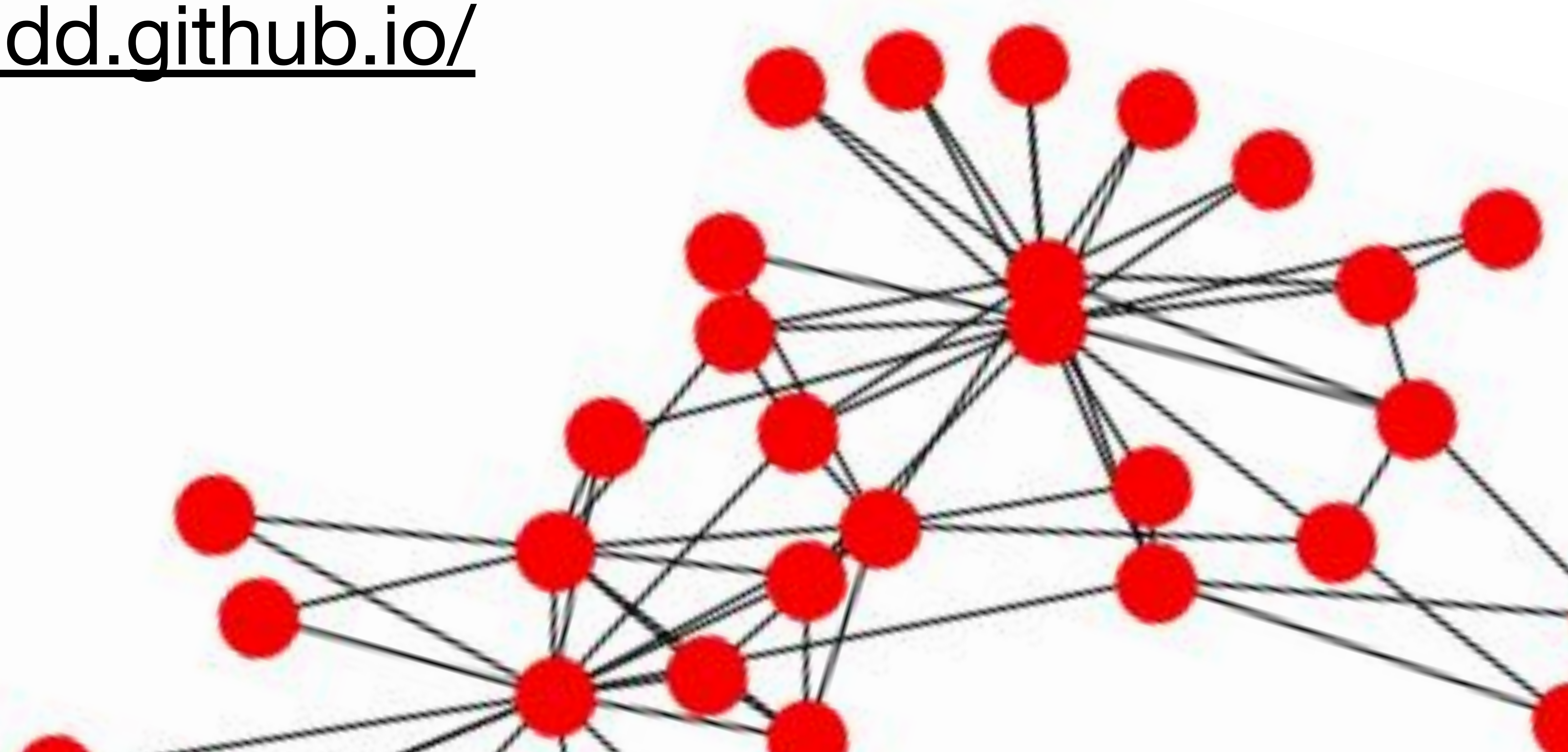
DMSN Tutorial 1: Networks and Random Graphs

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

Session will start at 9:05, see you soon! :-)



About me

PhD student in the Networks Group

Email
n.a.arnold@qmul.ac.uk
Webpage
https://narnolddd.github.io

Naomi Arnold

PhD Student

Queen Mary University of London
Email
Twitter
LinkedIn
Instagram
Github

Naomi Arnold

I am Naomi Arnold, a PhD student within the Networks group in the School of Electroni
Engineering and Computer Science at Queen Mary University of London. My supervisor
Richard Clegg and Raul Mondragon. My research interests are broadly in modelling dif
social and information systems as evolving graphs. Specific areas of interest to me are

- Network growth models: model selection and changepoint detection.
- Tools for temporal networks.

I maintain the FETA (Framework for Evolving Topology Analysis) codebase with Richar
which can be used for generating graphs from different growth models, and for model
(paper describing the background and process here).

I am also a contributor to the Raphtory software for the analysis of temporal graphs.

News

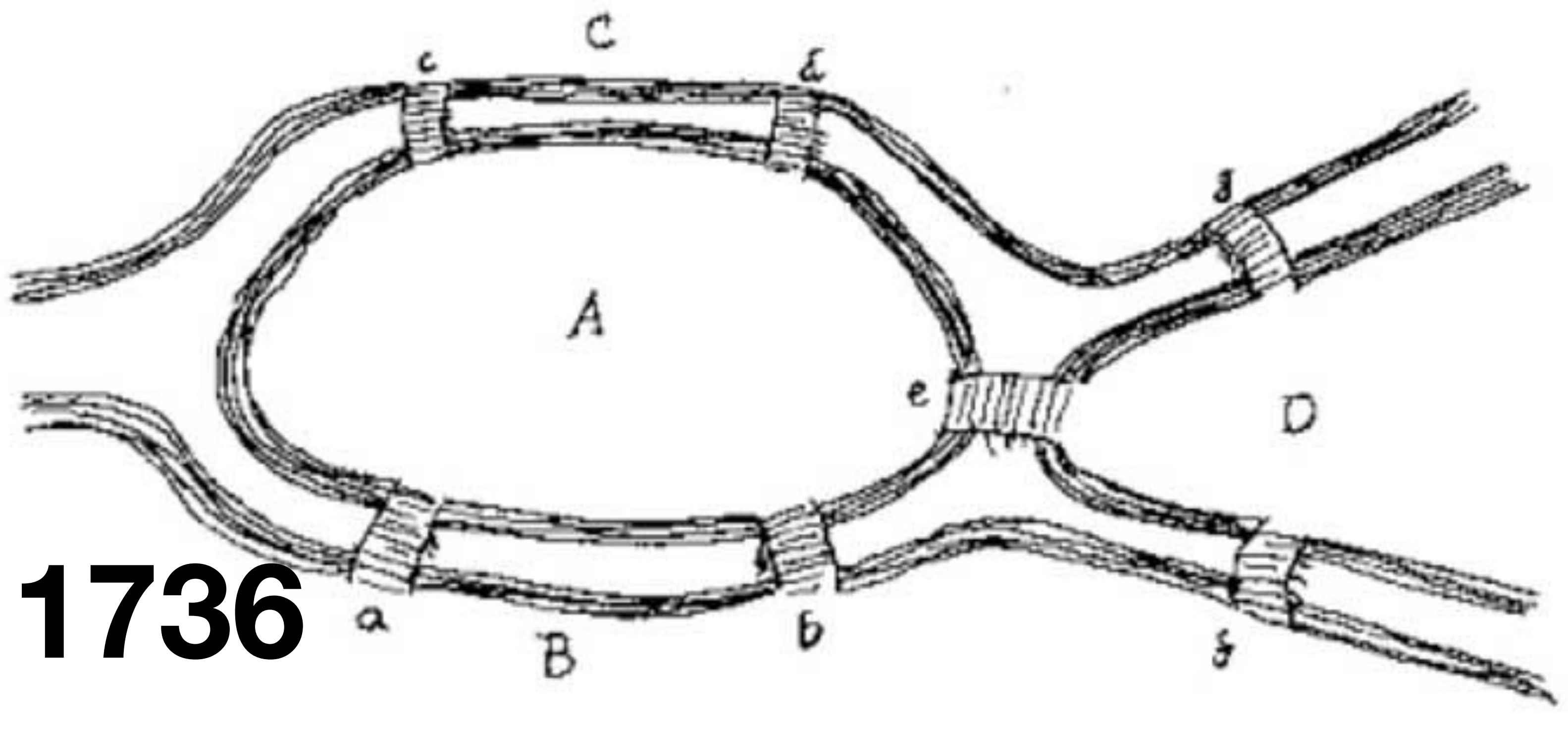
Housekeeping

- Asking questions -- “raise hand” feature
- Chat channels -- bear in mind that moderators can see these
- Session recordings -- each session will be recorded
- Tutorial materials -- will be uploaded after end of each session

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

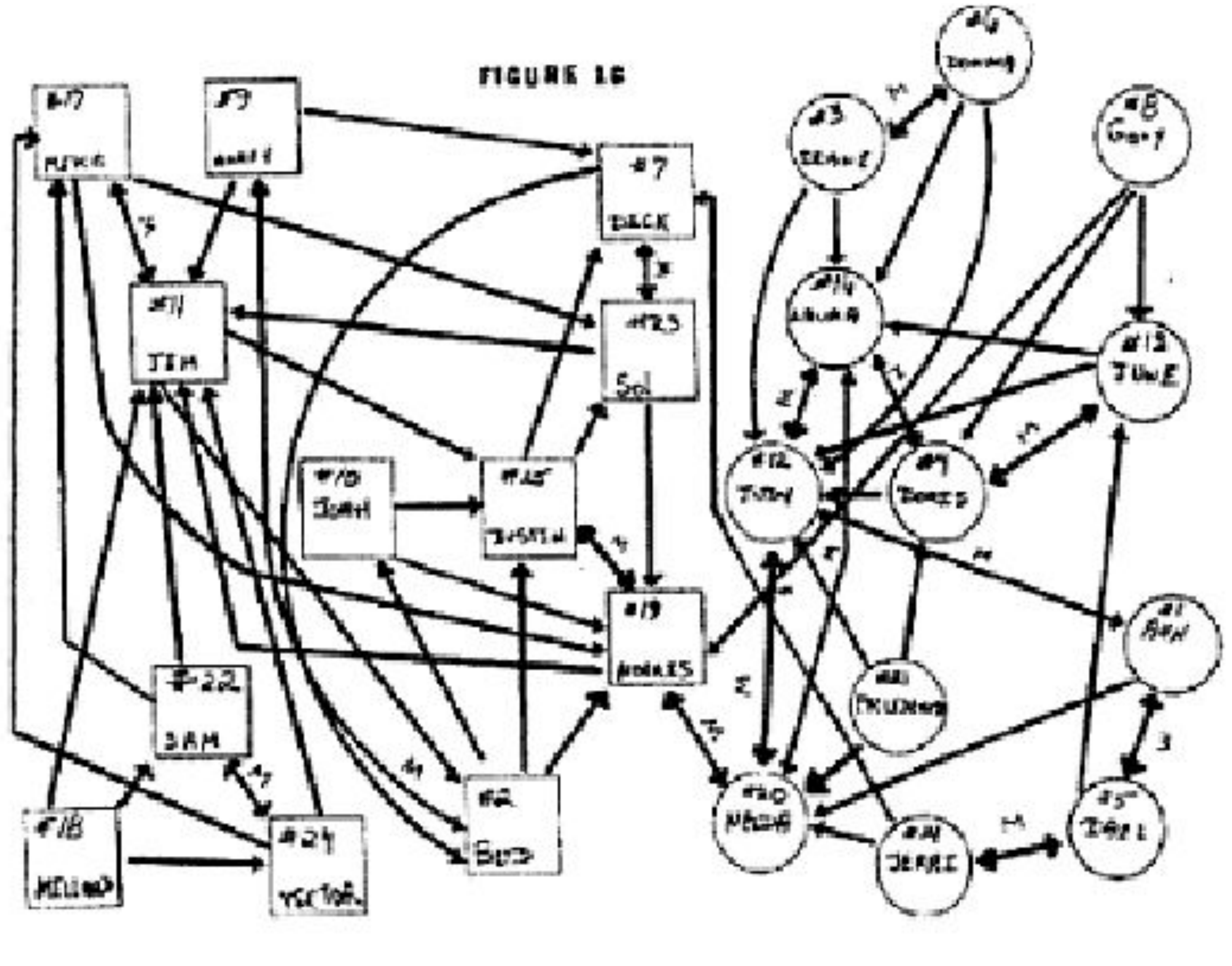
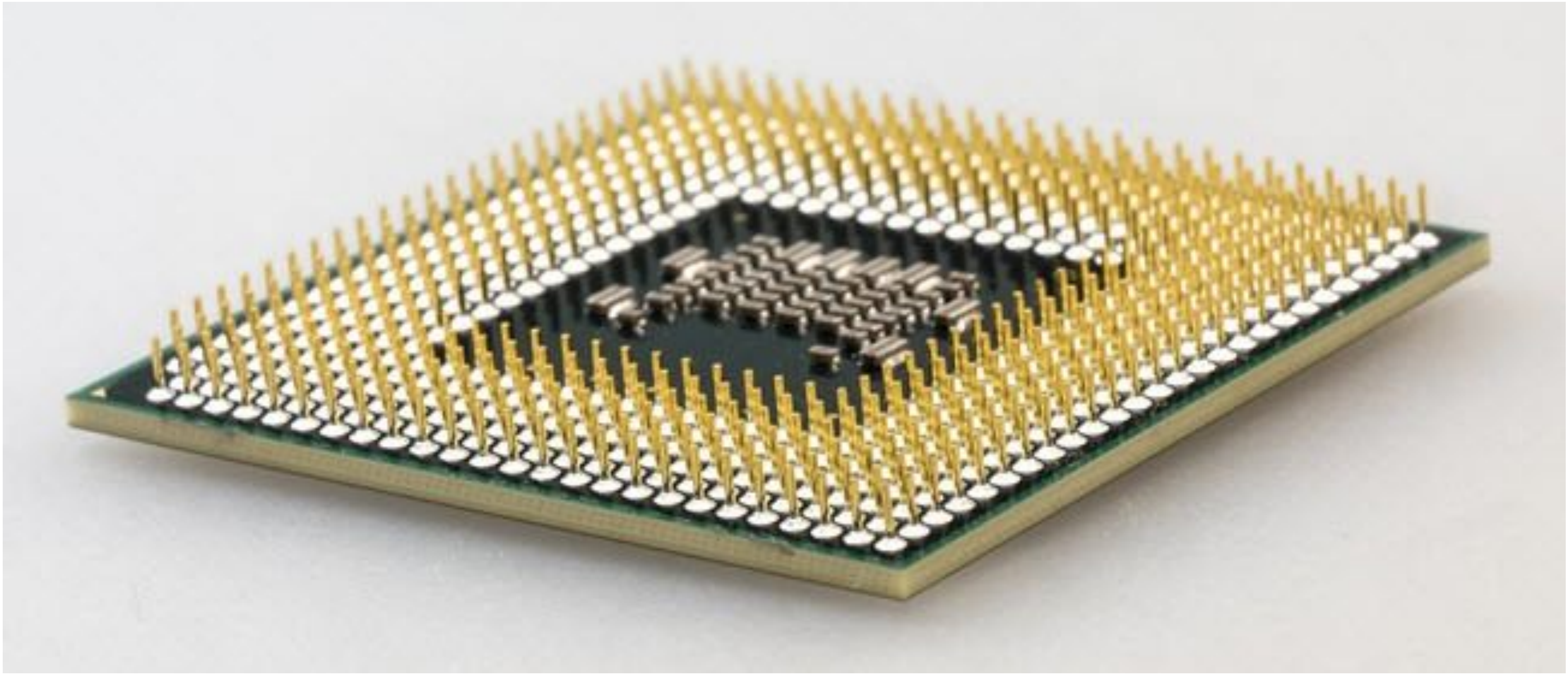


