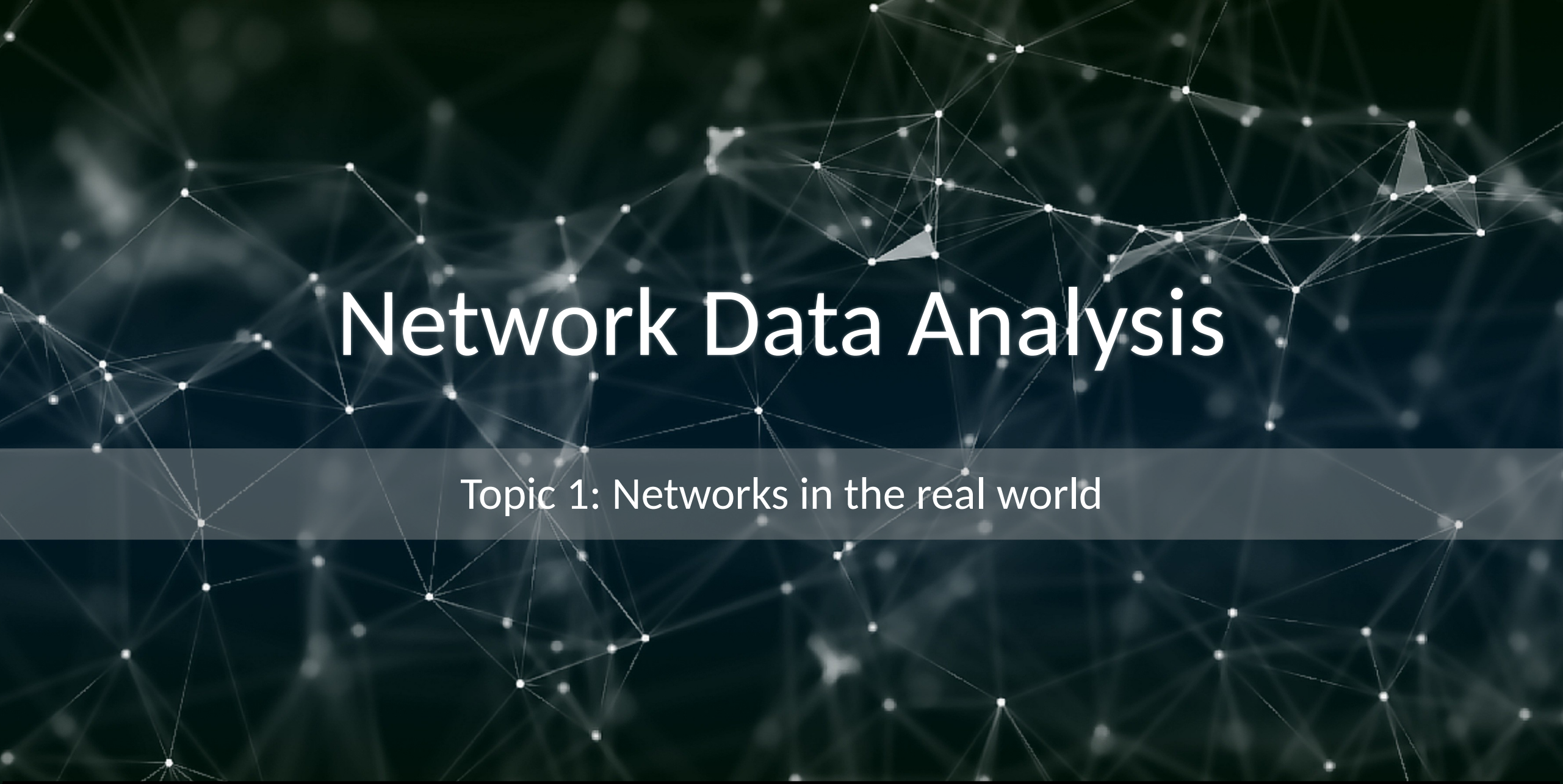
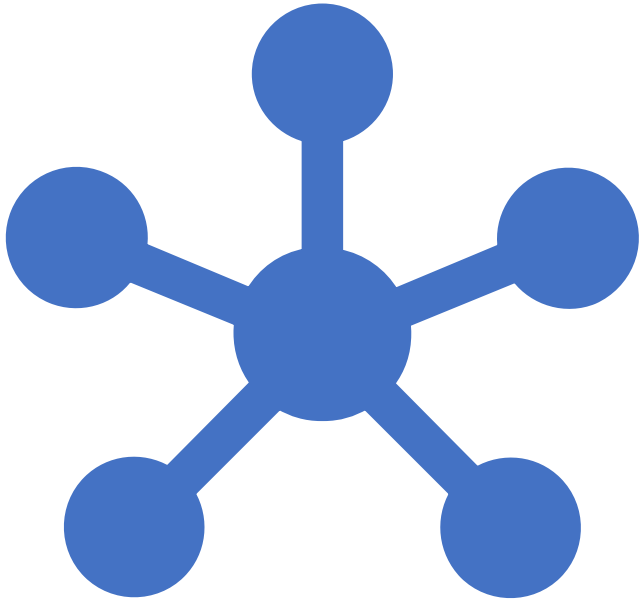


Network Data Analysis

Topic 1: Networks in the real world

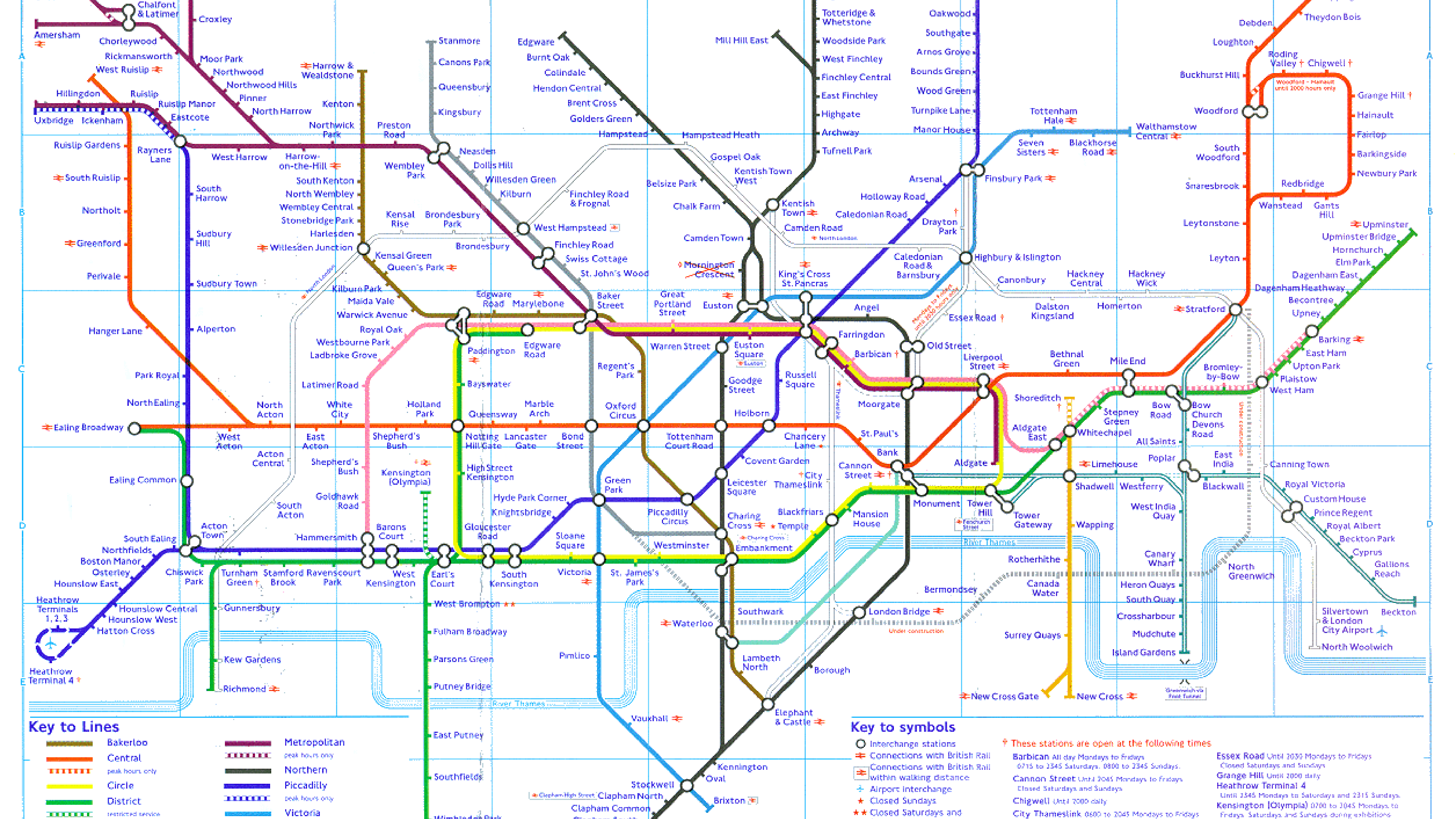
KING'S College LONDON

Albert Meroño-Peñuela
Jacopo de Berardinis