Figure 3.3 Moreno's 1933 Sociogram

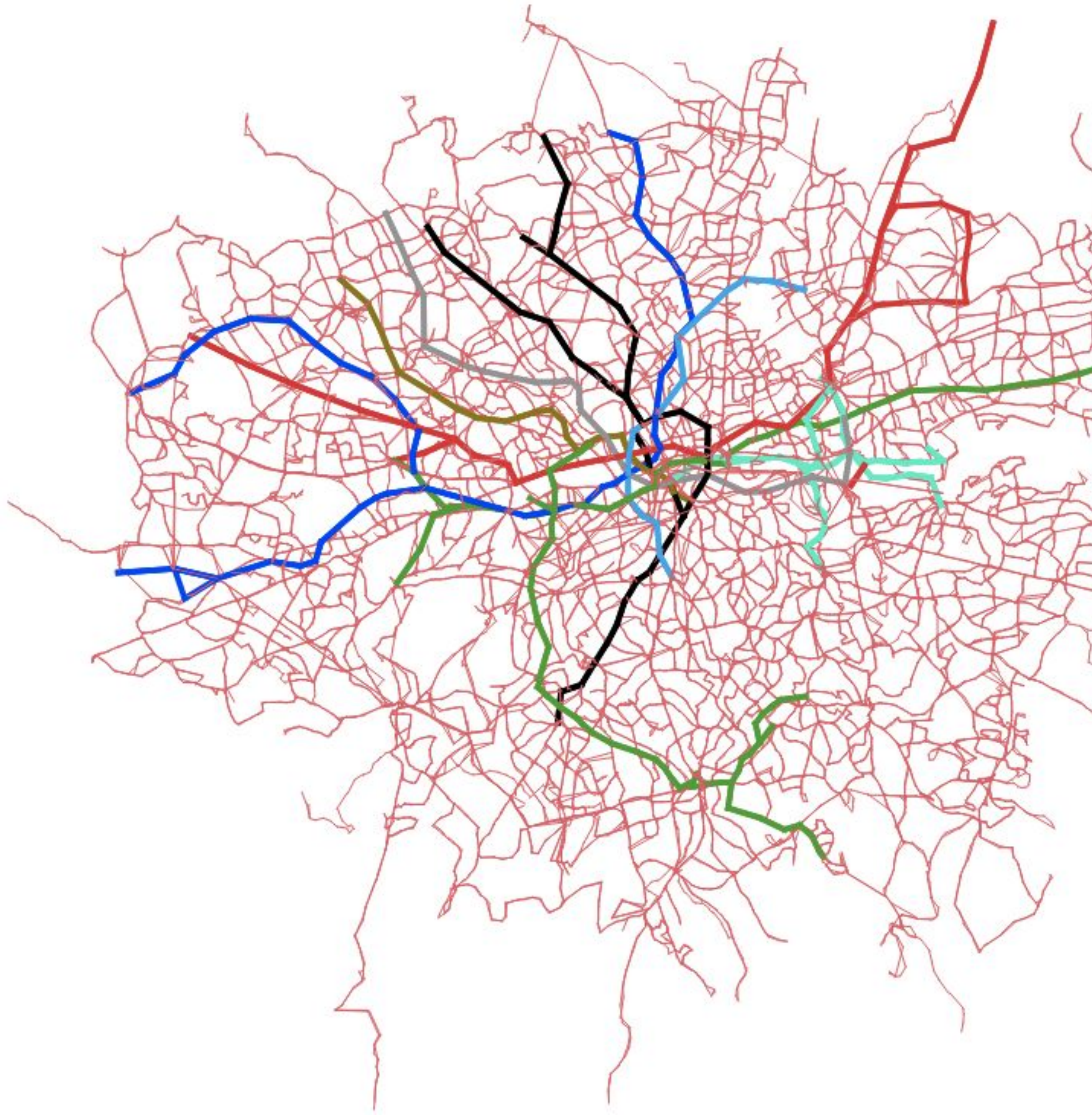
1933



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

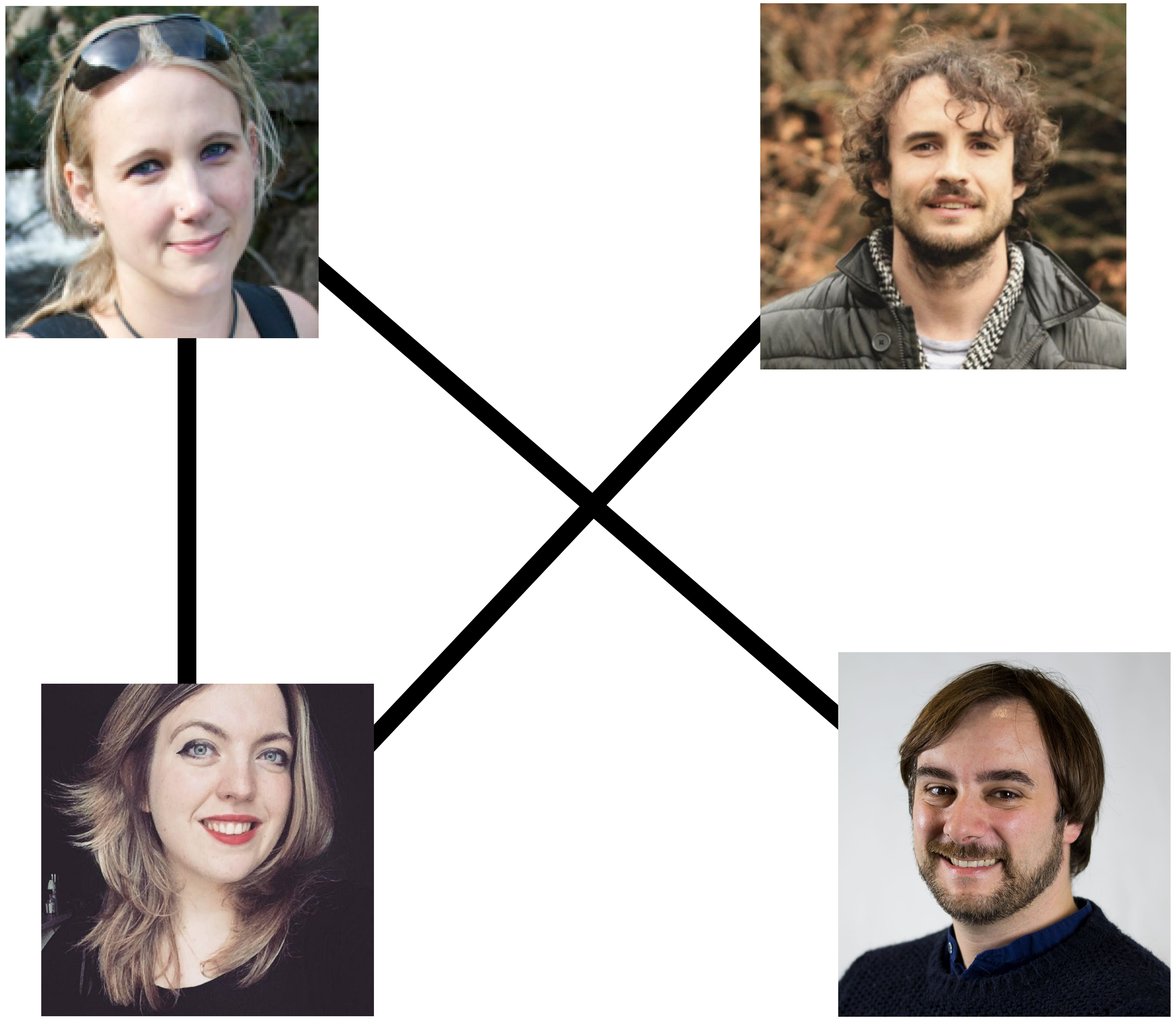
(Undirected) Graph

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

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

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

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

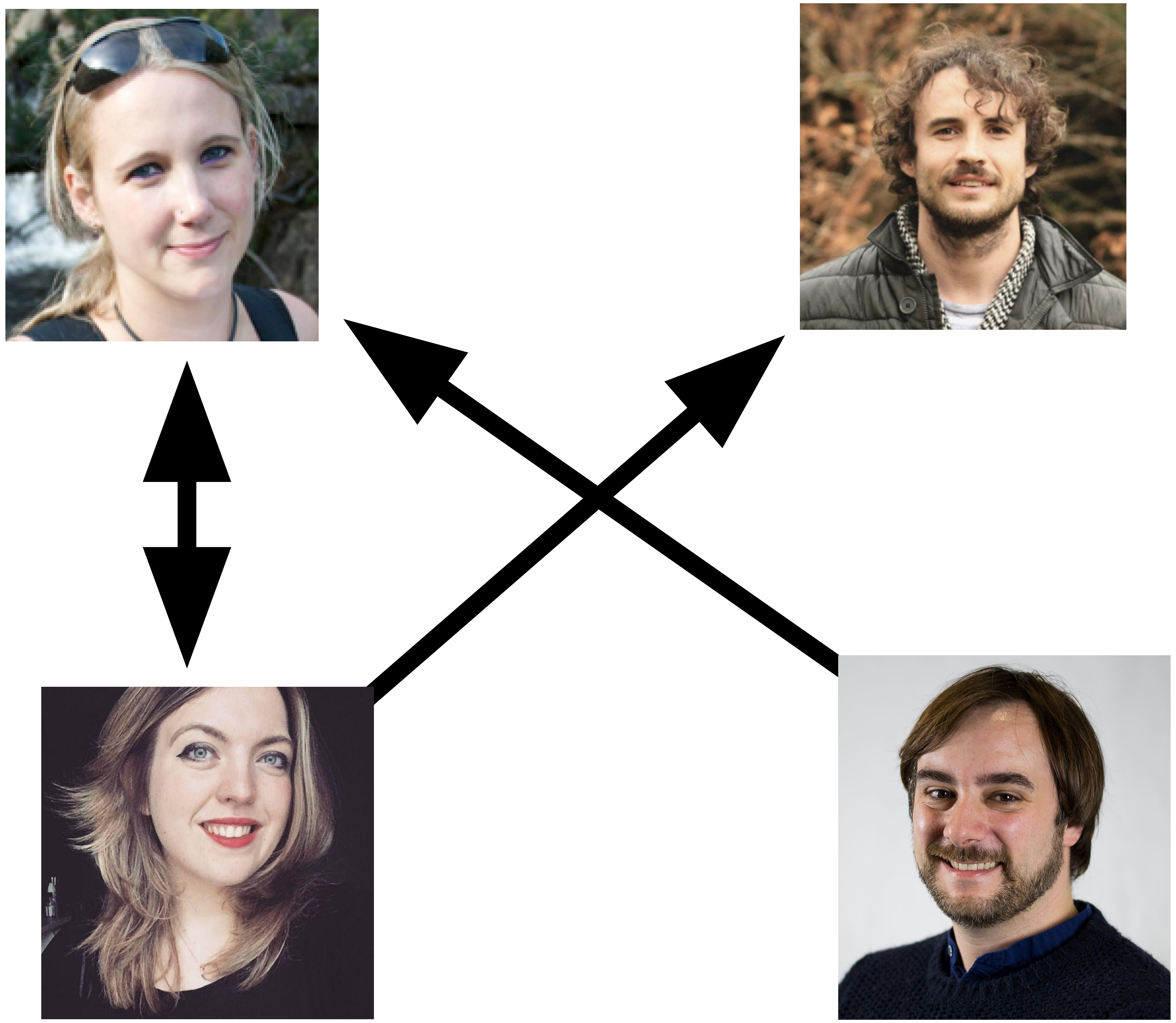


Directed Graph

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

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

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

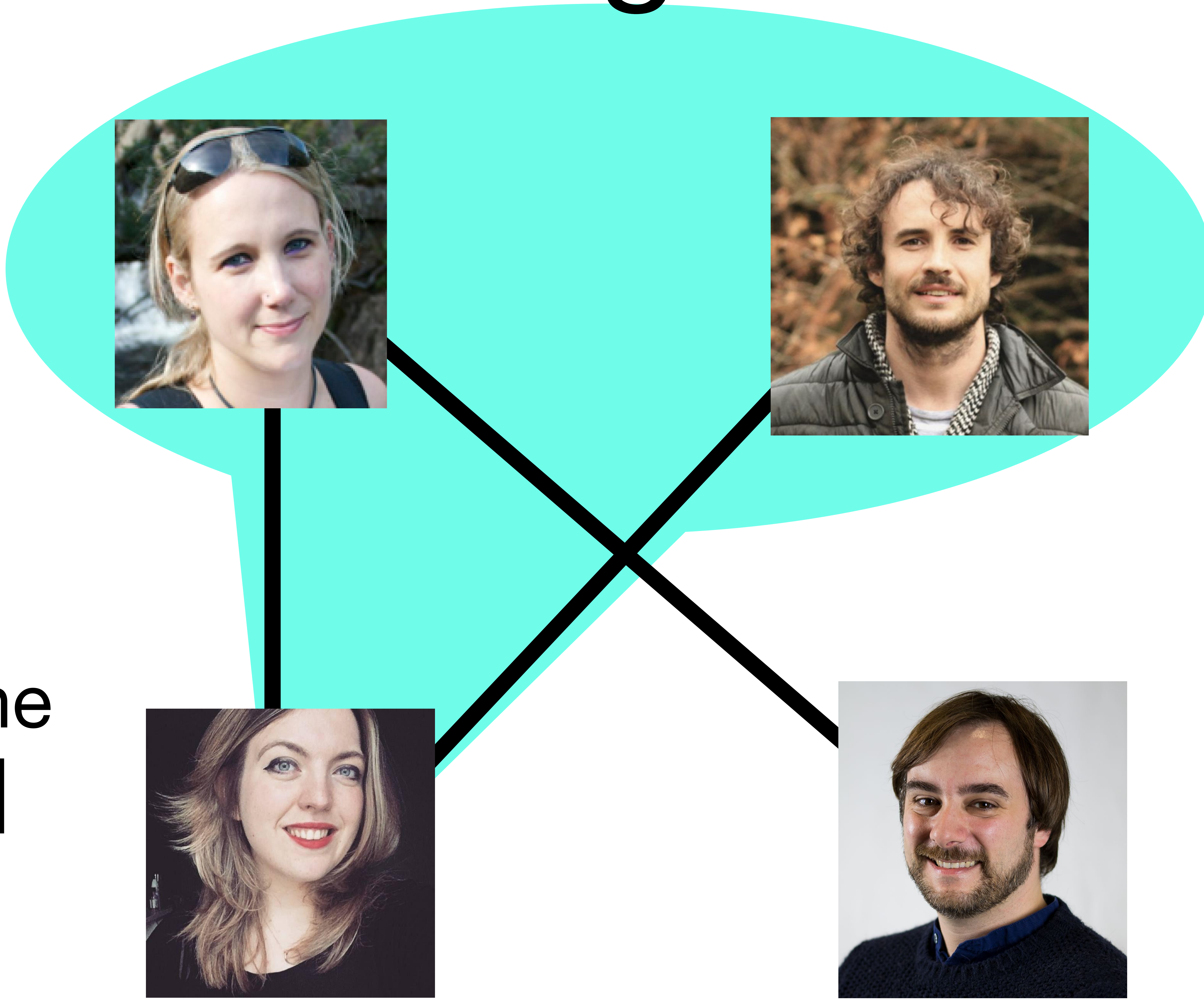


How do we measure graphs?
How do we compare them?

Neighbourhood and Degree

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

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



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

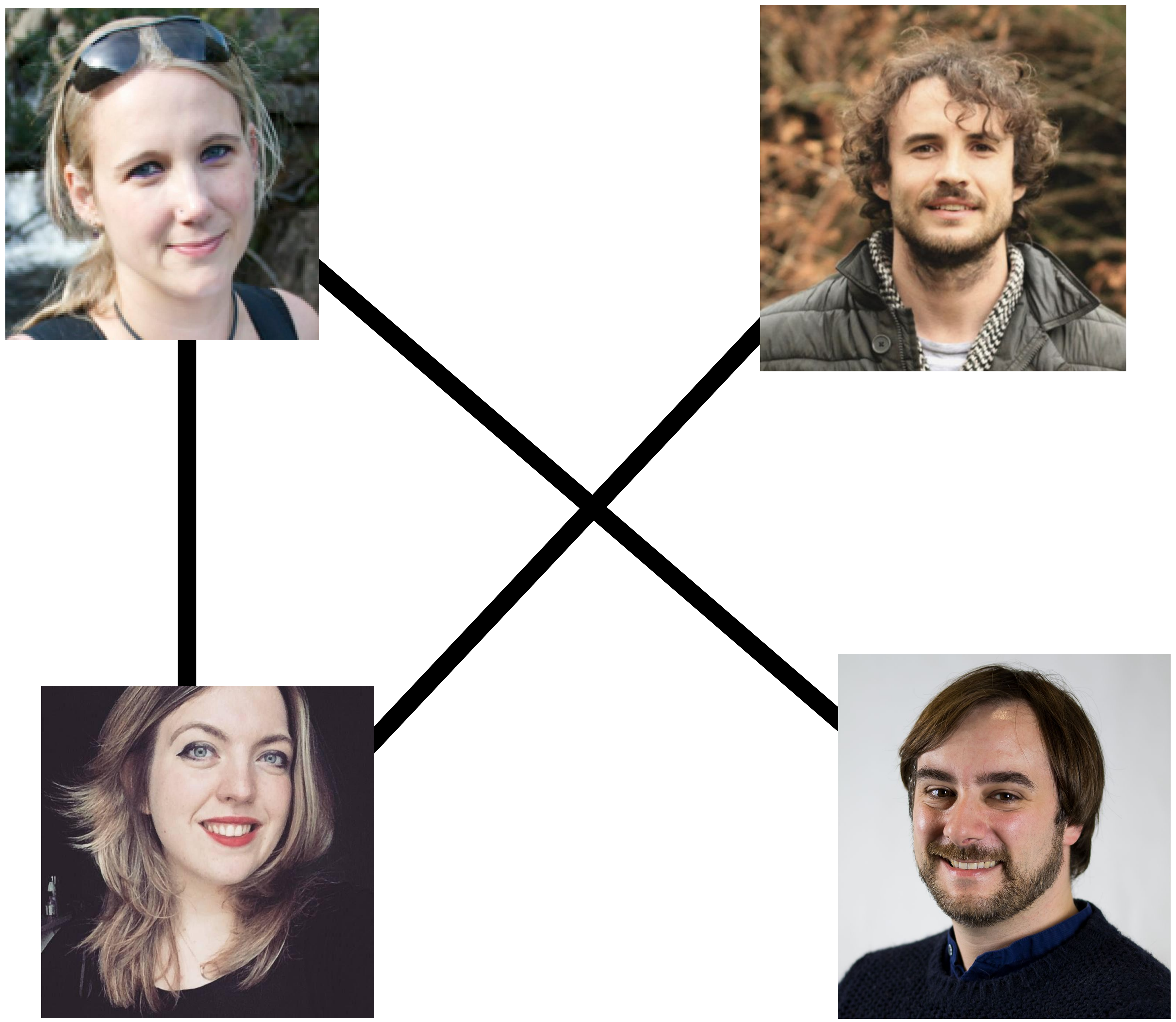
The **degree sequence** of a graph is the list of the vertex degrees for that graph

e.g. 2, 2, 1, 1

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

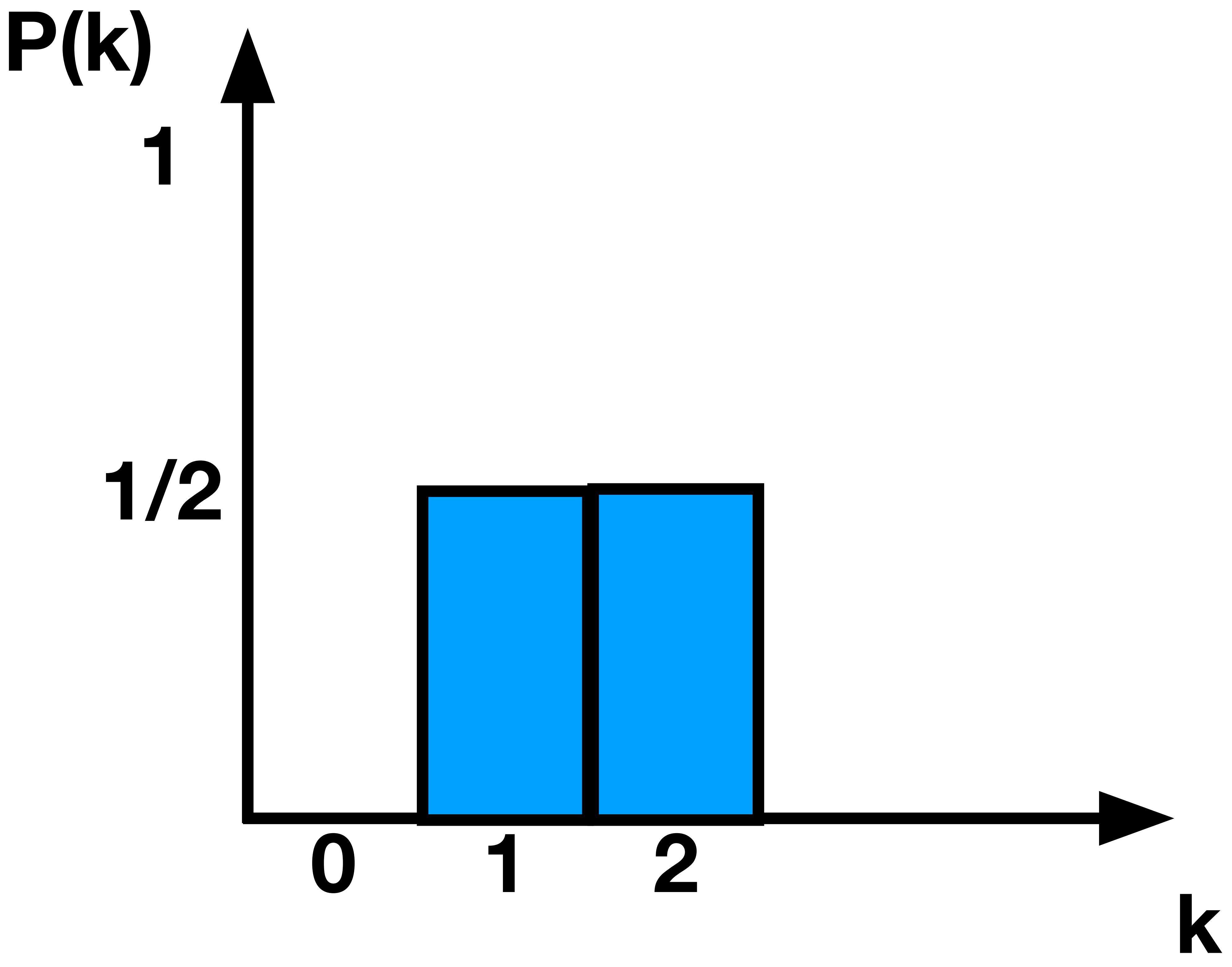
e.g. <k> = (2 + 2 + 1 + 1)/4 = 1.5

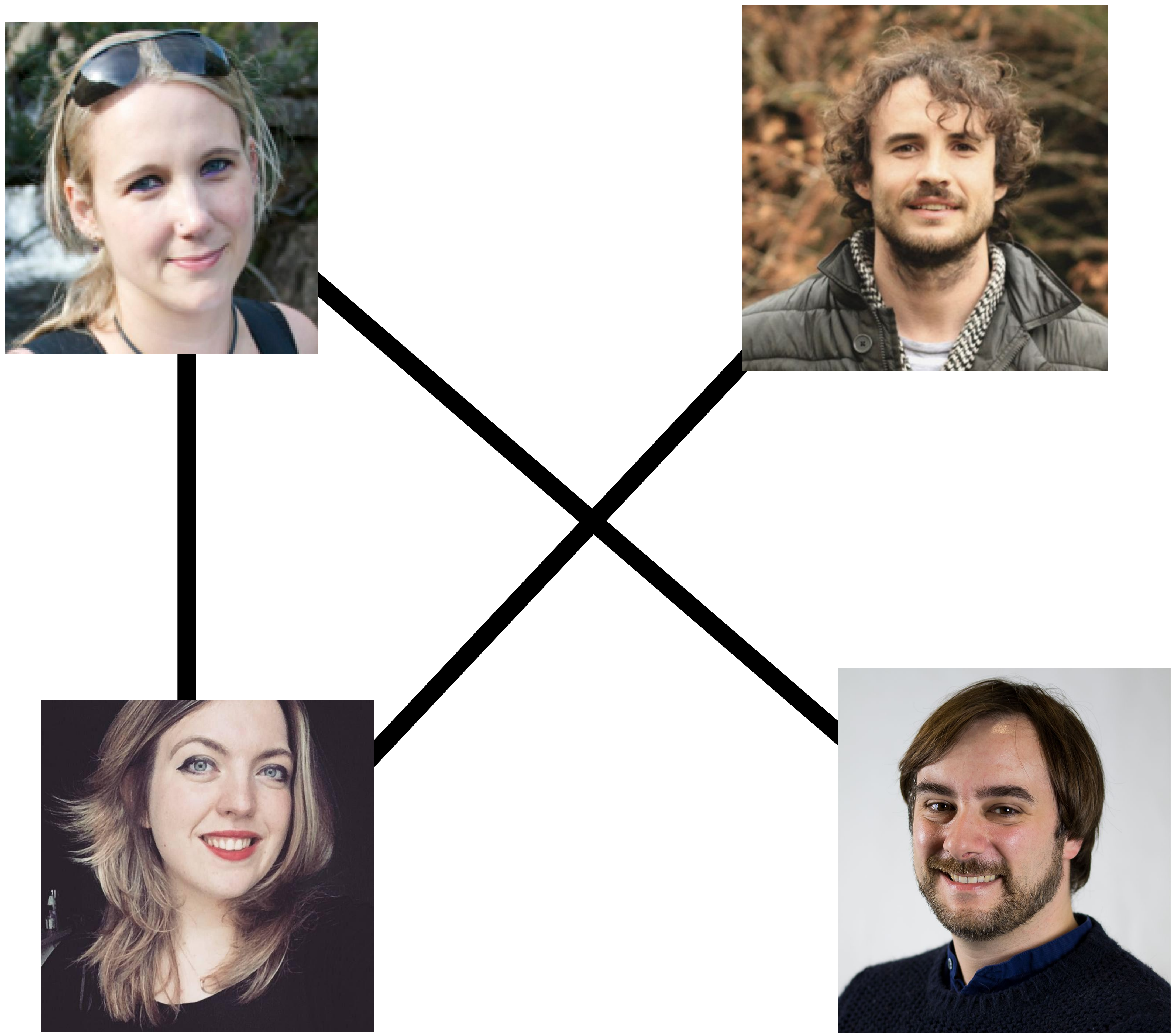
(also equal to 2*|edges|/|nodes|… why?)