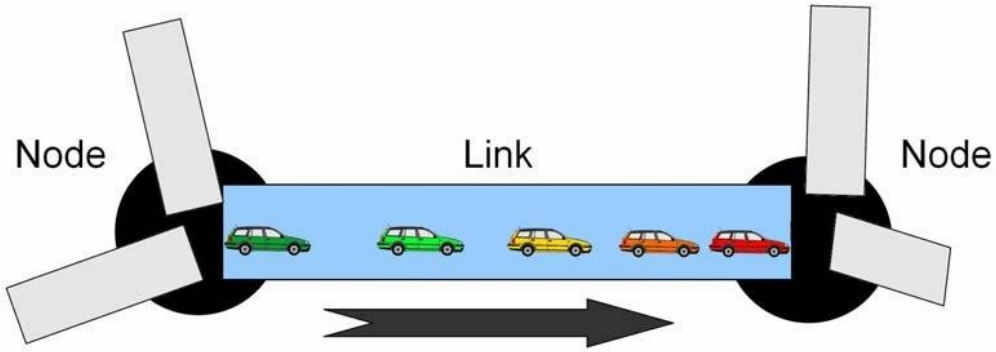


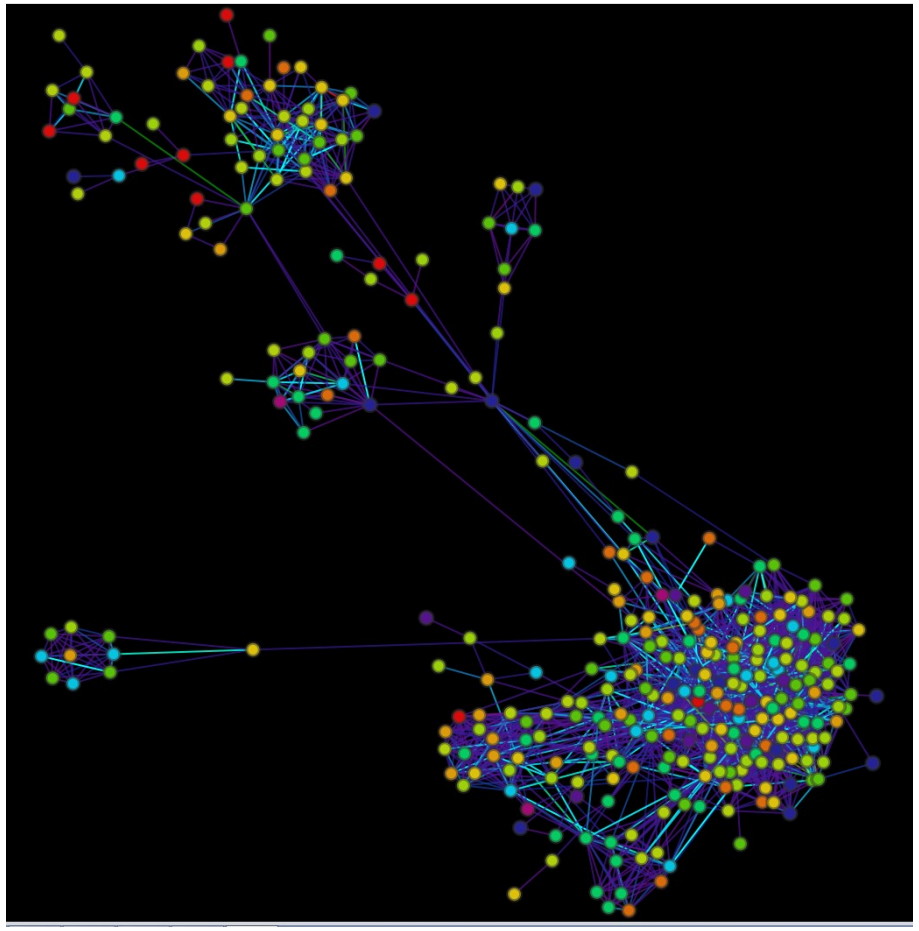
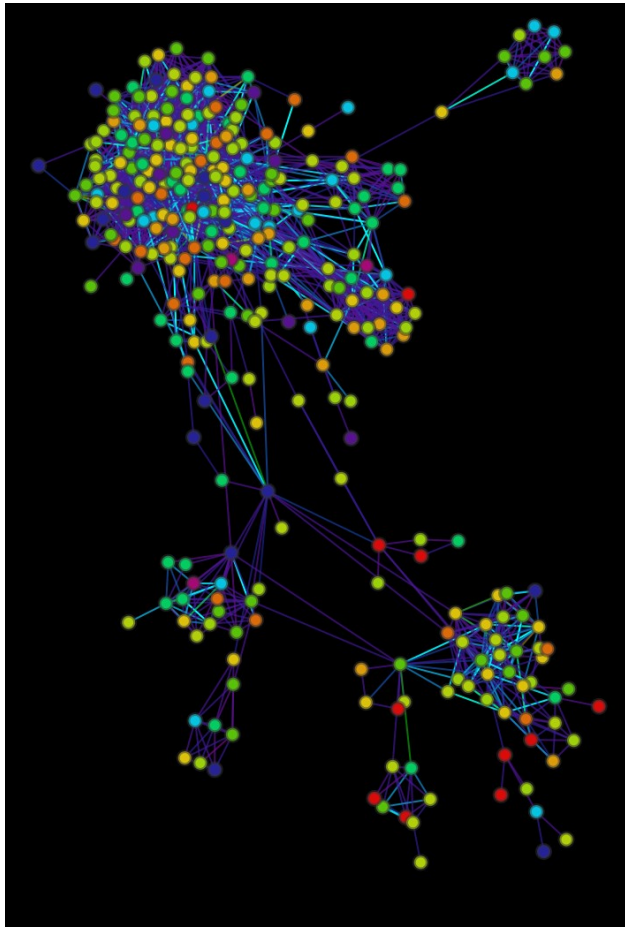
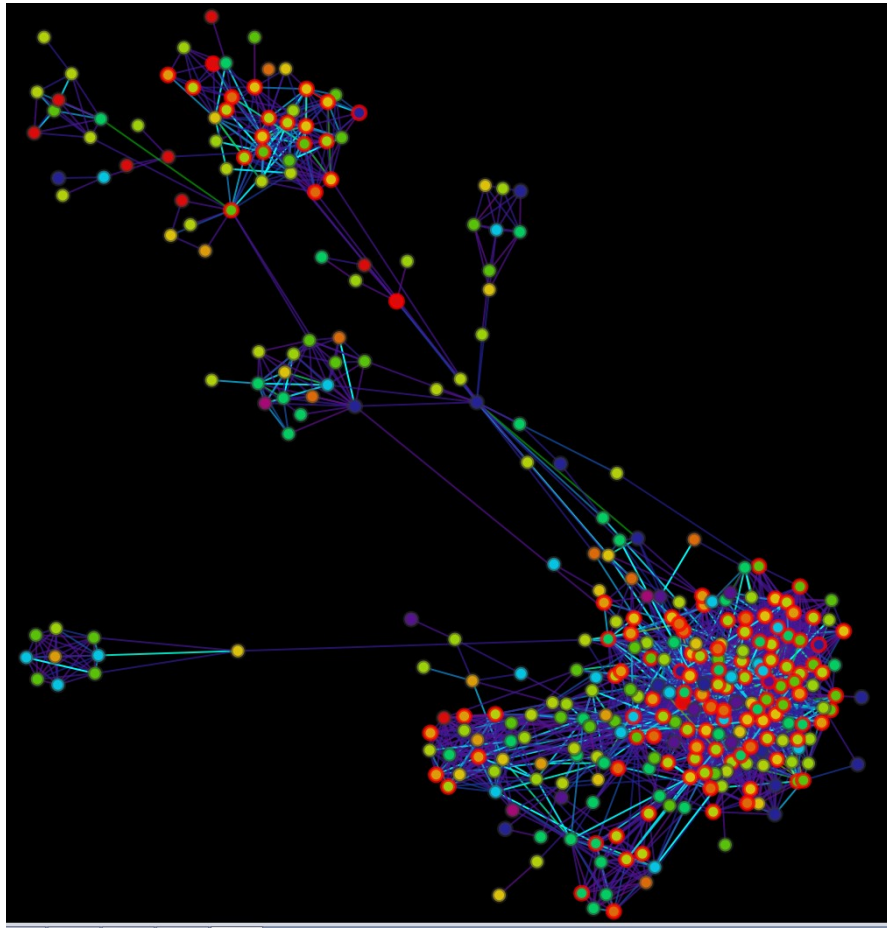
What is a graph?

- A graph, or network, is a set of items (***nodes*** or *vertices*) connected by their relations (***edges*** or *links*).
- Networks are present in the everyday life of all people.
  - Social: Friends & family, society
  - Transport: Rail, road, public transport
  - Information: World economy
  - Biology: Brain cells
  - Communication: Internet
- Consciously or not, we use networks and their properties as a universal language for describing complex data.
  - Networks from science, nature, and technology are more similar than one would expect.
  - Shared vocabulary between fields, including computer science, social science, physics, economics, statistics, biology.

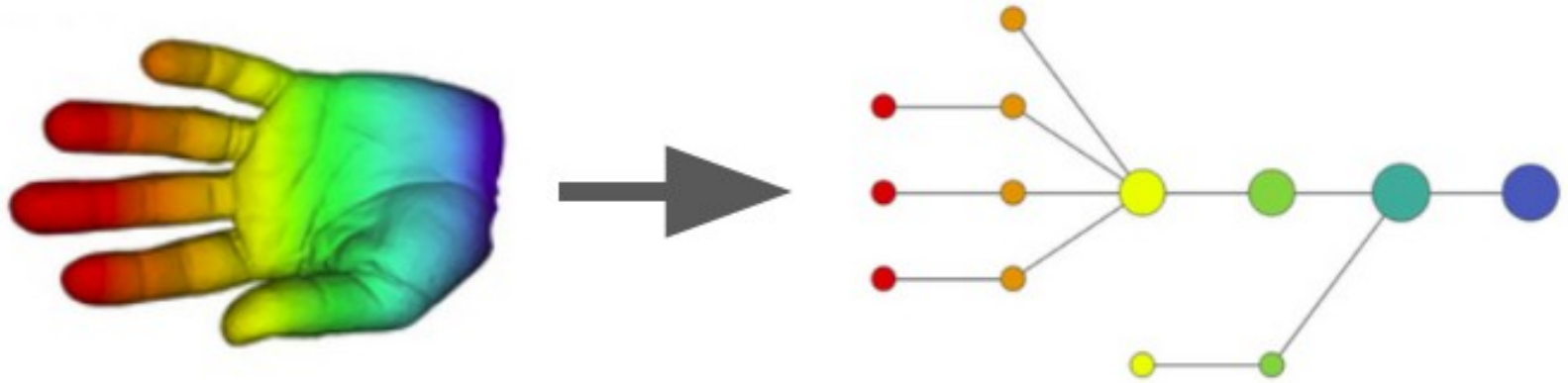

Amersham
Chalfont & Latimer
Chorleywood
Croxley
Rickmansworth
Moor Park
Northwood
Northwood Hills
Pinner
North Harrow
West Ruislip
Hillingdon
Ruislip
Ruislip Manor
Eastcote
Uxbridge
Ickenham
Ruislip Gardens
Rayners Lane
West Harrow
Harrow-on-the-Hill
South Ruislip
Northolt
Greenford
Perivale
Hanger Lane
South Harrow
Sudbury Hill
Sudbury Town
Alperton
Park Royal
North Ealing
Ealing Broadway
Ealing Common
West Acton
North Acton
East Acton
Acton Central
South Acton
Acton Town
South Ealing
Northfields
Boston Manor
Osterley
Hounslow East
Hounslow Central
Hounslow West
Hatton Cross
Heathrow Terminals 1, 2, 3
Heathrow Terminal 4
Chiswick Park
Turnham Green
Stamford Brook
Ravenscourt Park
Gunnersbury
Kew Gardens
Richmond
Harrow & Wealdstone
Kenton
Northwick Park
Preston Road
Wembley Park
South Kenton
North Wembley
Wembley Central
Stonebridge Park
Harlesden
Willesden Junction
Kensal Rise
Brondesbury Park
Brondesbury
Kensal Green
Queen's Park
Kilburn Park
Maida Vale
Warwick Avenue
Royal Oak
Westbourne Park
Ladbroke Grove
Latimer Road
White City
Shepherd's Bush
Holland Park
Goldhawk Road
Hammersmith
Kensington (Olympia)
Barons Court
West Kensington
Earl's Court
West Brompton
Fulham Broadway
Parsons Green
Putney Bridge
East Putney
Southfields
Wimbledon Park
Stanmore
Canons Park
Queensbury
Kingsbury
Neasden
Dollis Hill
Willesden Green
Kilburn
West Hampstead
Finchley Road
Swiss Cottage
St. John's Wood
Edgware
Burnt Oak
Colindale
Hendon Central
Brent Cross
Golders Green
Hampstead
Hampstead Heath
Finchley Road & Frognal
Belsize Park
Chalk Farm
Camden Town
Mornington Crescent
Kentish Town
Kentish Town West
Gospel Oak
Camden Road
Mill Hill East
Totteridge & Whetstone
Woodside Park
West Finchley
Finchley Central
East Finchley
Highgate
Archway
Tufnell Park
Oakwood
Southgate
Arnos Grove
Bounds Green
Wood Green
Turnpike Lane
Manor House
Arsenal
Holloway Road
Caledonian Road
Finsbury Park
Drayton Park
Caledonian Road & Barnsbury
Highbury & Islington
Canonbury
King's Cross St. Pancras