Degree distribution

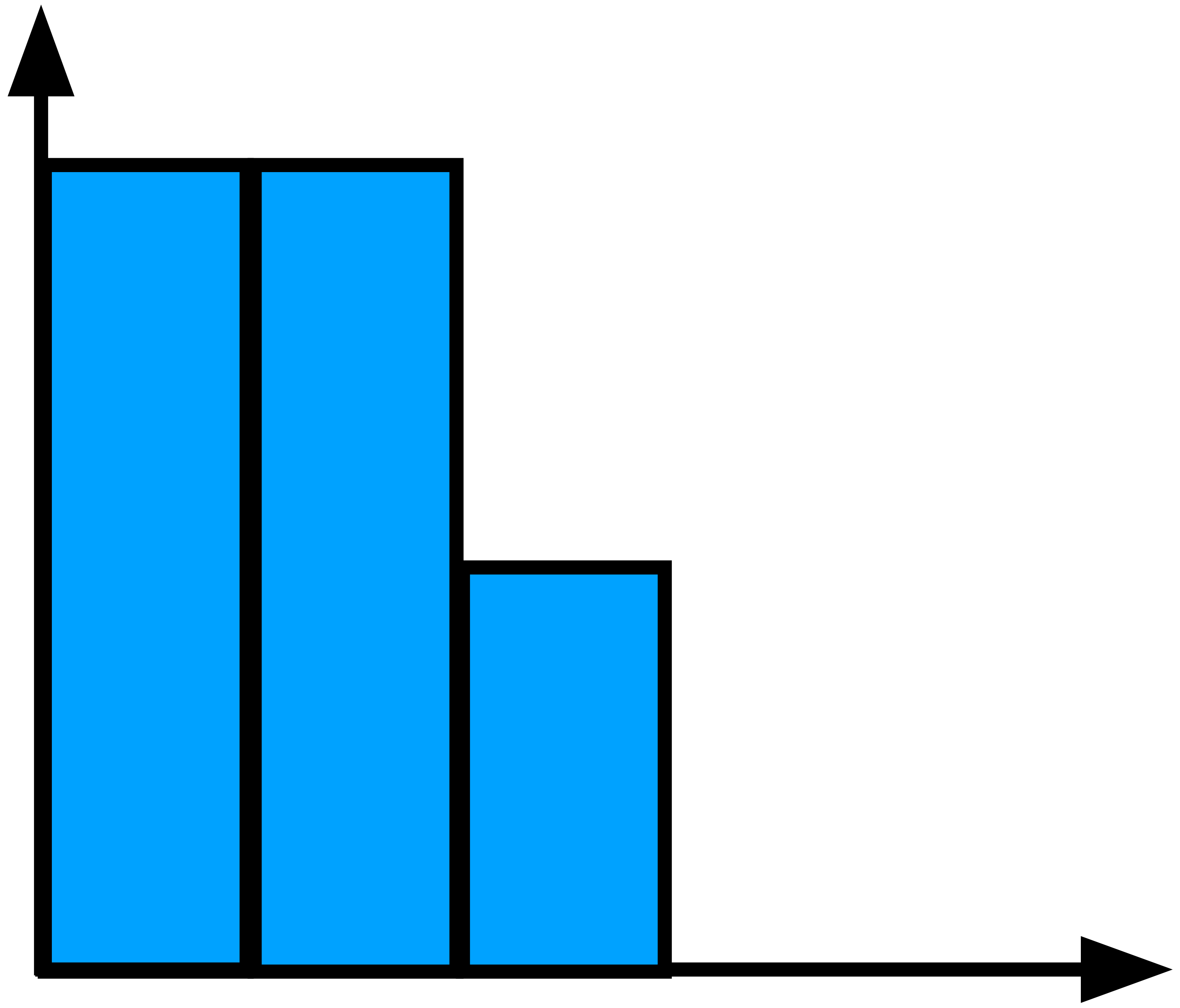
The degree distribution **P(k)** is the proportion of nodes with degree **equal to k**

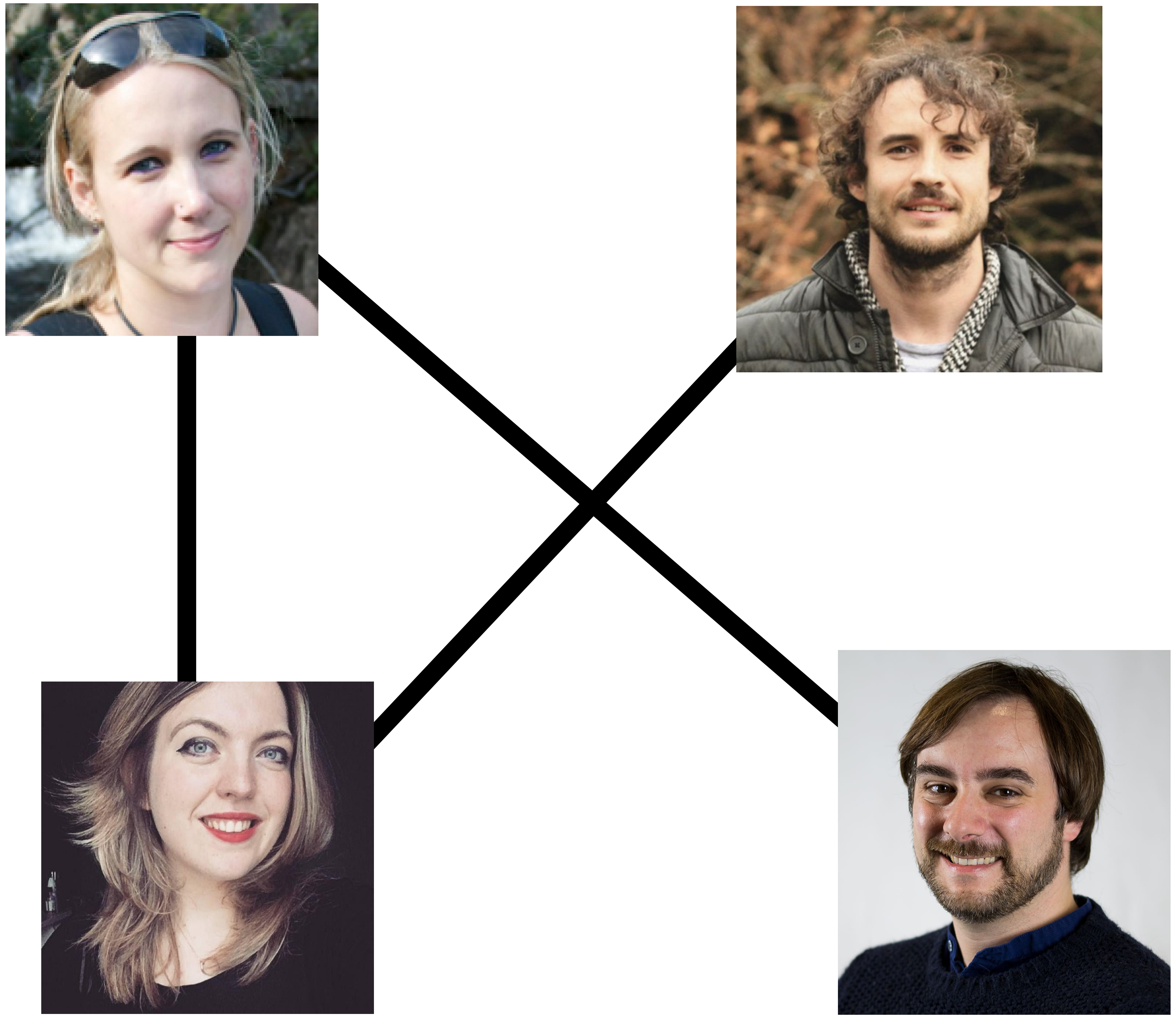




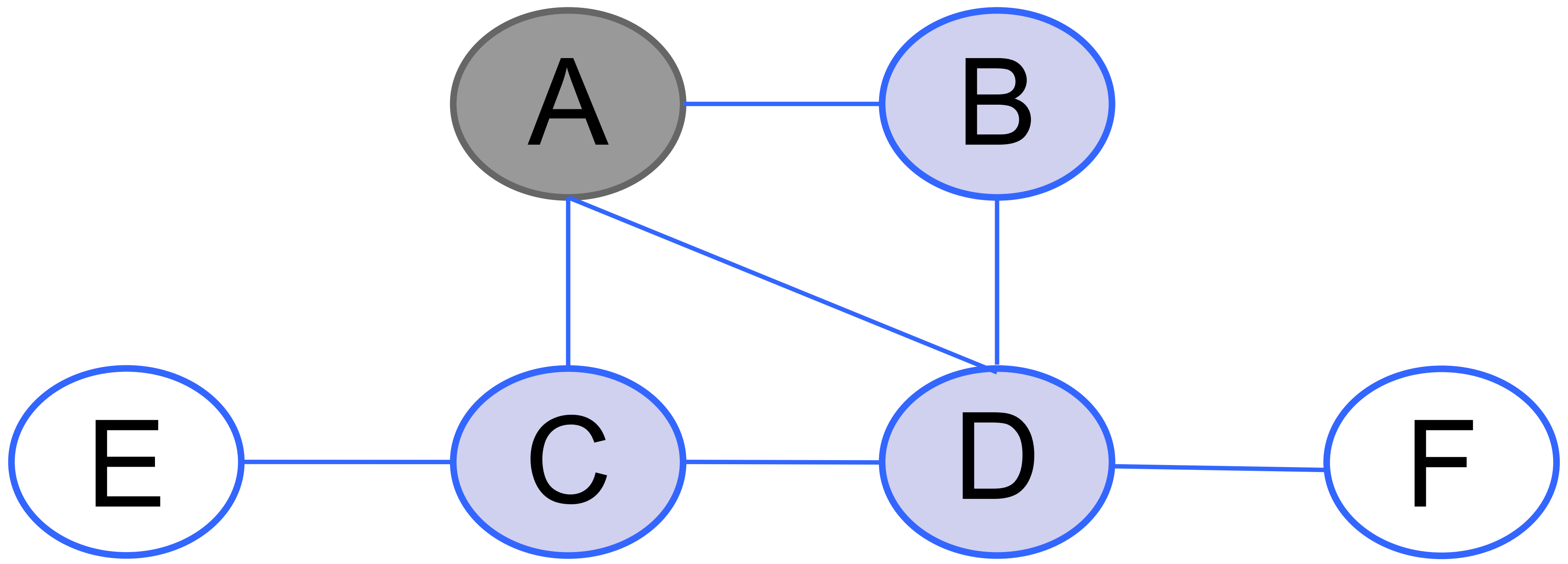
Degree distribution

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





CLUSTERING COEFFICIENT



The **clustering coefficient** defines the proportion of A's neighbours (N(A)) which are connected by an edge (are friends).

The number of triangles in which A is involved wrt to the ones it could be involved in.

FORMALLY: CLUSTERING COEFFICIENT

Local Clustering Coefficient

$$C_i = \frac{2\,|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

Proportion of my friends who are also friends with my other friends...

Network Clustering Coefficient

$$CG = \frac{1}{N}\sum_i C_i$$

The average all all the node's local clustering coefficients

FORMALLY: CLUSTERING COEFFICIENT

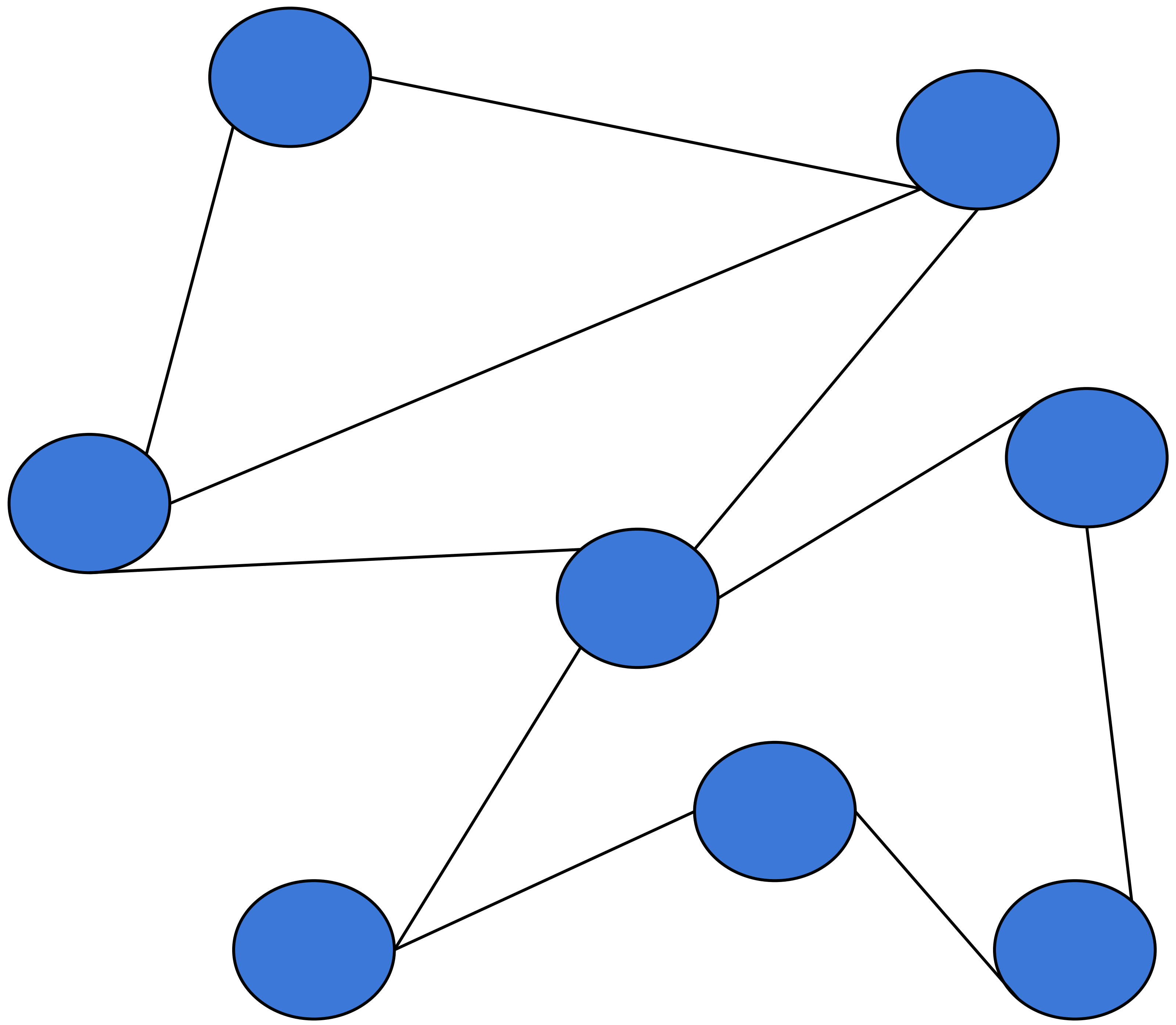
Local Clustering Coefficient

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

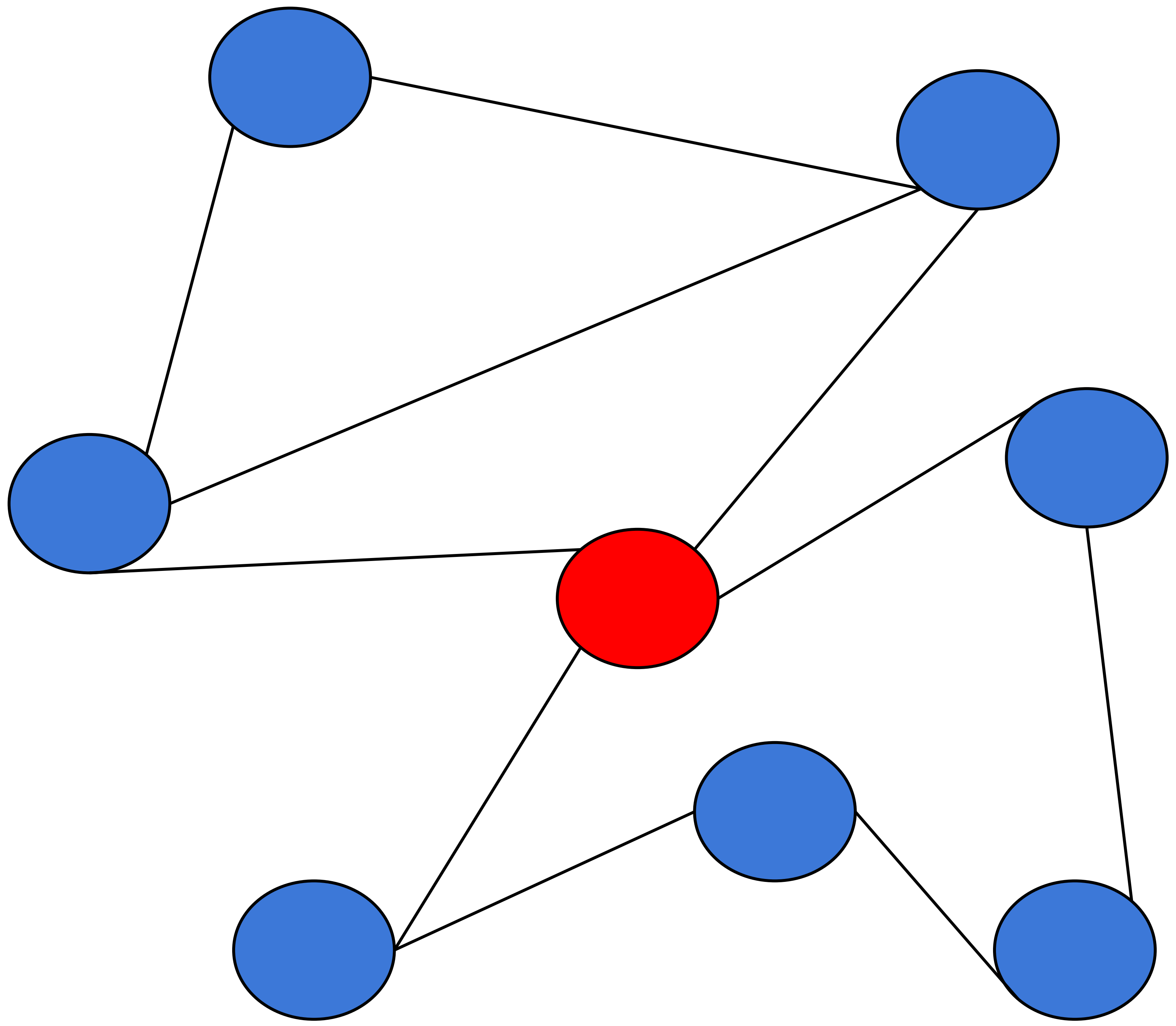
Network Clustering Coefficient

$$CG = \frac{1}{N}\sum_i C_i$$

CLUSTERING COEFFICIENT: EXAMPLE

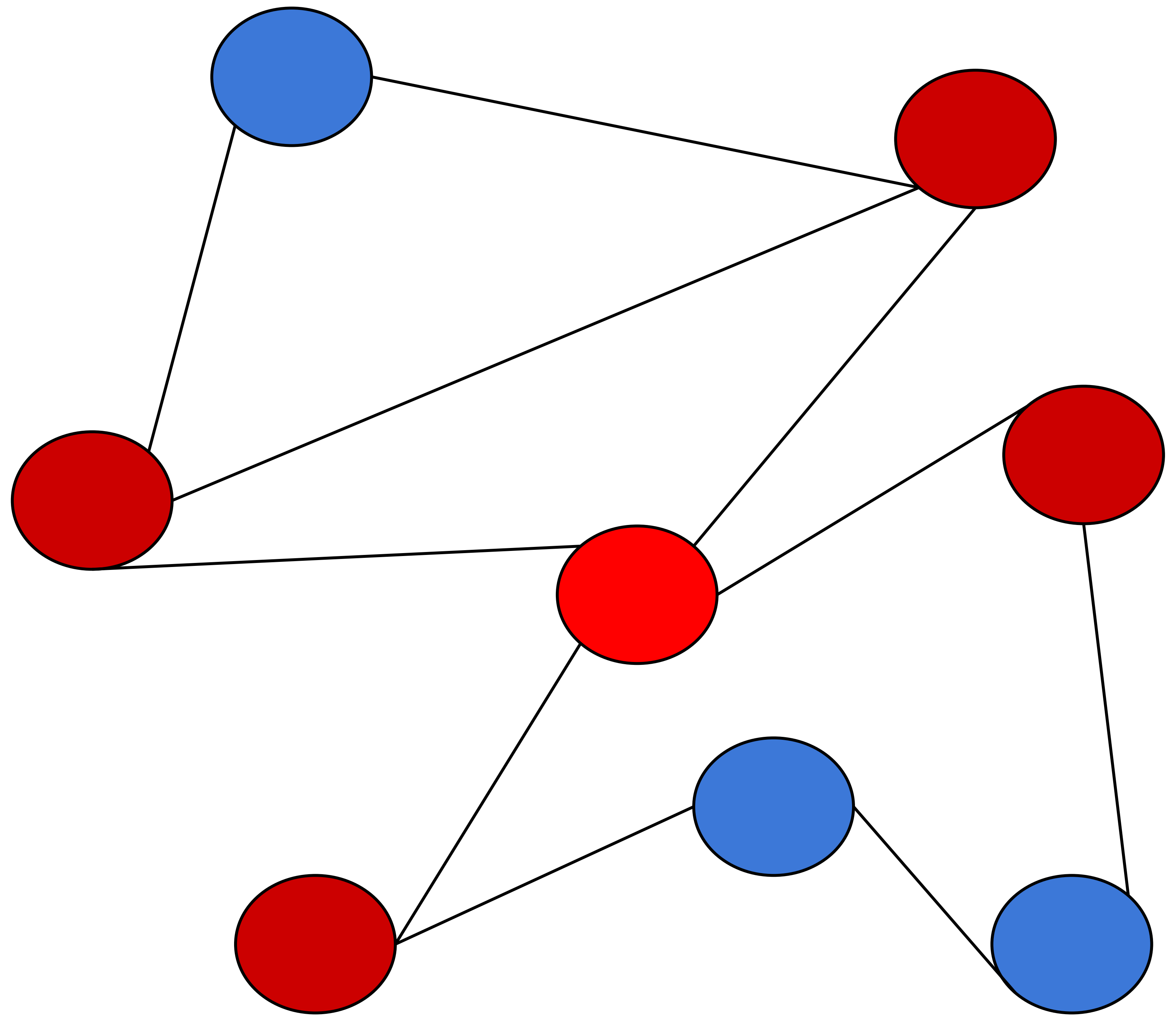


CLUSTERING COEFFICIENT: EXAMPLE



CLUSTERING COEFFICIENT: EXAMPLE

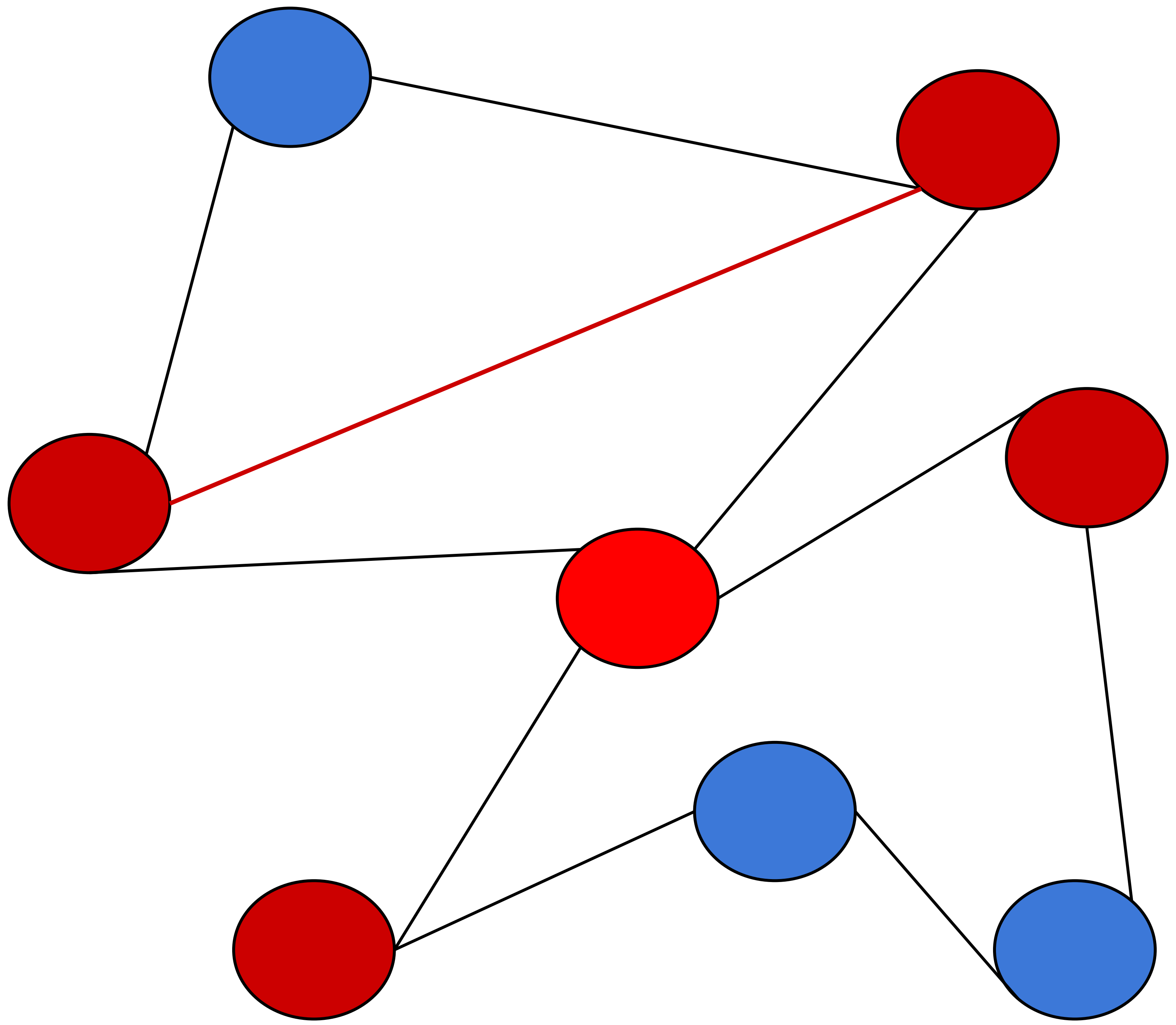
Degree = 4



CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

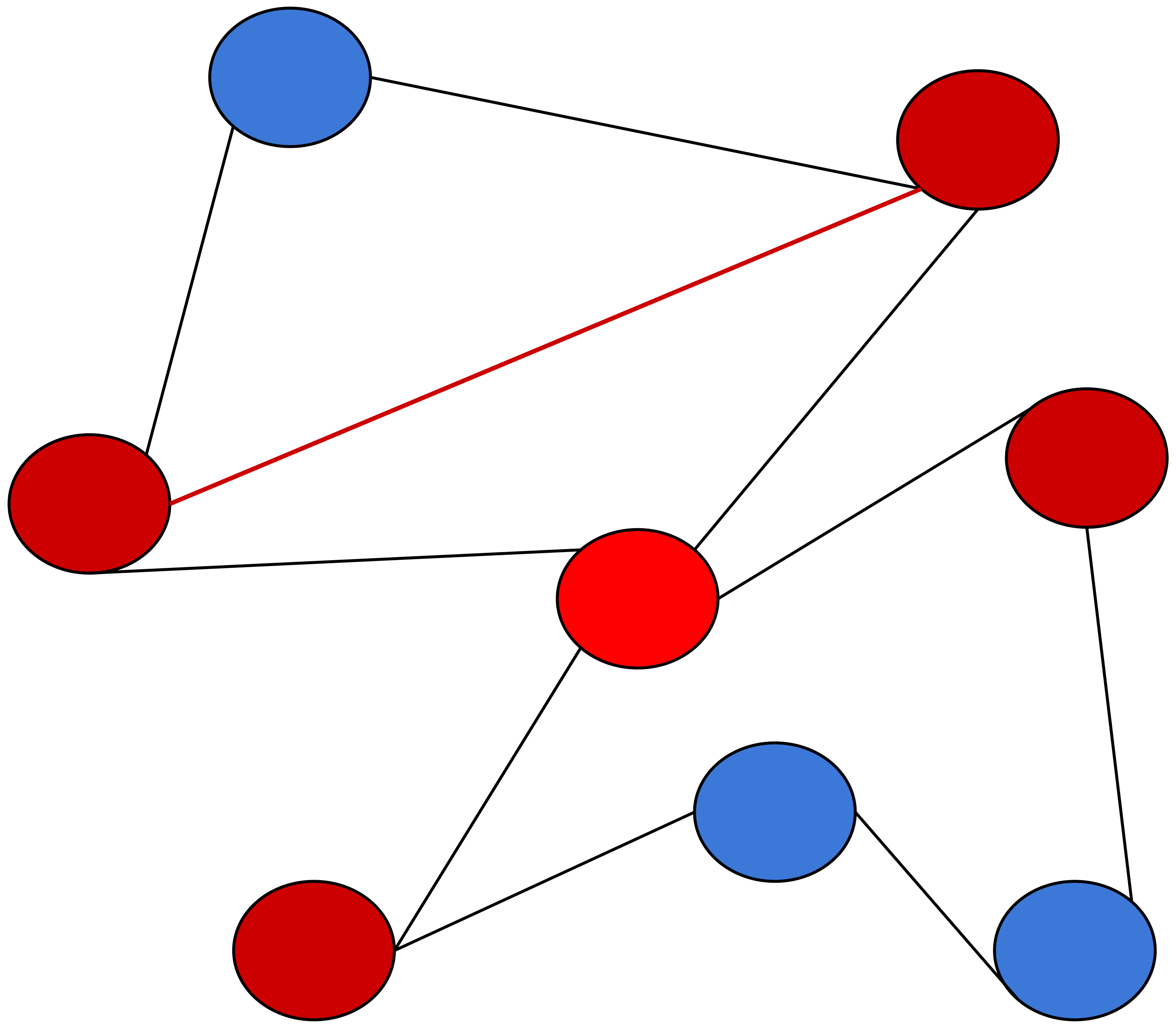


CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_{j,} v_k \in N_i, e_{j,k} \in E$$

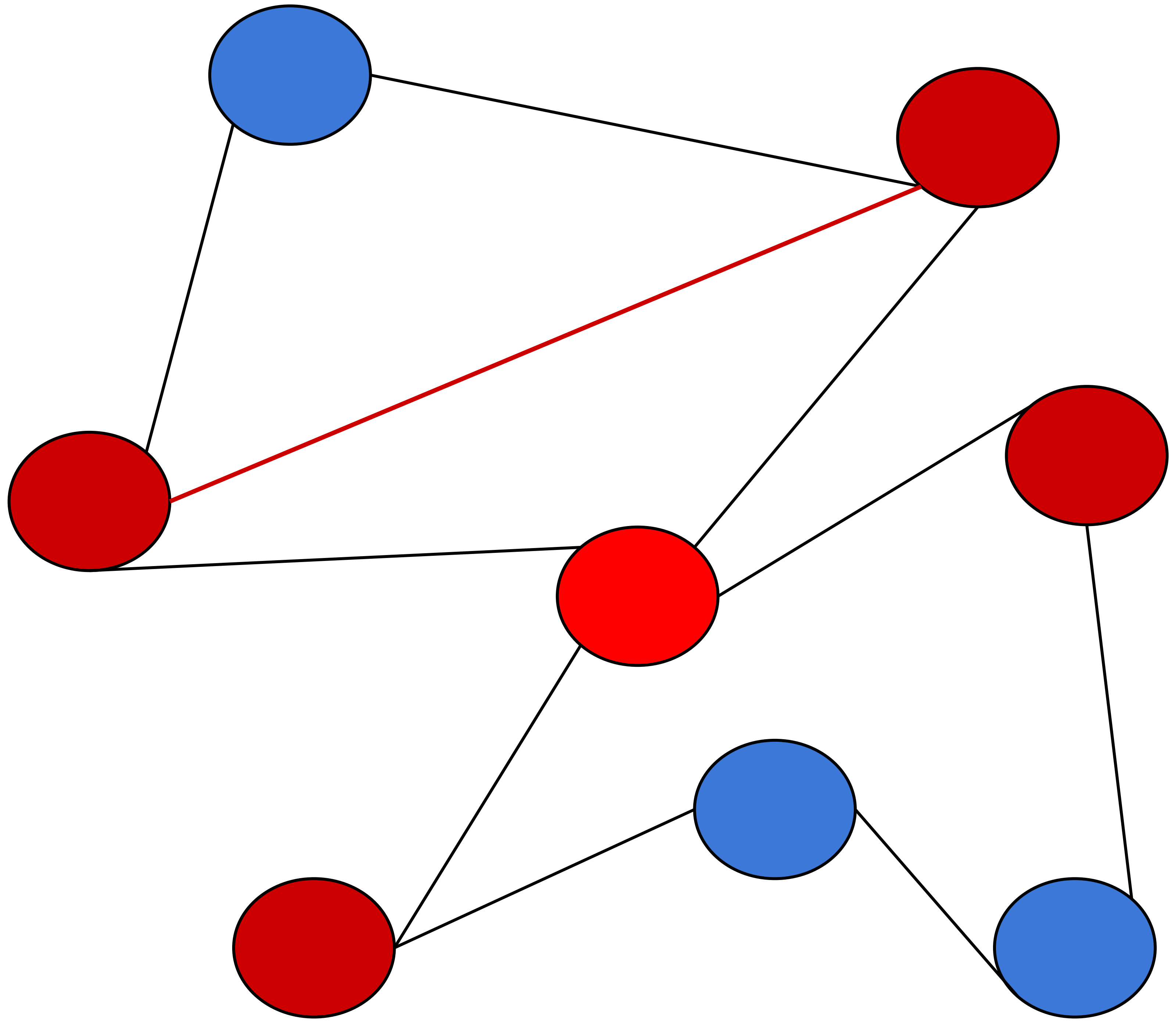


CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_{j,} v_k \in N_i, e_{j,k} \in E$$

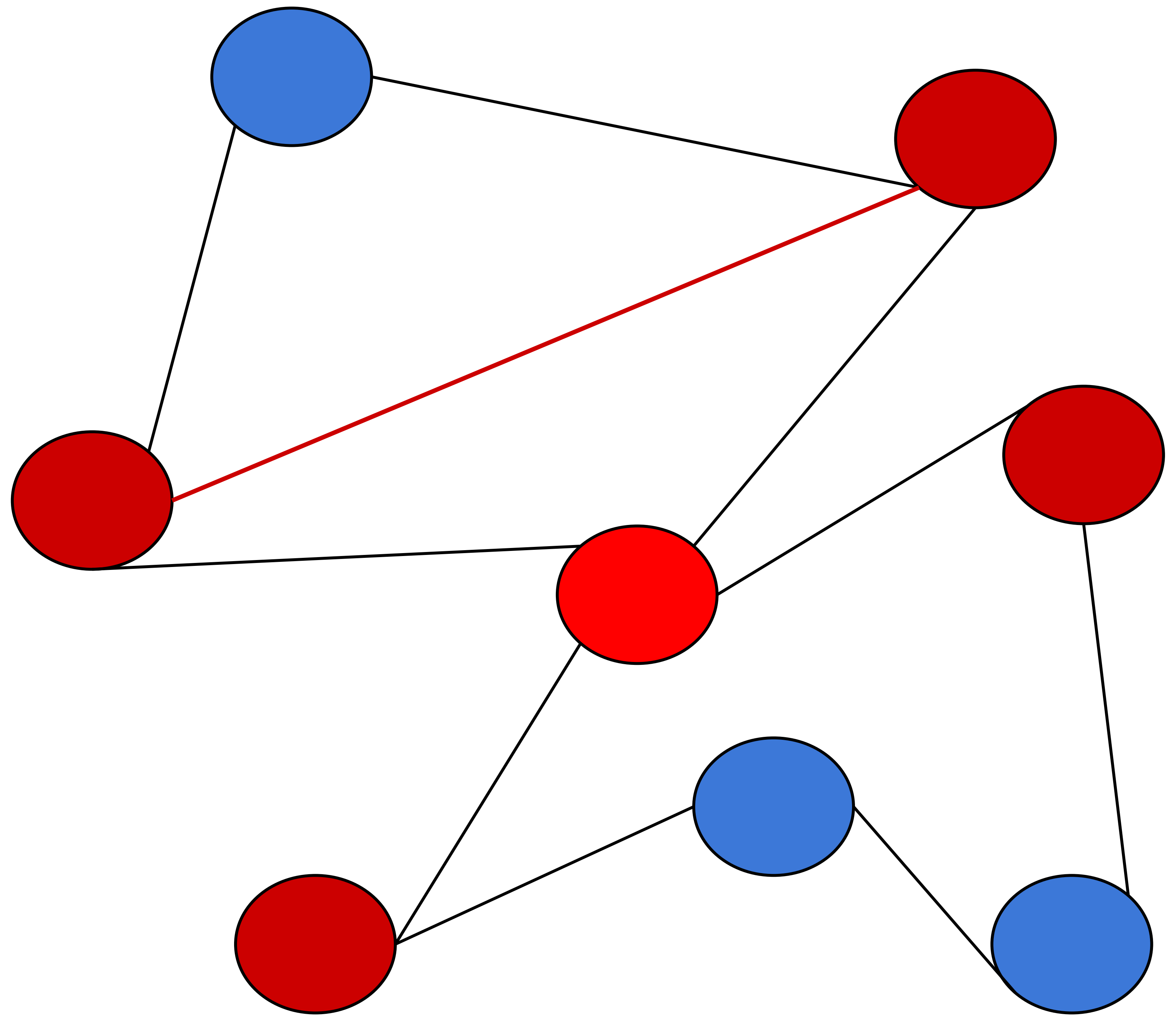


CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_{j,} v_k \in N_i, e_{j,k} \in E$$



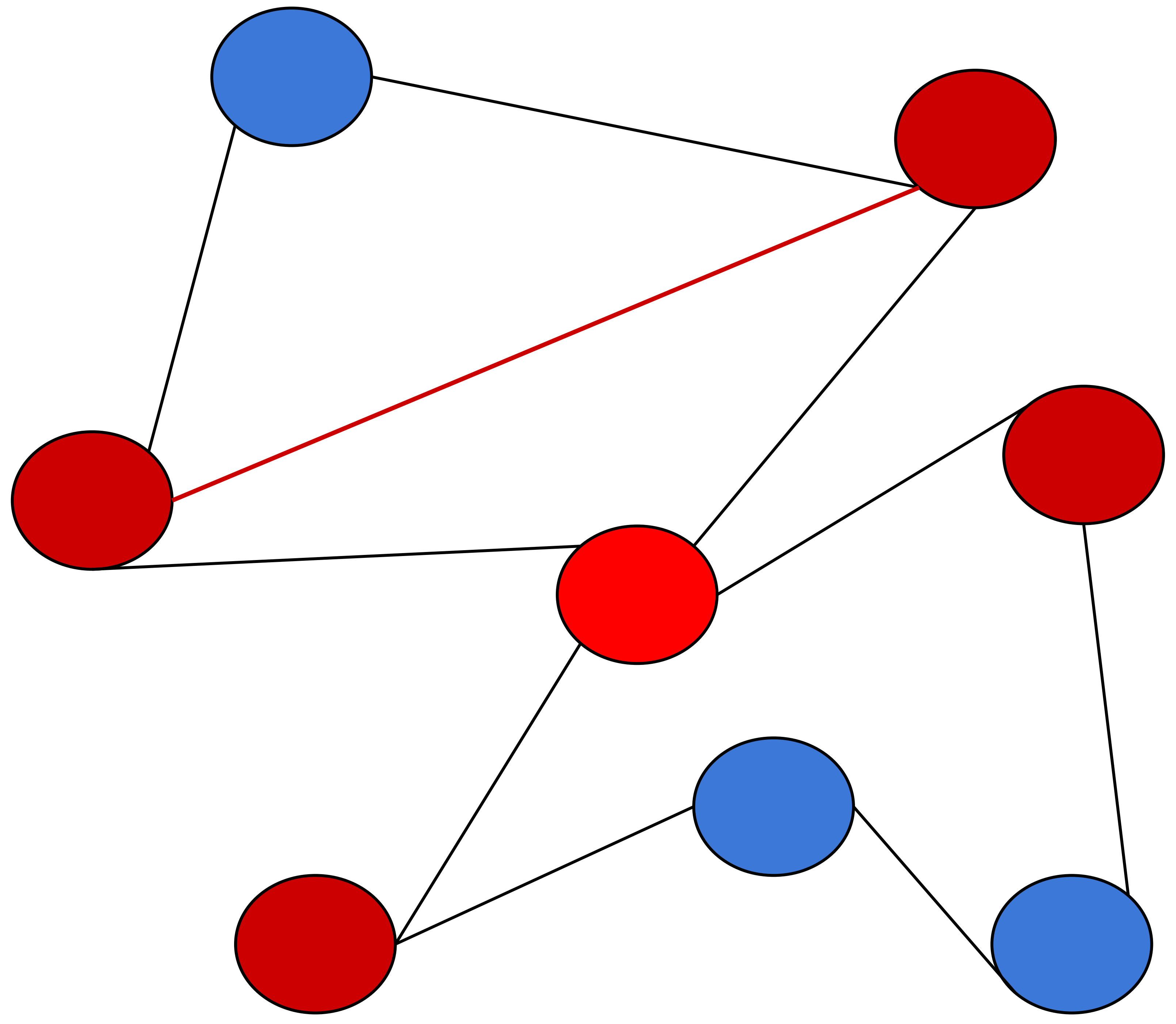
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2\,|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