Euston
Great Portland Street
Baker Street
Marylebone
Edgware Road
Paddington
Bayswater
Queensway
Notting Hill Gate
Lancaster Gate
Marble Arch
Bond Street
Oxford Circus
Regent's Park
Warren Street
Euston Square
Goodge Street
Tottenham Court Road
Holborn
Russell Square
Chancery Lane
Covent Garden
Leicester Square
Piccadilly Circus
Green Park
Hyde Park Corner
Knightsbridge
High Street Kensington
Gloucester Road
South Kensington
Sloane Square
Victoria
St. James's Park
Westminster
Charing Cross
Embankment
Temple
Blackfriars
City Thameslink
Waterloo
Lambeth North
Southwark
Pimlico
Vauxhall
Stockwell
Oval
Kennington
Clapham North
Clapham Common
Brixton
Elephant & Castle
Borough
London Bridge
Angel
Farringdon
Barbican
Moorgate
Old Street
Essex Road
Liverpool Street
St. Paul's
Bank
Cannon Street
Mansion House
Monument
Tower Hill
Tower Gateway
Aldgate
Aldgate East
Shoreditch
Whitechapel
Bethnal Green
Mile End
Stepney Green
Shadwell
Wapping
Rotherhithe
Canada Water
Surrey Quays
New Cross Gate
New Cross
Bermondsey
Limehouse
Westferry
West India Quay
Canary Wharf
Heron Quays
South Quay
Crossharbour
Mudchute
Island Gardens
Poplar
All Saints
Bow Road
Bow Church
Devons Road
Bromley-by-Bow
East India
Blackwall
Canning Town
Royal Victoria
Custom House
Prince Regent
Royal Albert
Beckton Park
Cyprus
Gallions Reach
Beckton
North Greenwich
Silvertown & London City Airport
North Woolwich
Tottenham Hale
Seven Sisters
Blackhorse Road
Walthamstow Central
Dalston Kingsland
Hackney Central
Hackney Wick
Homerton
Stratford
West Ham
Plaistow
Upton Park
East Ham
Barking
Upney
Becontree
Dagenham Heathway
Dagenham East
Elm Park
Hornchurch
Upminster Bridge
Upminster
Leyton
Leytonstone
Wanstead
Snaresbrook
Redbridge
Gants Hill
Newbury Park
Barkingside
Fairlop
Hainault
Grange Hill
Chigwell
Roding Valley
South Woodford
Woodford
Buckhurst Hill
Loughton
Debden
Theydon Bois
Woodford - Hainault until 2000 hours only
River Thames
Under construction
Key to Lines
Bakerloo
Central
peak hours only
Circle
District
restricted service
Metropolitan
peak hours only
Northern
Piccadilly
peak hours only
Victoria
Key to symbols
Interchange stations
Connections with British Rail
Connections with British Rail within walking distance
Airport interchange
Closed Sundays
Closed Saturdays and
† These stations are open at the following times
Barbican All day Mondays to Fridays 0715 to 2345 Saturdays. 0800 to 2345 Sundays.
Cannon Street Until 2045 Mondays to Fridays Closed Saturdays and Sundays
Chigwell Until 2000 daily
City Thameslink 0600 to 2045 Mondays to Fridays
Essex Road Until 2030 Mondays to Fridays Closed Saturdays and Sundays
Grange Hill Until 2000 daily
Heathrow Terminal 4
Kensington (Olympia)