$C_i = \frac{2}{}$



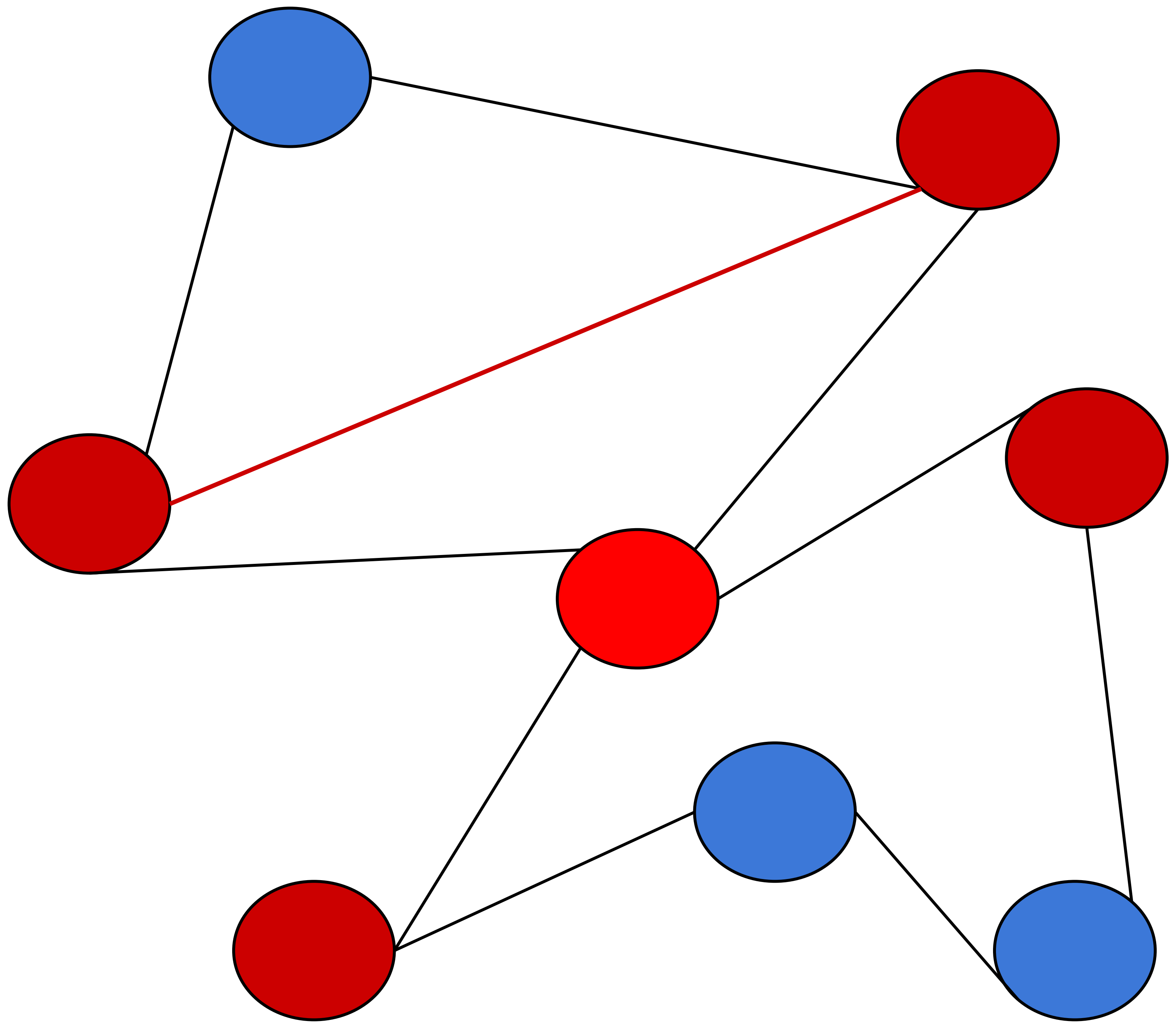
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2\,|\boxed{\{e_{jk}\}}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

$$C_i = \frac{2 * 1}{}$$



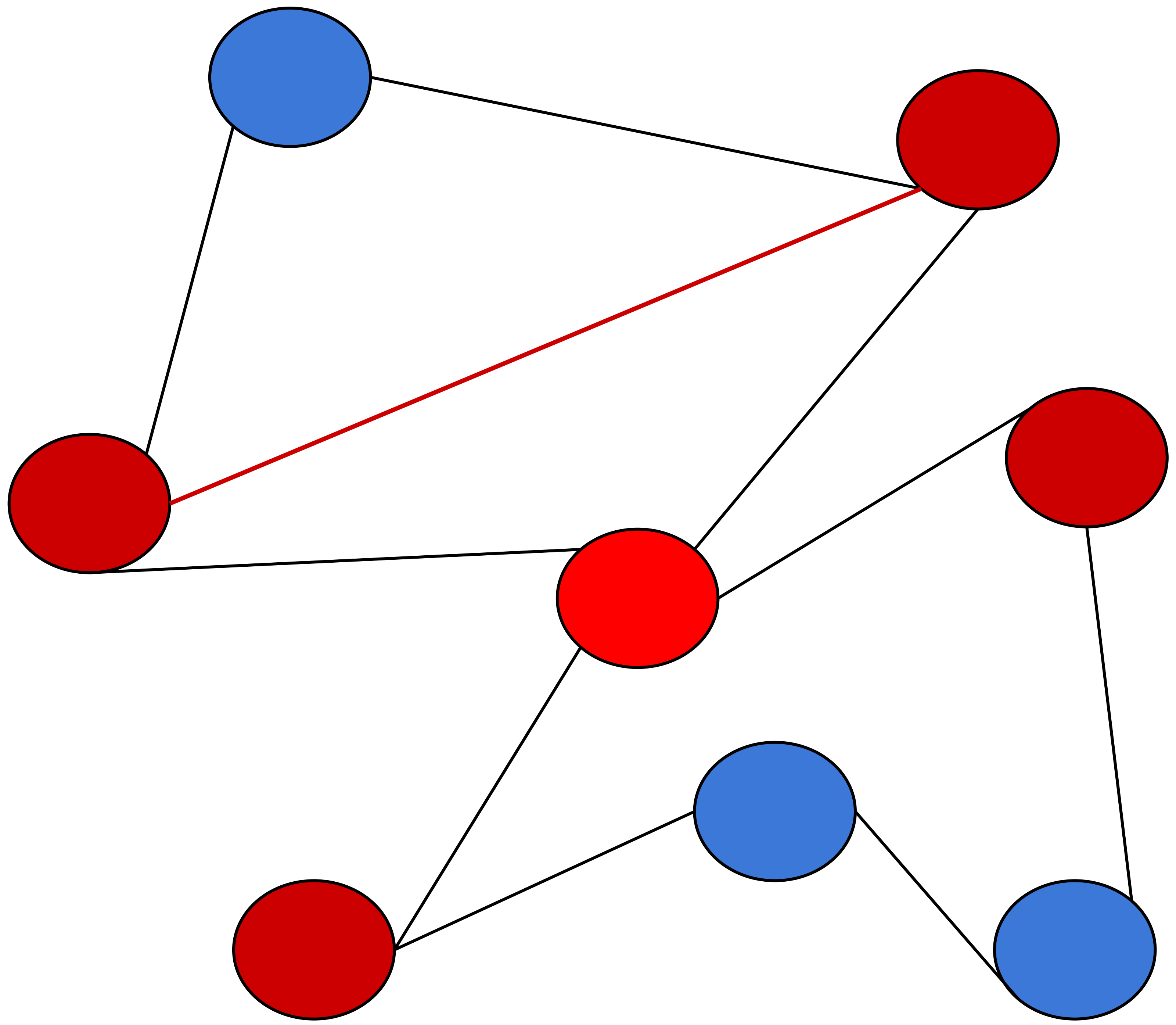
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

$$C_i = \frac{2 * 1}{4}$$



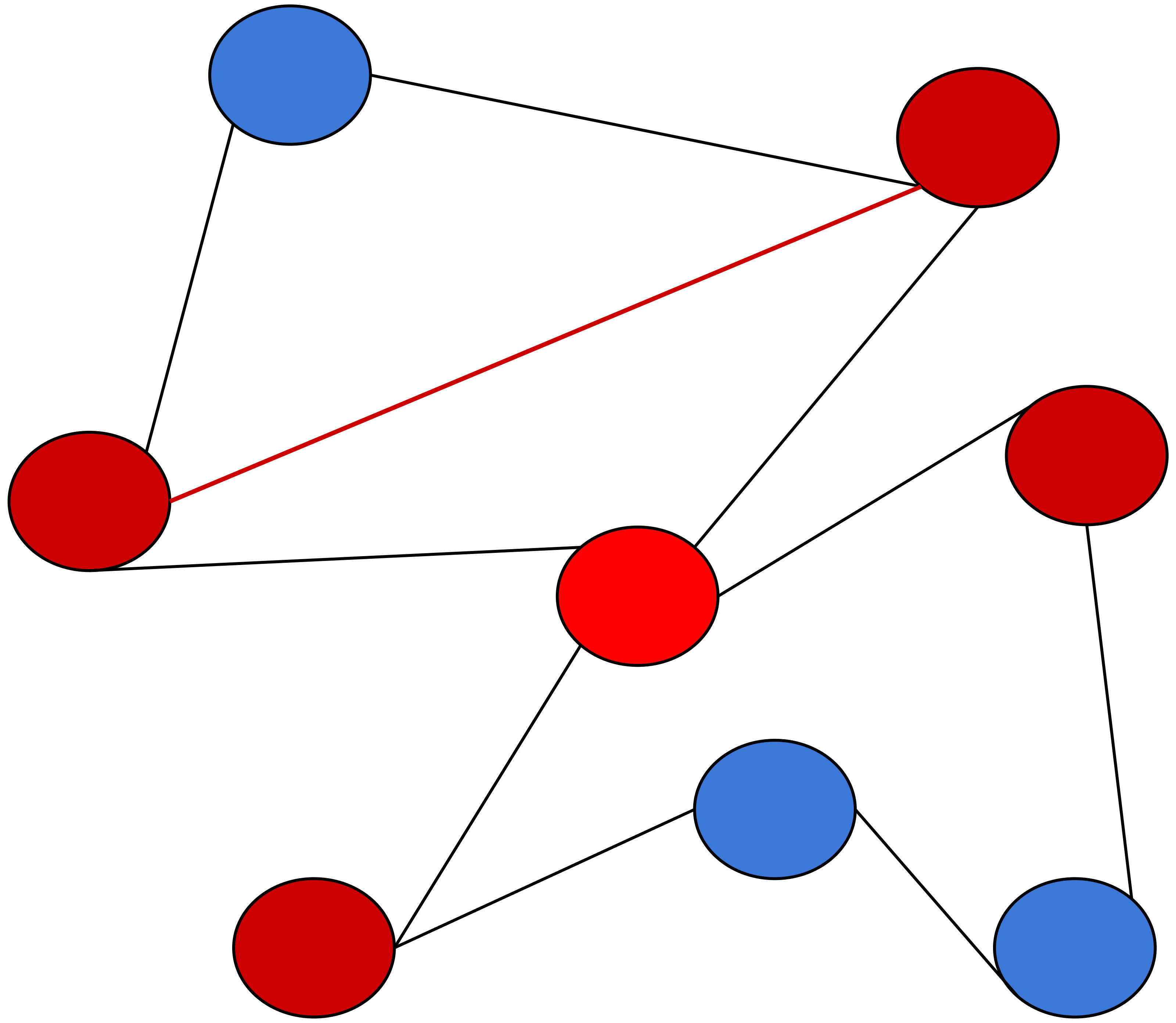
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i\boxed{(k_i - 1)}} : v_j, v_k \in N_i, e_{j,k} \in E$$

$$C_i = \frac{2 * 1}{4 * 3}$$



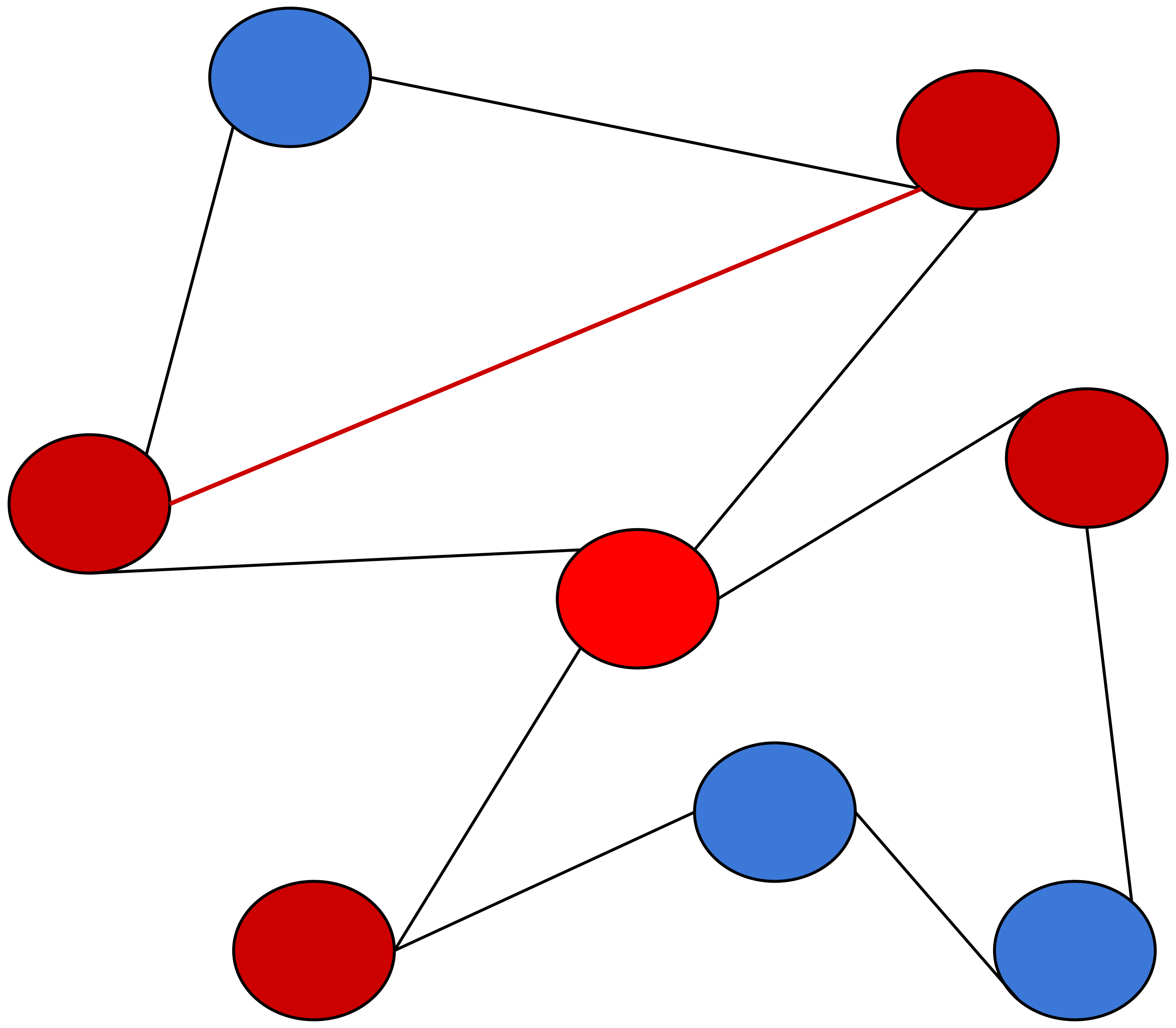
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

$$C_i = \frac{2}{12}$$



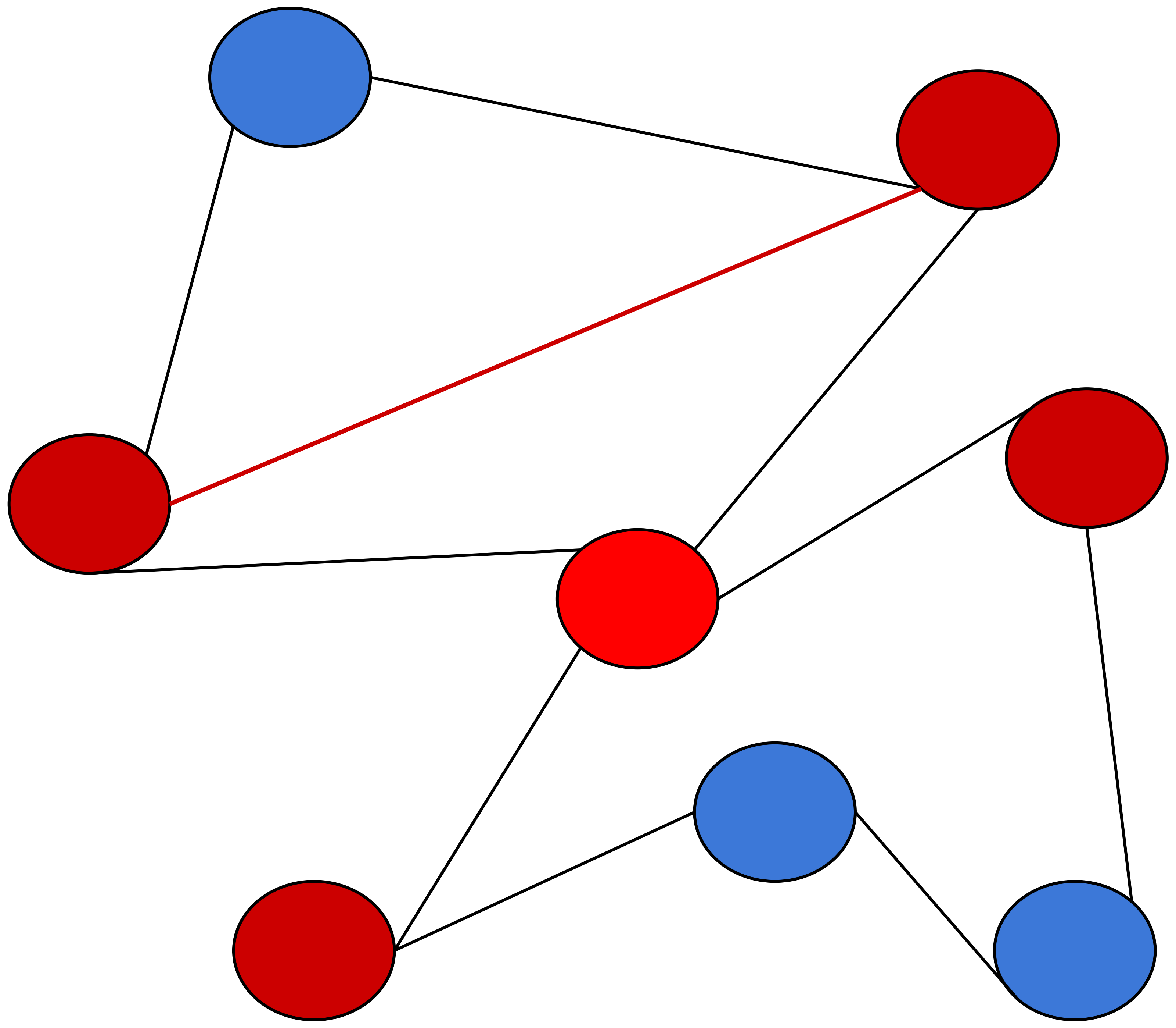
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

$$C_i = \frac{1}{6}$$



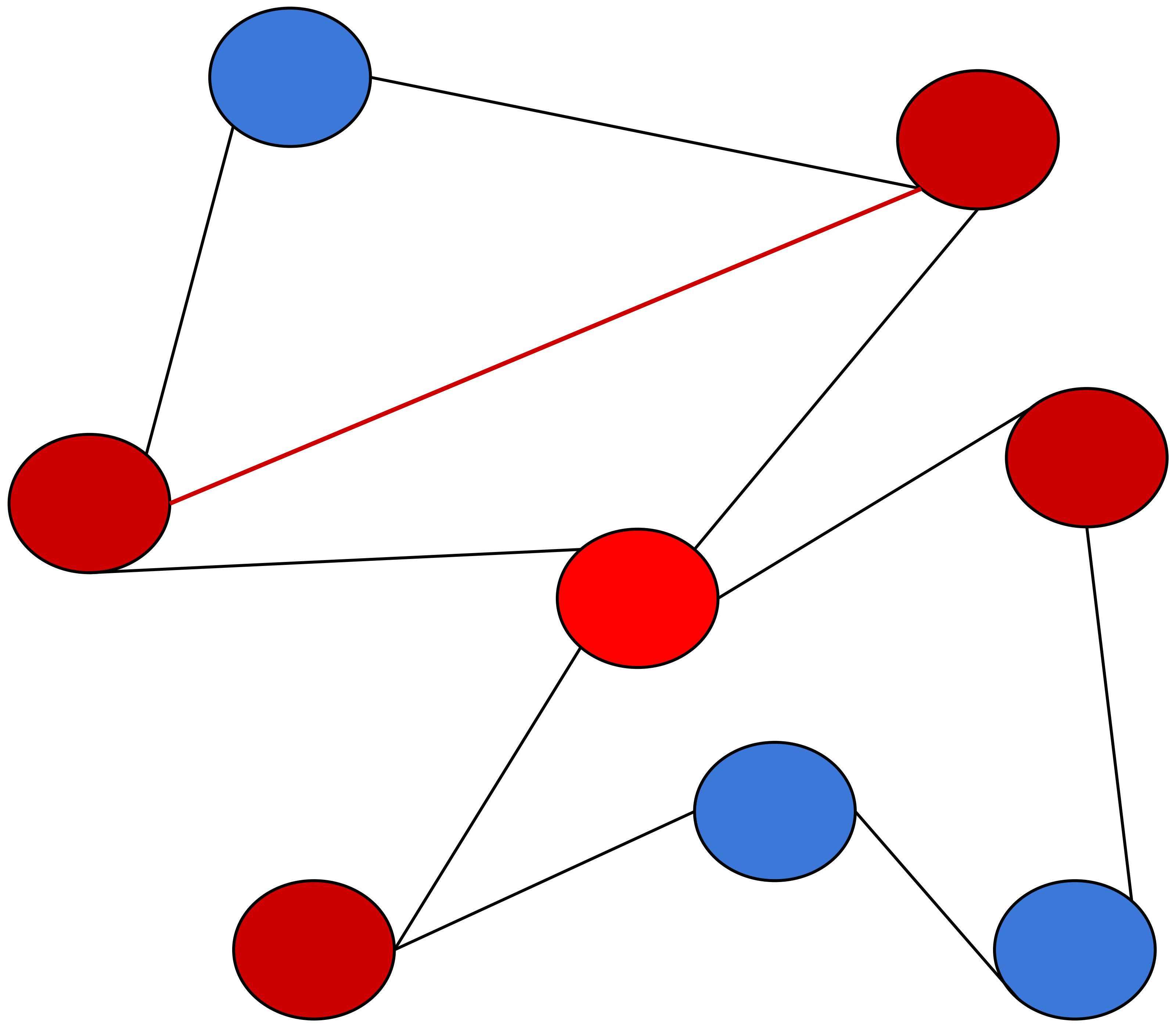
CLUSTERING COEFFICIENT: EXAMPLE

Degree = 4

Links between neighbours = 1

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

$$C_i = \frac{1}{6}$$



Fraction of possible interconnections between my neighbour!

Clustering Coefficient

Proportion of possible interconnections between neighbours

Node clustering coefficient Ci

$$C_i = \frac{2\,|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{j,k} \in E$$

Special case:

if k(v) = 1 or 0,

C(v) = 0

Clustering Coefficient

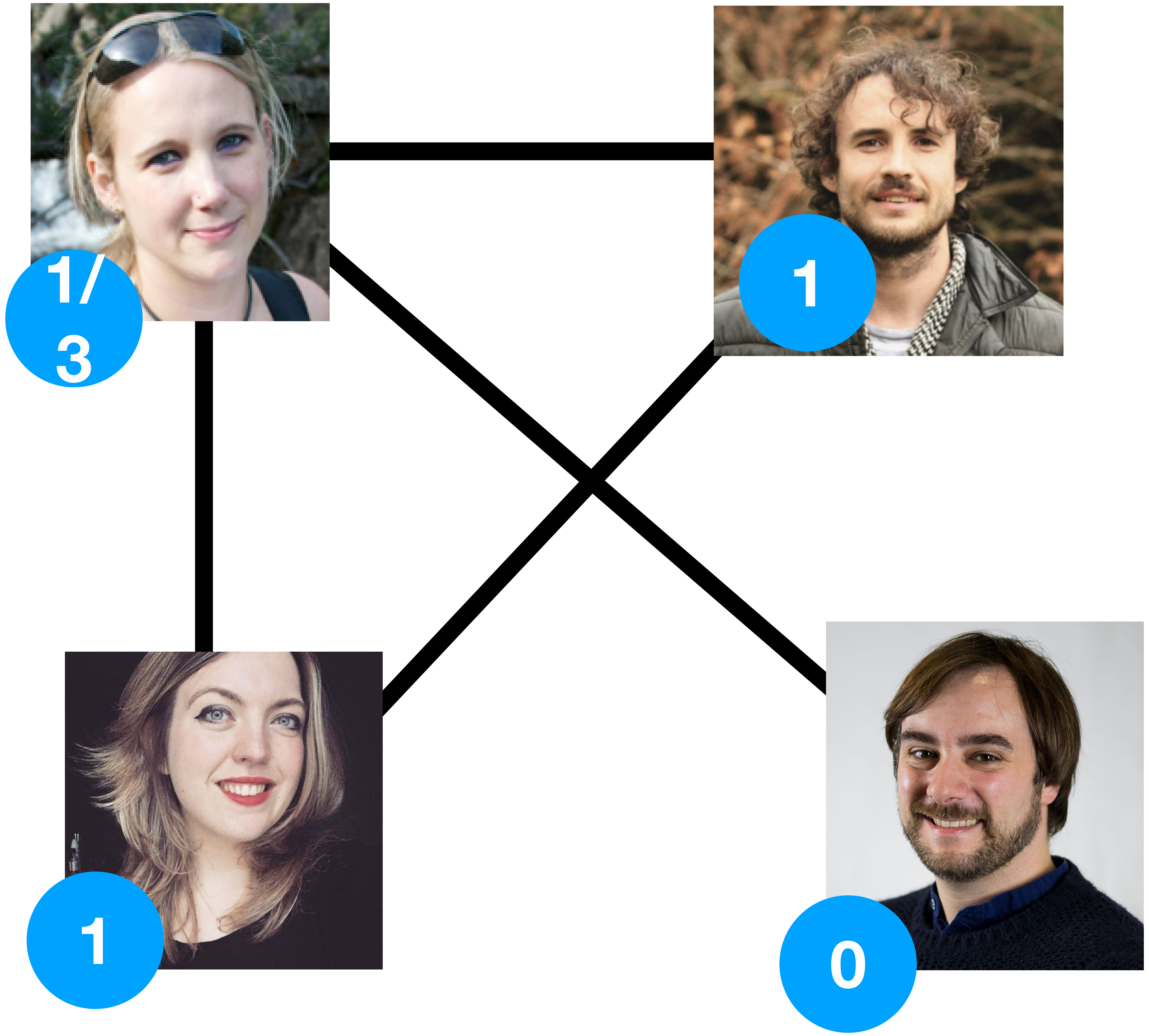
What is Laurissa's clustering coefficient?

Numerator: Only <u>one</u> pair of Laurissa's neighbours are connected (Naomi, Teo), so <u>2*1</u>