Agent-based modelling of traffic



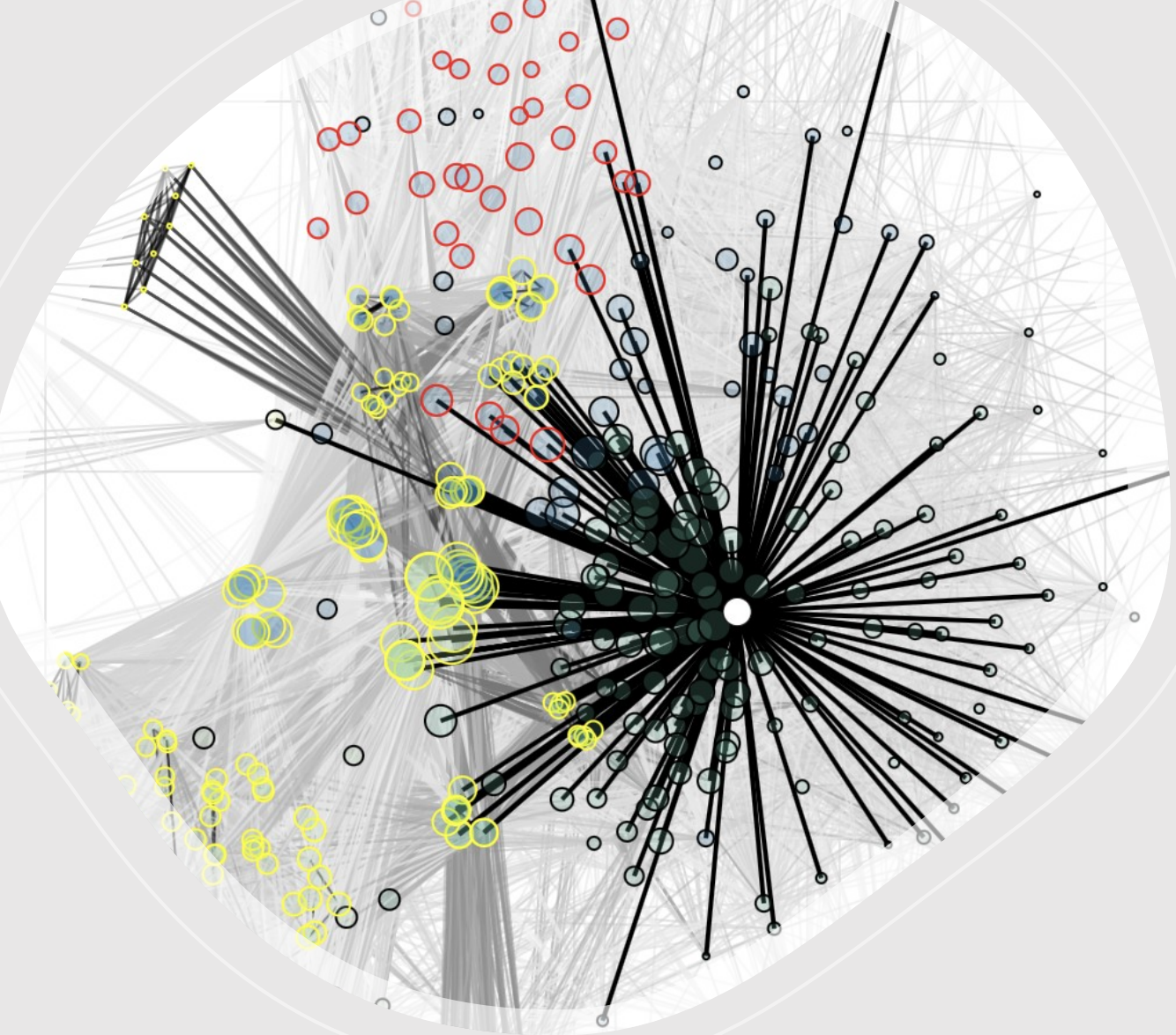




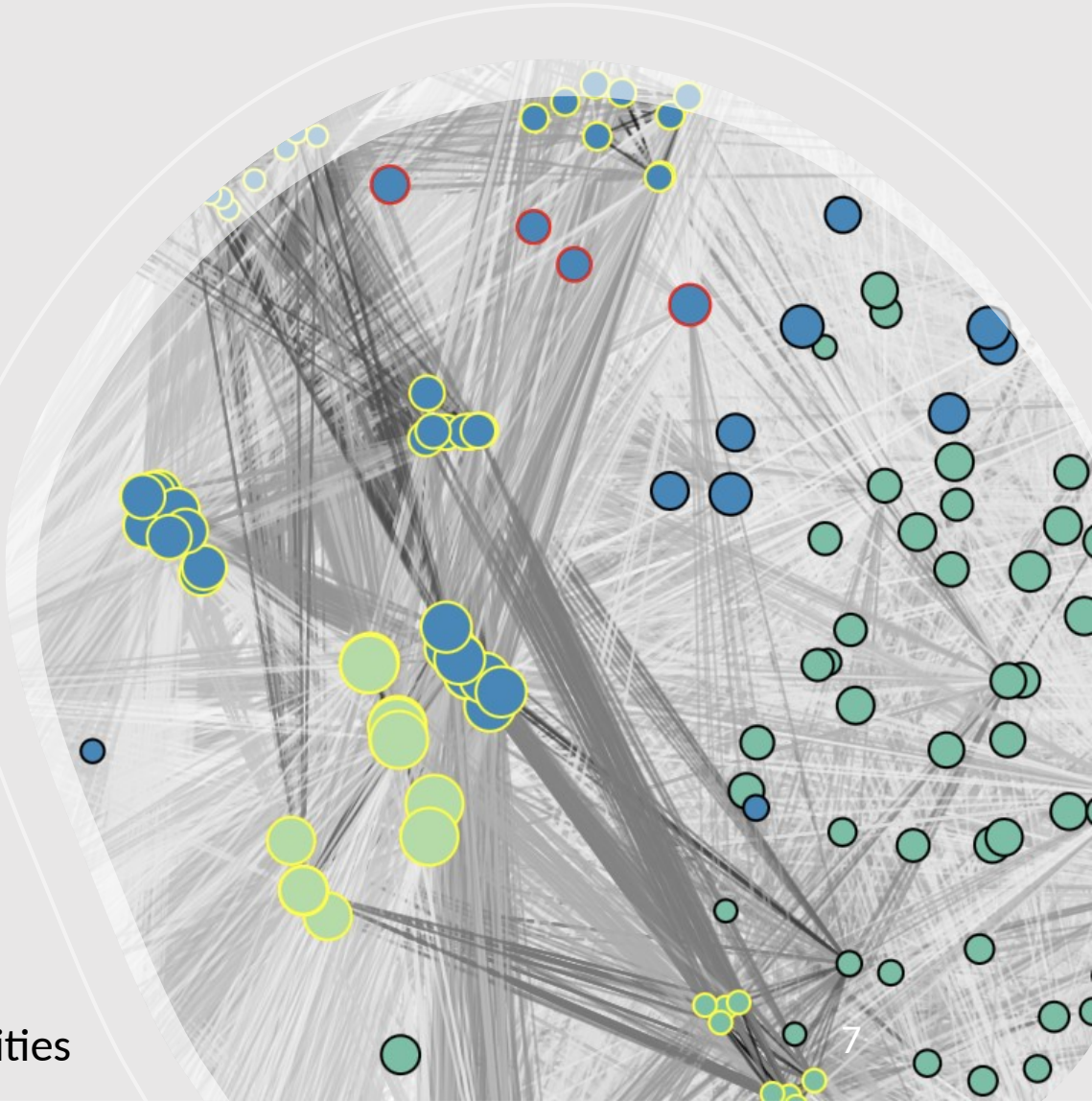
Information spread in social networks



Topological Data Analysis: looking at “the shape of data”

Polifonia





The Harmonic Network

Nodes are tracks, Edges express harmonic similarities

Networks in the world

Social and economic networks: People or groups with contacts or interactions between them

- Friendship networks, business relations between companies, intermarriages between families, labour markets
- **Example**: To understand disease spread, you need to know social networks.
- **Questions**: Degree of connectedness, small-world effects

Information networks: Connections of information objects

- Academic paper citations, network of webpages with links to others
- **Example**: To understand news dissemination, you need to know information networks.
- **Questions**: Ranking, navigation

Spatial networks: Networks where nodes and edges have location

- Road networks, migration destinations, river systems
- **Example**: To understand whether crime is distributed according to space, you need to know street networks
- **Questions**: Network-based clustering or correlations

Study of networks

Empirical: Study network data to find organisational principles

How do we measure and quantify networks?