Denominator: Laurissa's degree is 3, so 3*2 <u>= 6</u>

So C(Laurissa) = <u>2/6 = 1/3</u>

Average clustering C(G) = 7/12



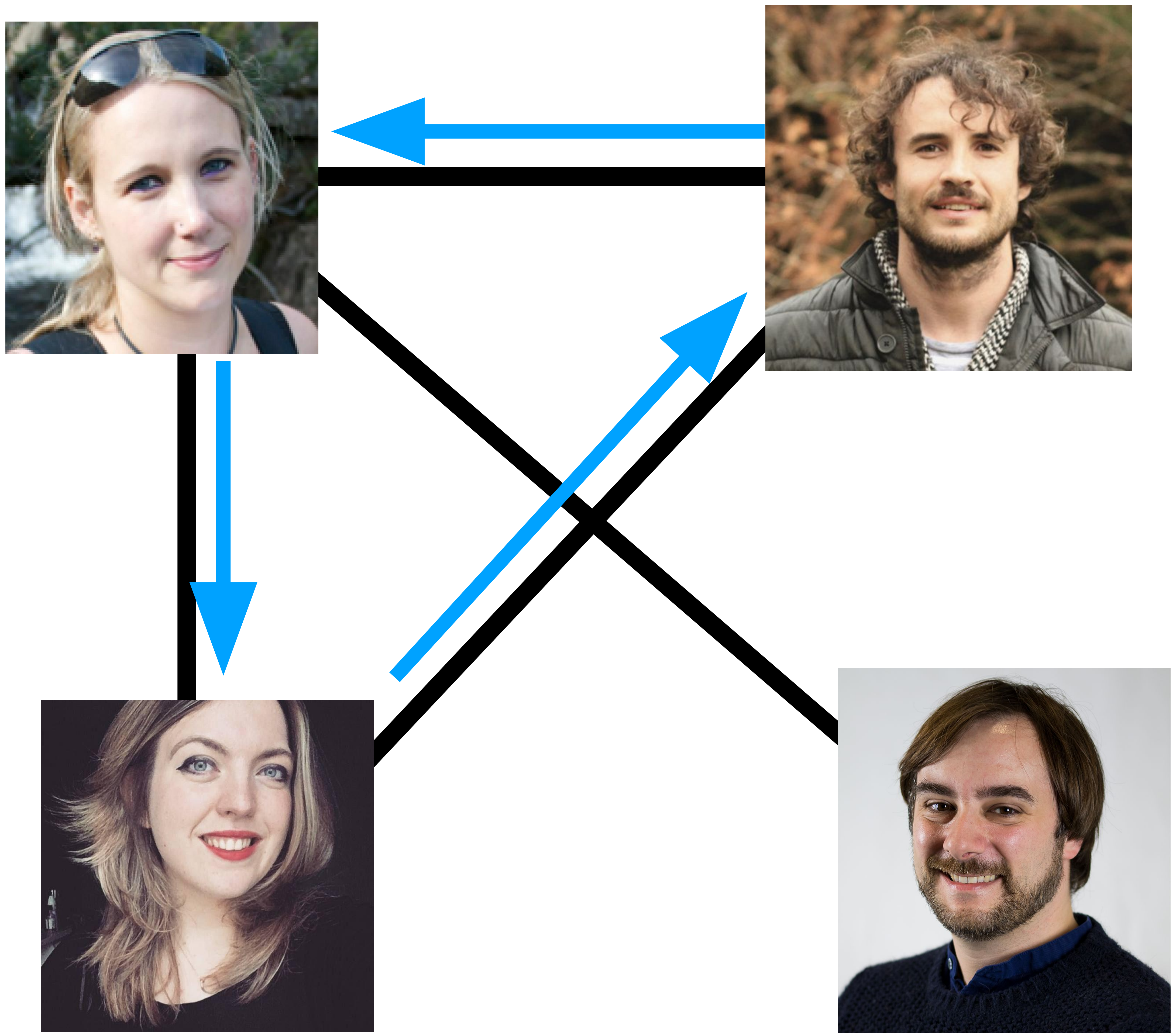
Paths and Cycles

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

Teo, Laurissa, Naomi

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

Teo, Laurissa, Naomi, Teo



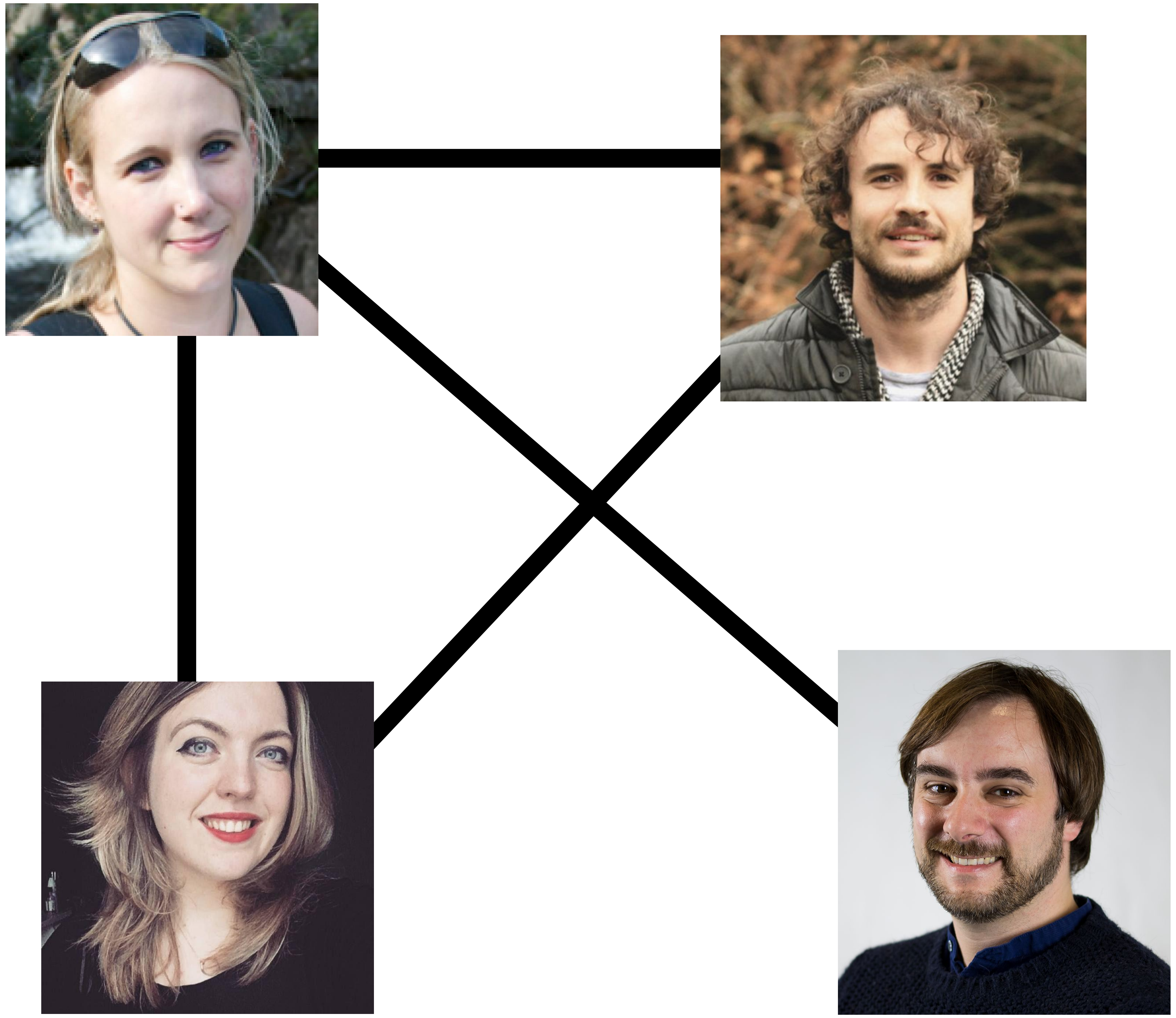
Paths and Cycles

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

d(Teo, Mathieu) = 2

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

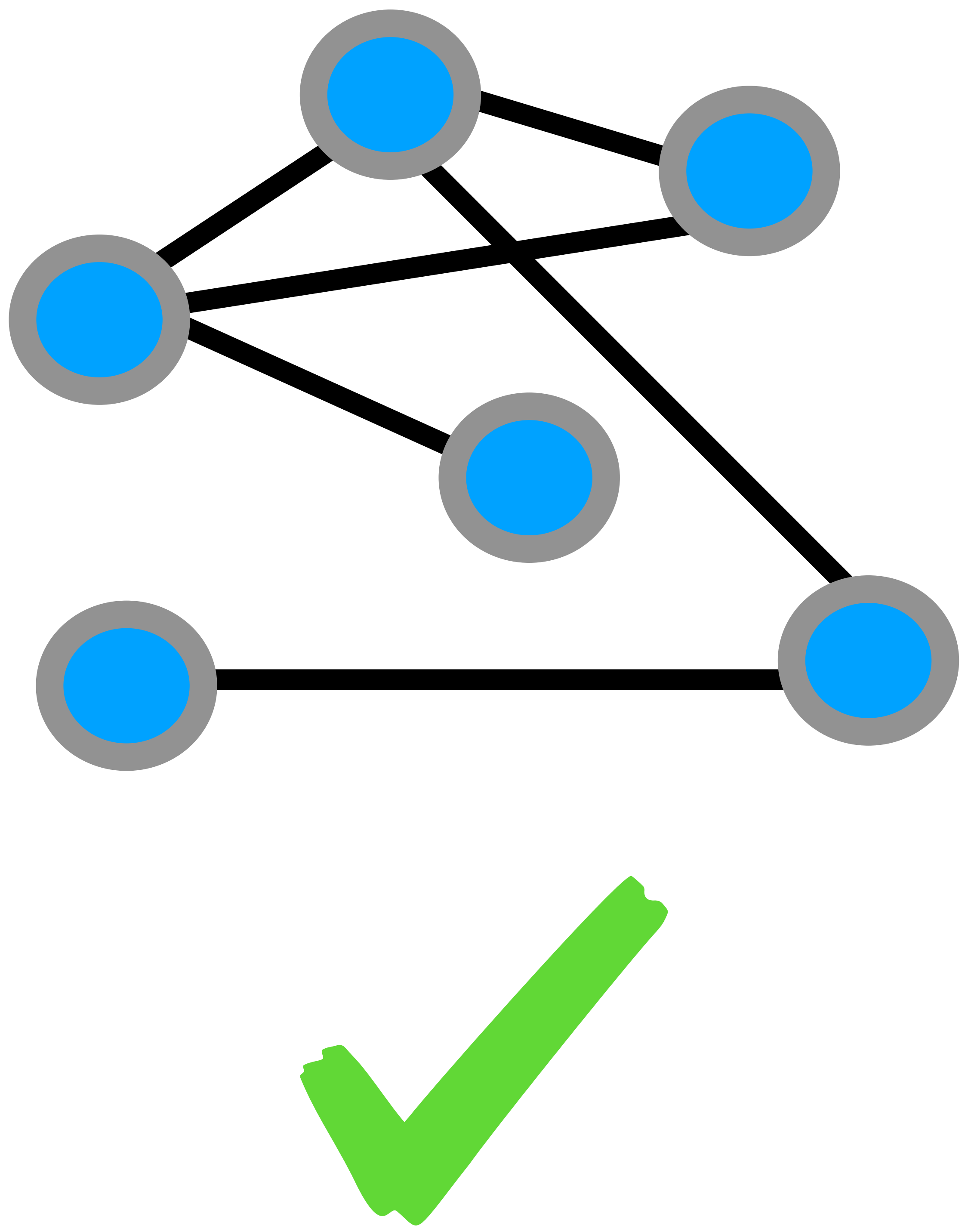
d(G) = 2

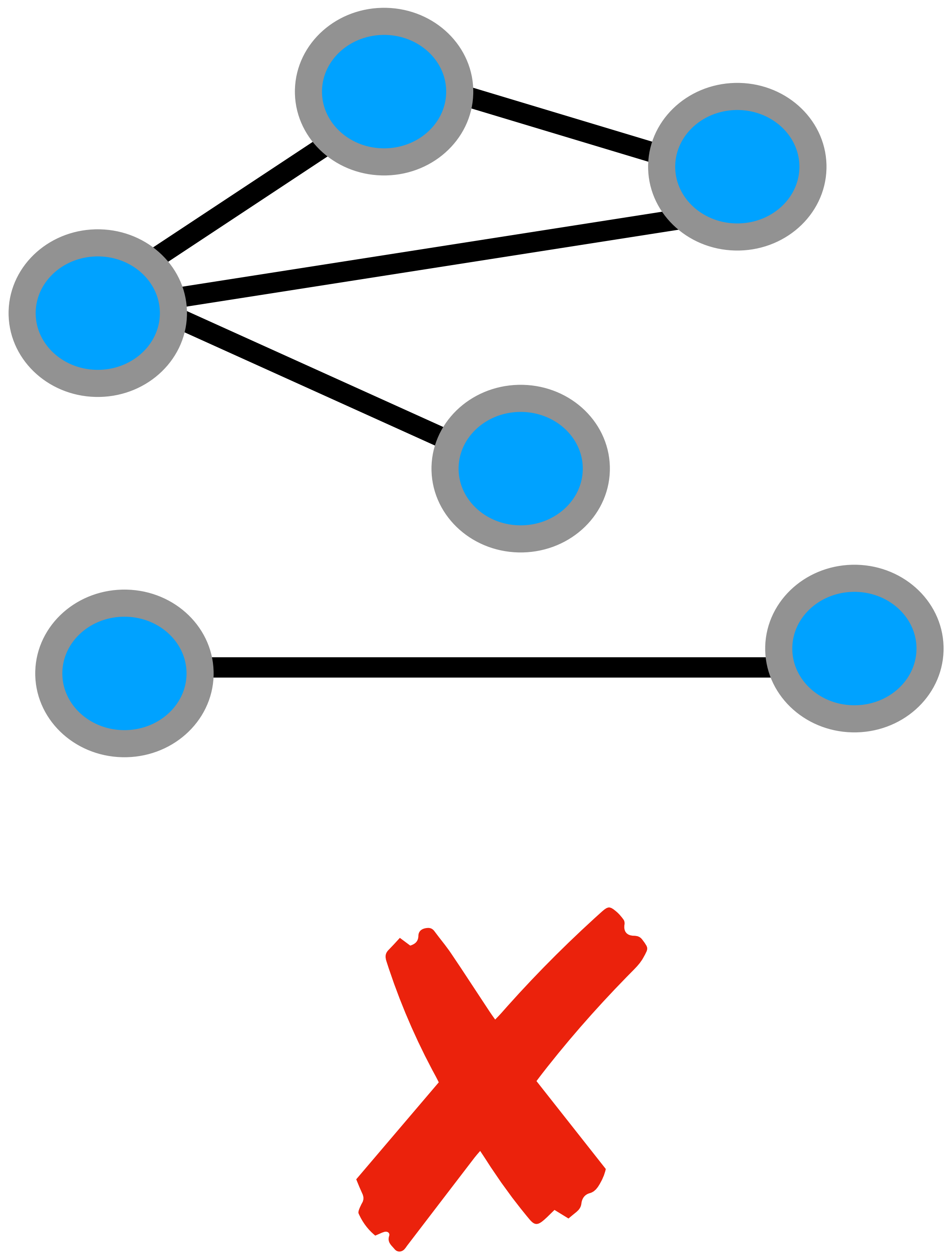


Often more meaningful to look at average path length

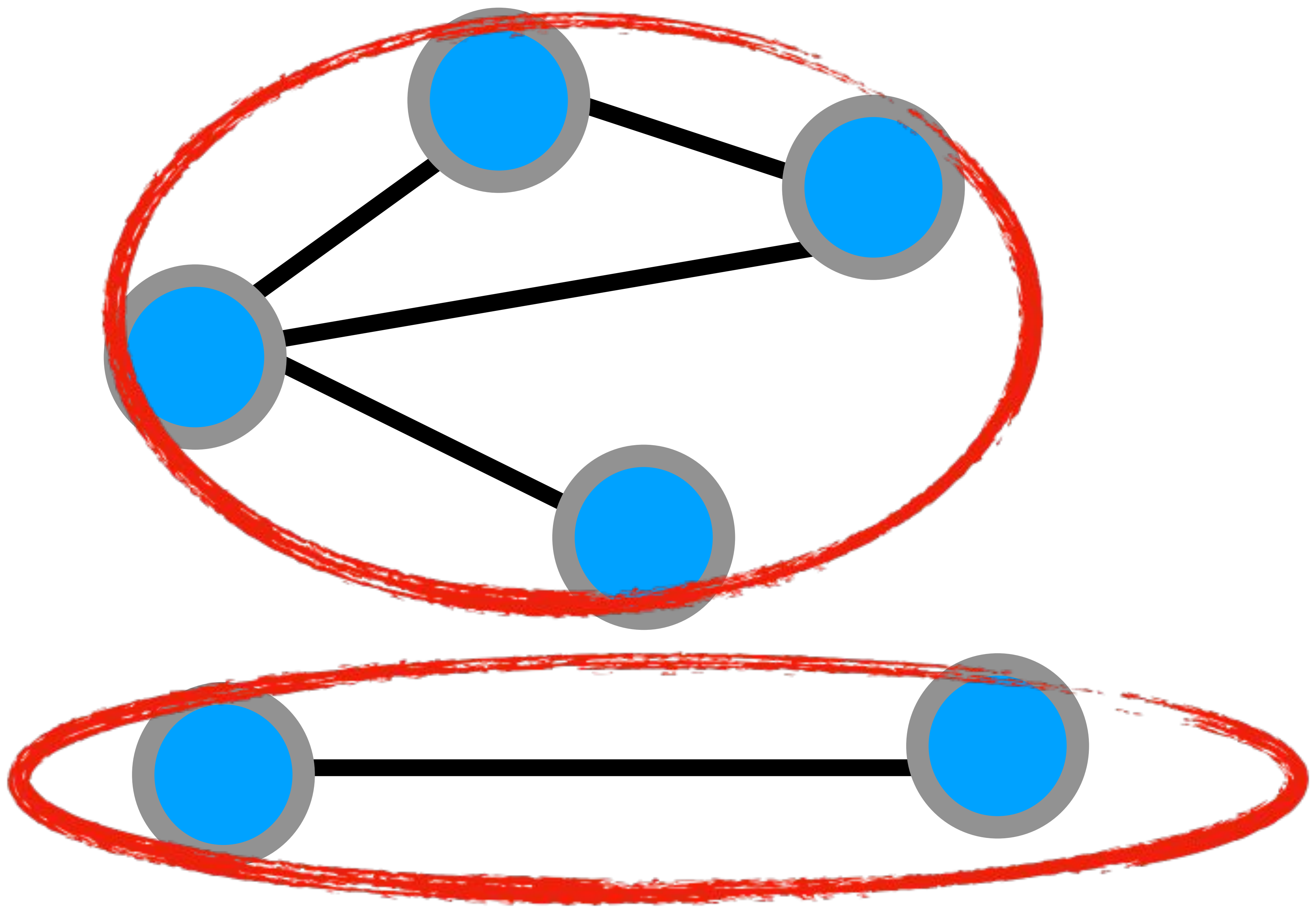
Connected Graph

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?

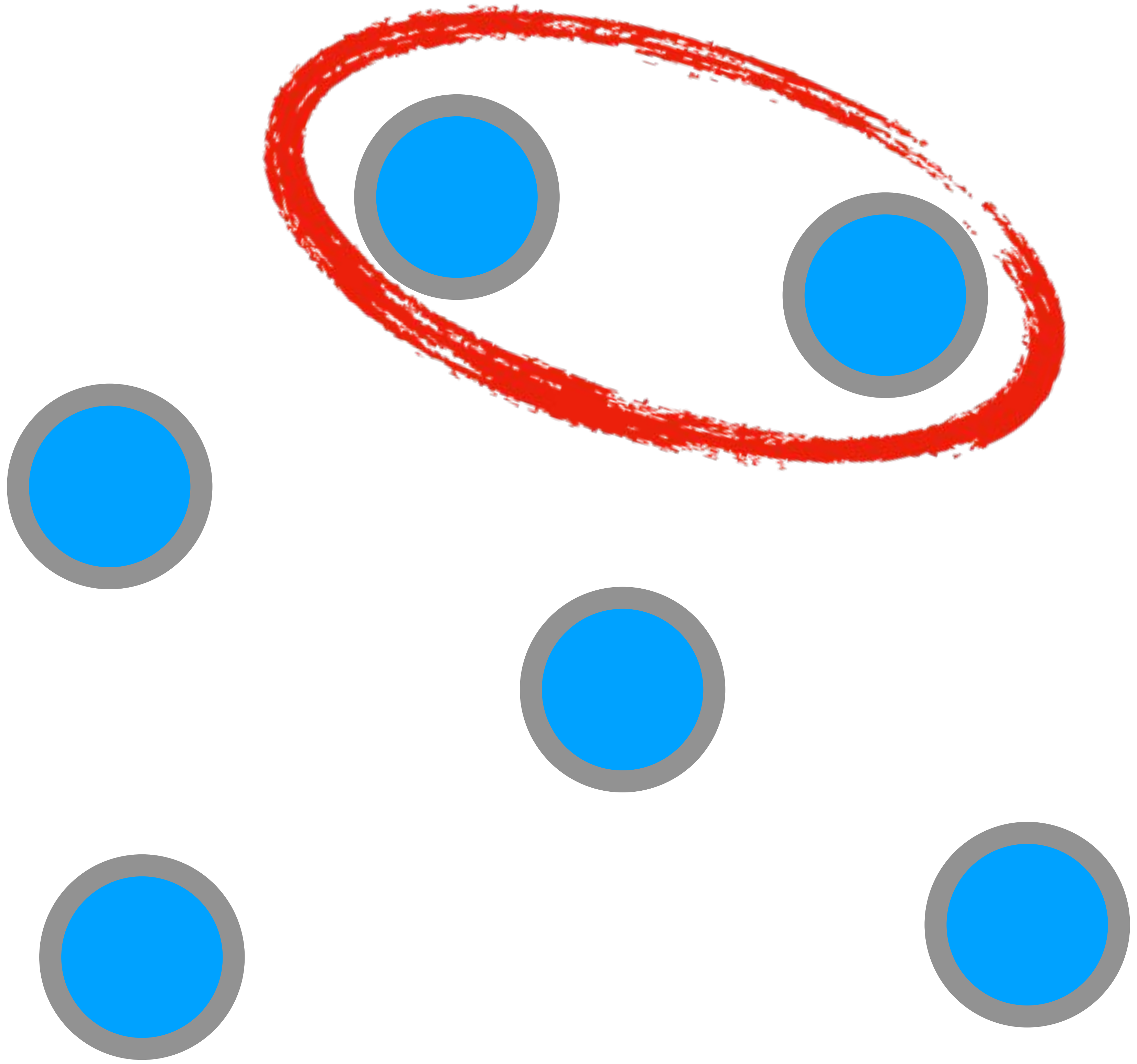
Erdos-Renyi Random Graph Model

- 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**

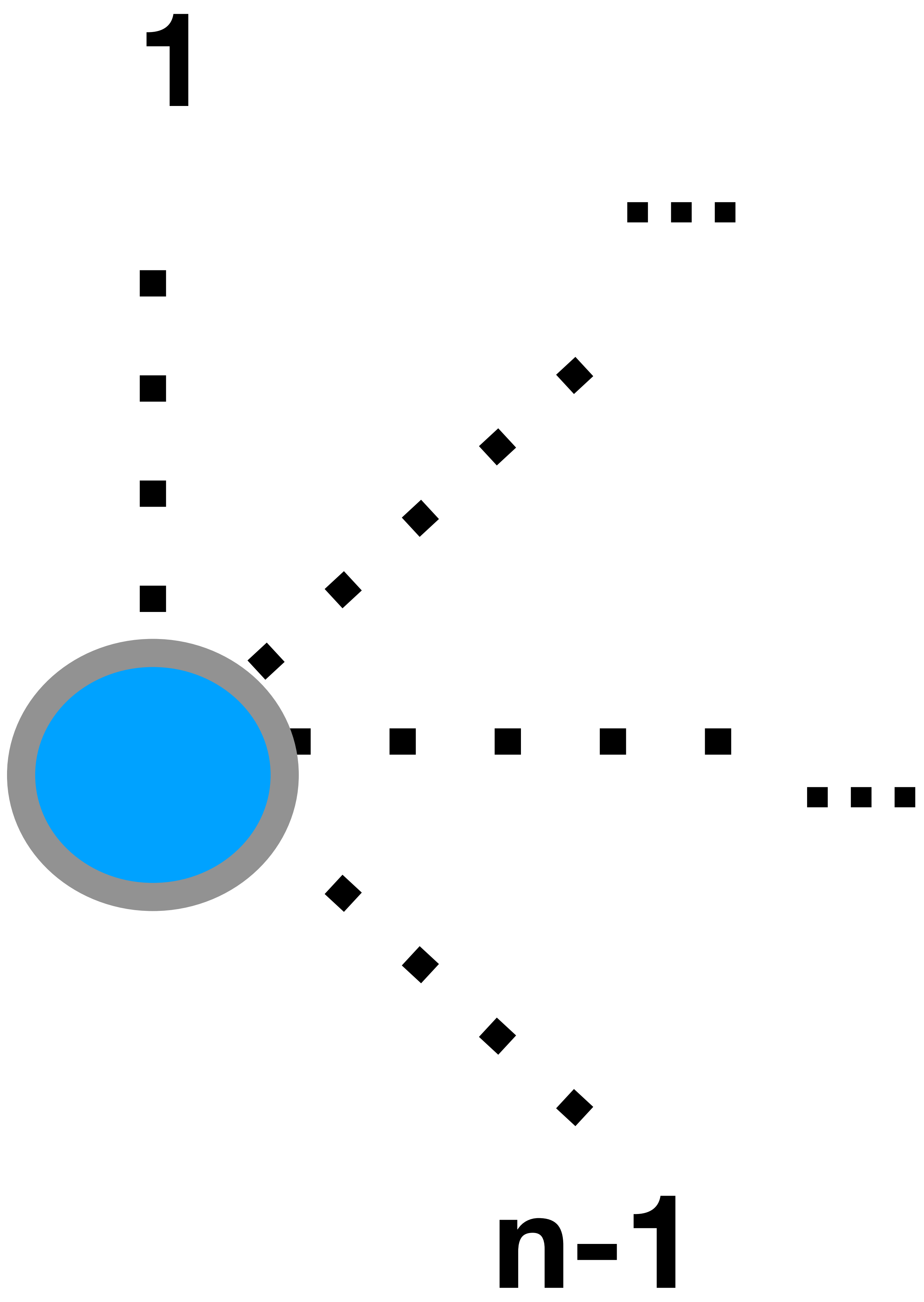
3. For each pair of nodes, do a coin toss. If heads, draw an edge between them. If not, move on.

Erdos-Renyi G(n,p) model

p=0 **p=1**

Increasing **p**

Average degree of ER networks



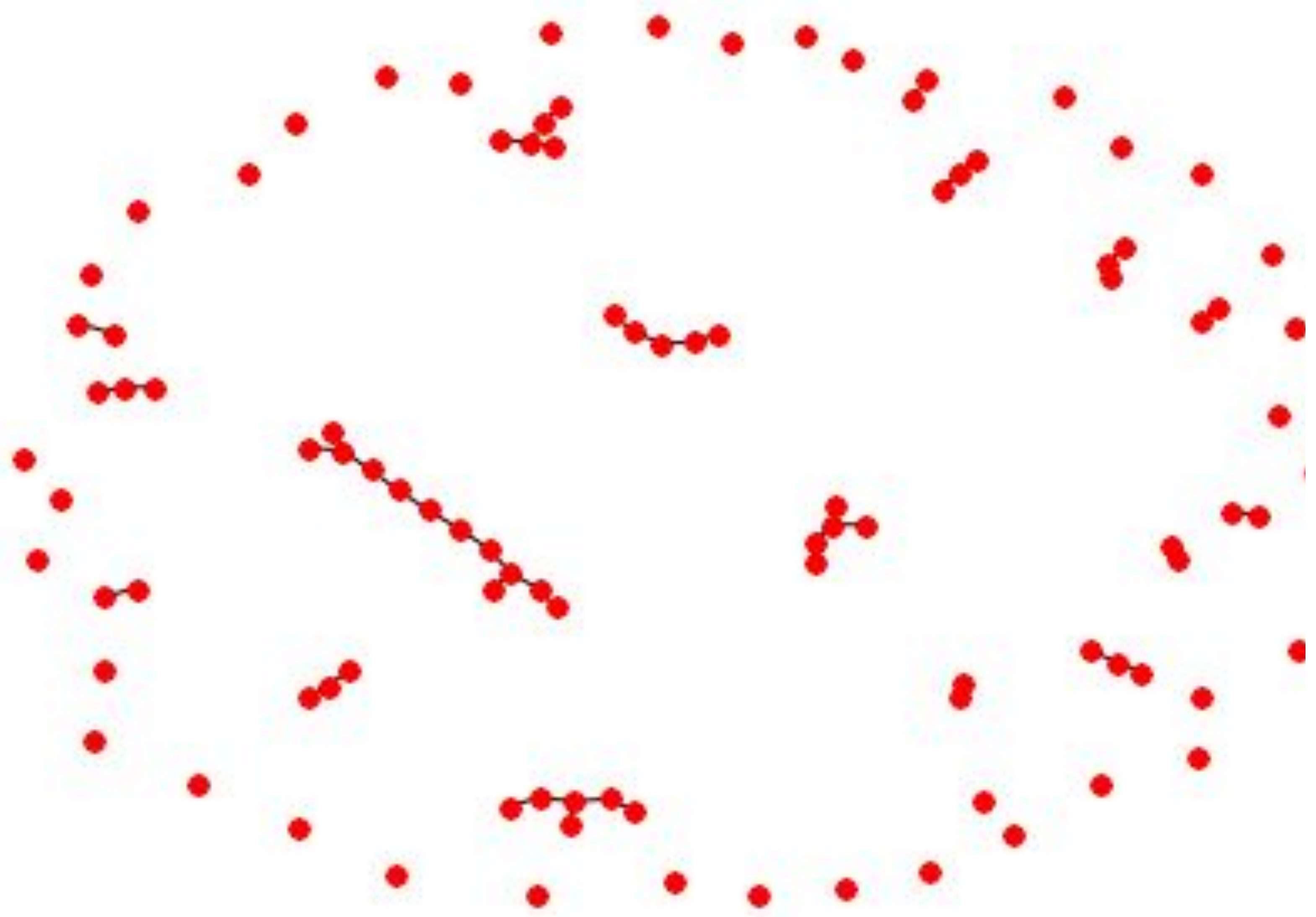
For each node, there are n-1 others in the graph it could connect to.

Each of those connections can happen with probability p

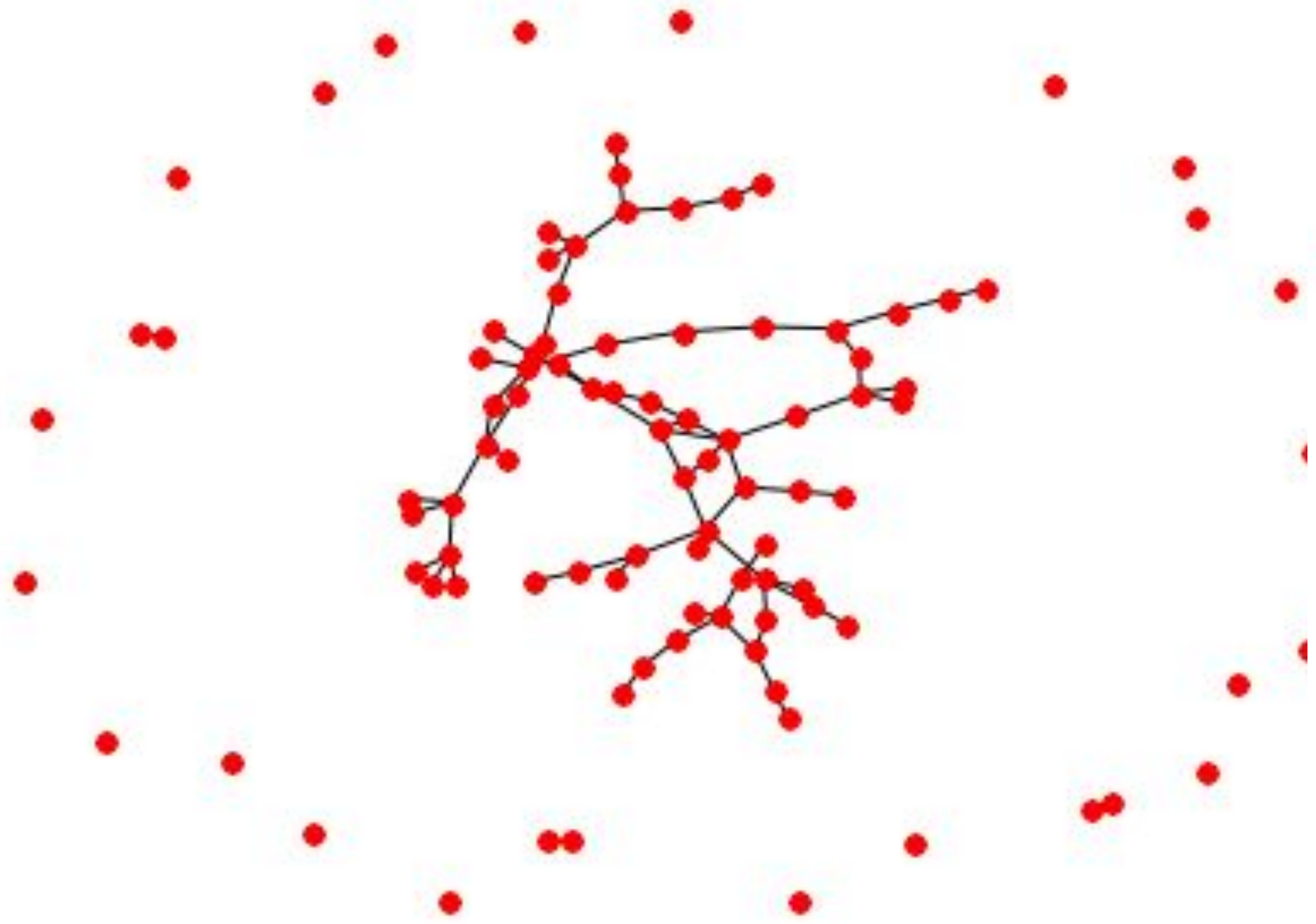
(If you were a fan of Probability and Matrices, this is a binomial with **n-1 trials** and **success probability p**)

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

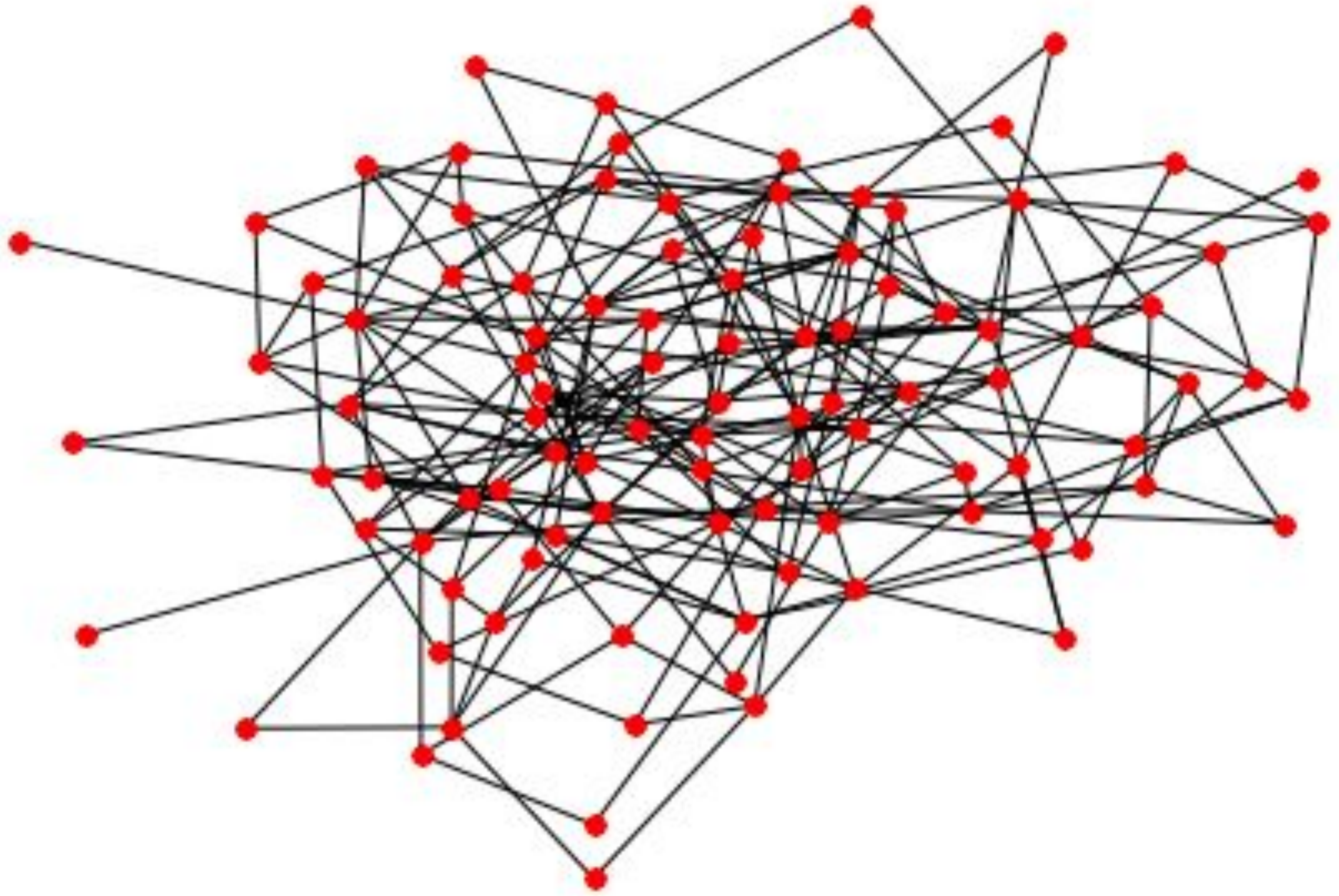
What do ER graphs look like?



Very disconnected graph, only tiny connected components

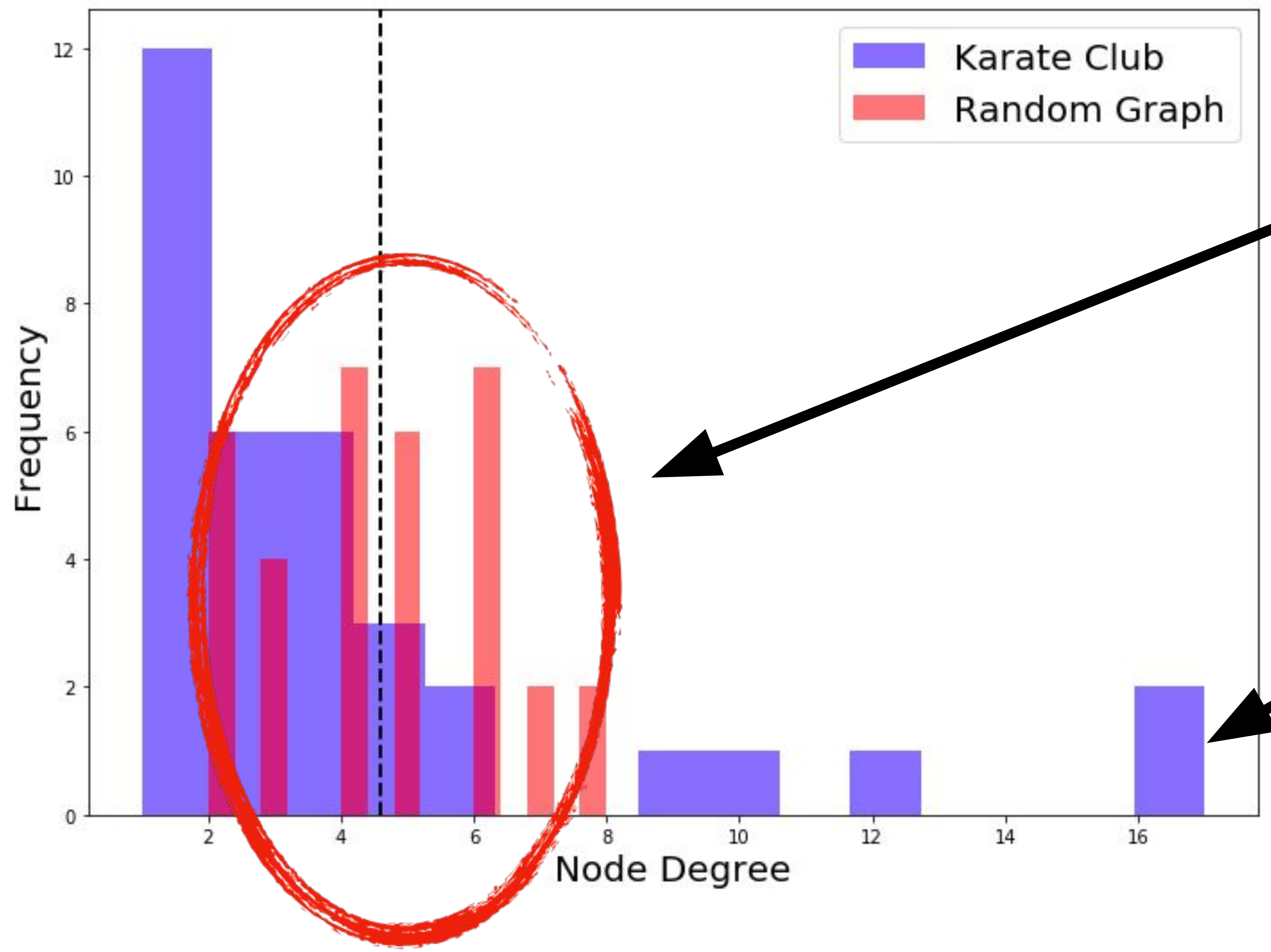


A giant component appears, no/very few cycles



Whole graph is connected, some cycles present

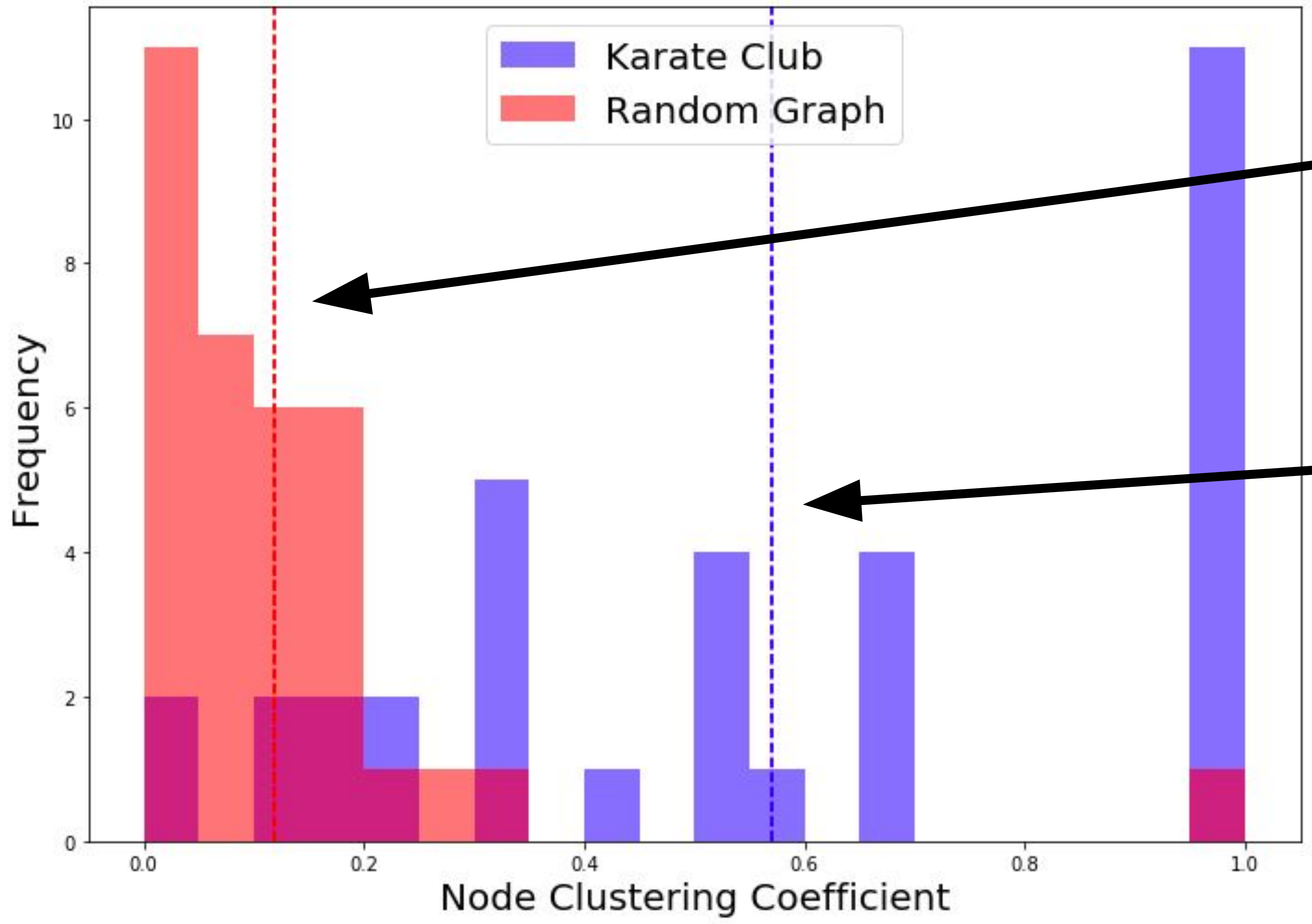
Random Graphs vs Real Networks



Random: node degrees all clustered round the average value

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

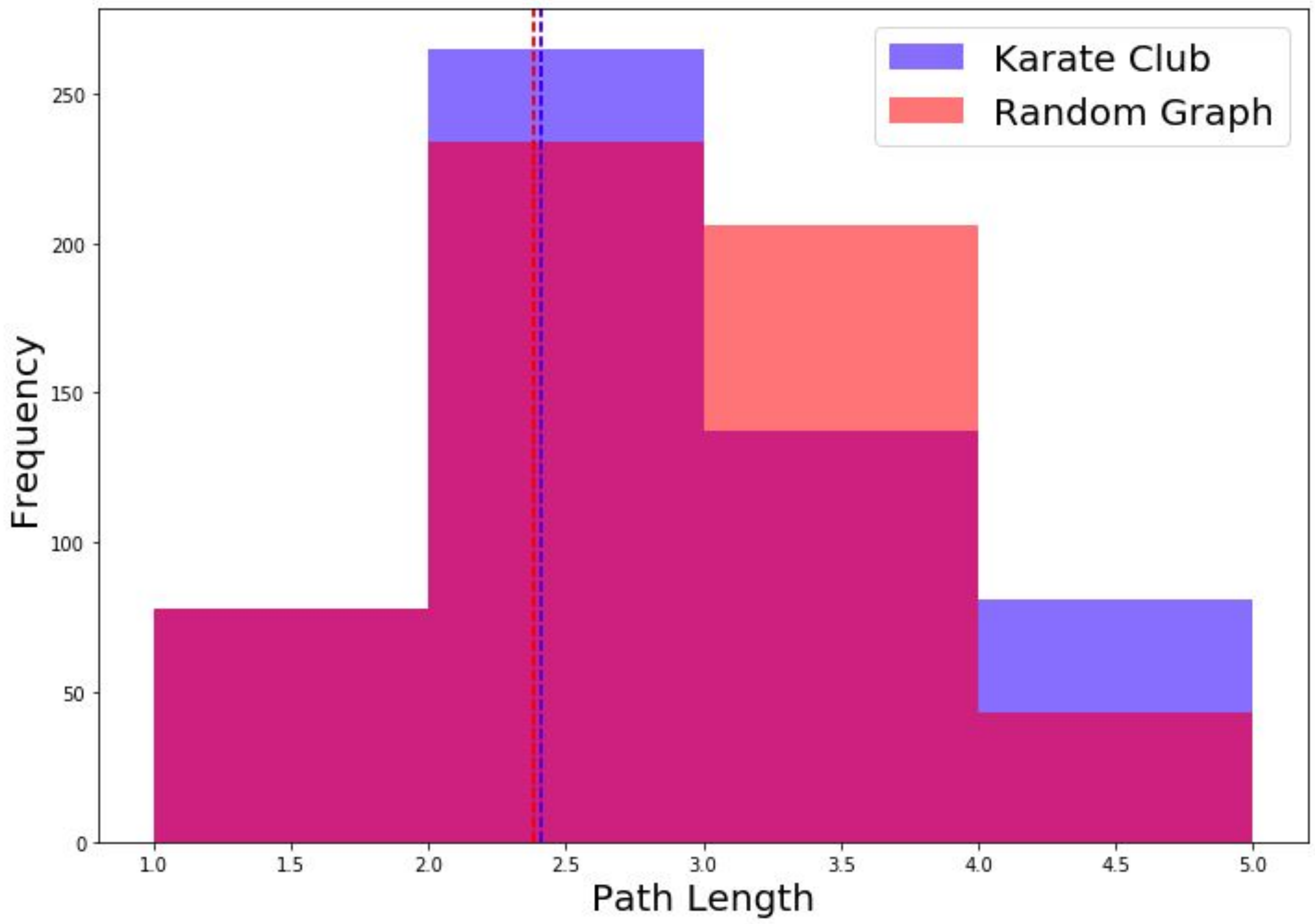
Random Graphs vs Real Networks



Random: very low average clustering coefficent

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

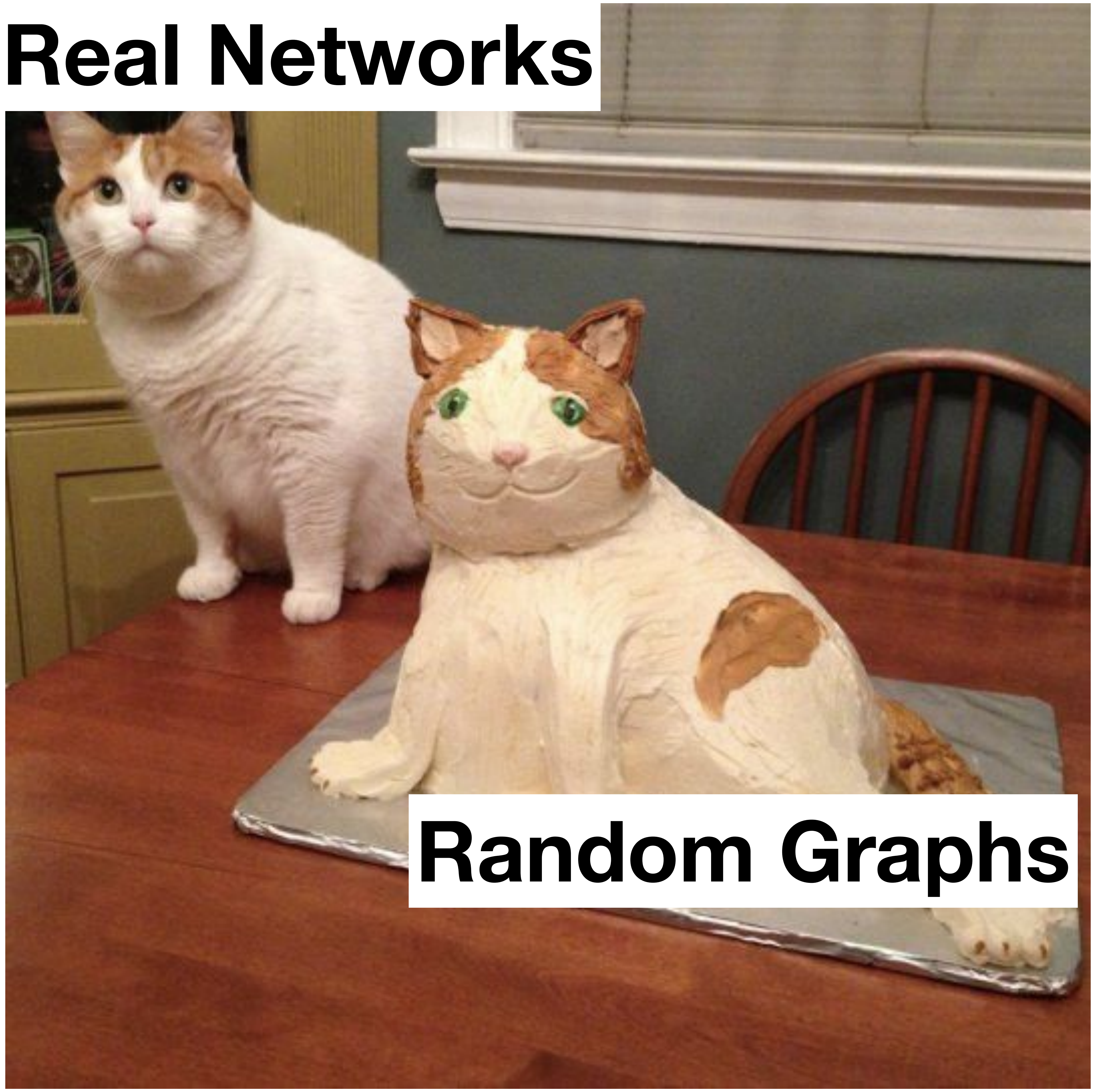
Random Graphs vs Real Networks



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?