Mathematical models: Graph theory, statistical models

Models allow us to understand behaviors and distinguish surprising from expected phenomena

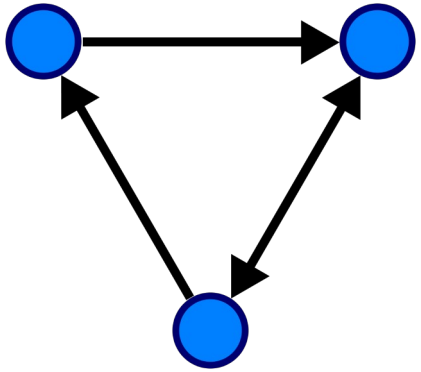
Algorithms for analysing graphs

Hard computational challenges!

Historical study of networks with mathematical graph theory

Graphs and their properties

Graphs



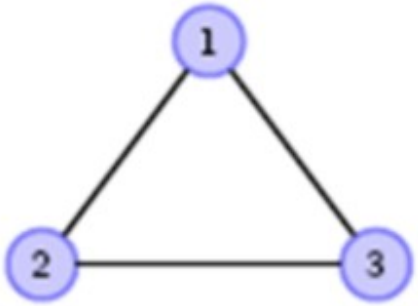
- A **graph** is a mathematical representation of a network, and the basic data structure for analysing network data
- A graph consists of a set of **nodes** and a set of **edges**

$$G = (V, E)$$

Graph representations

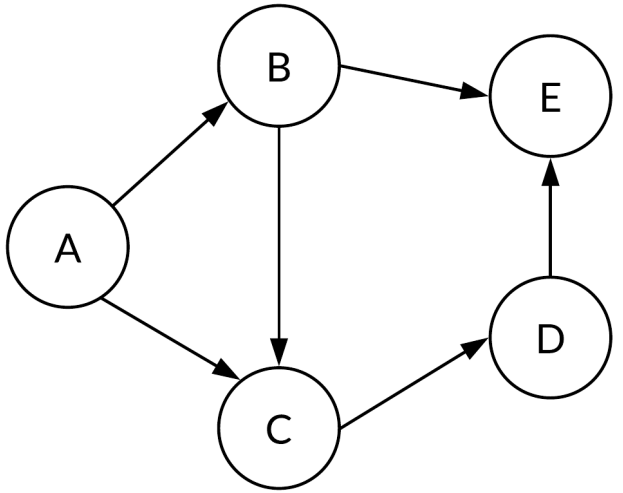
- **Adjacency matrix:** the graph is represented by a tuple **(N, g)**, where **N = {1, 2, …, n}** is a set of **n** nodes, and **g** is a **n x n** matrix (the adjacency matrix). Each element of the matrix $\mathbf{g_{ij}}$ has value **1** if there is an edge between nodes **i** and **j**, and **0** if there is not.

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \Rightarrow$$



Graph representations

- **Adjacency list:** represented by a tuple **(N, E)**, where **N = {1, 2, ..., n}** is a set of **n** nodes as before and **E** is a set of edges, where each edge is a tuple of the two nodes connected by the edge, e.g. **E = {(1, 2), (1, 3), (2, 3)}**.

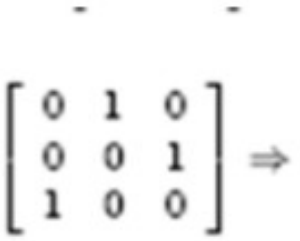


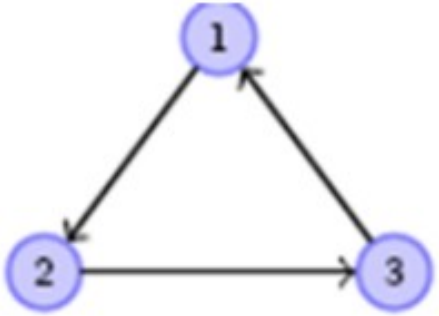

```
Adjacency list:

graph = {
'A': ['B', 'C'],
'B': ['C', 'E'],
'C': ['D'],
'D': ['E'],
}
```

Directed and undirected graphs

- Directed graph: each edge has a direction, going from the **source** node to the **target** (or sink) node. For example, in the directed graph with edges **E = {(1, 2), (1, 3), (2, 3)}**, there is an edge from node 1 (the source) to node 2 (the target) but no edge from node 2 to node 1
- Undirected graph: edges have no direction, e.g. in an undirected graph with edges **E = {(1, 2), (1, 3), (2, 3)}**, there is a connection between nodes 1 and 2 but no notion of whether that goes ‘from’ or ‘to’ each node
- **Underlying undirected graph** of a directed graph is the undirected graph with the same nodes and edges but ignoring any direction



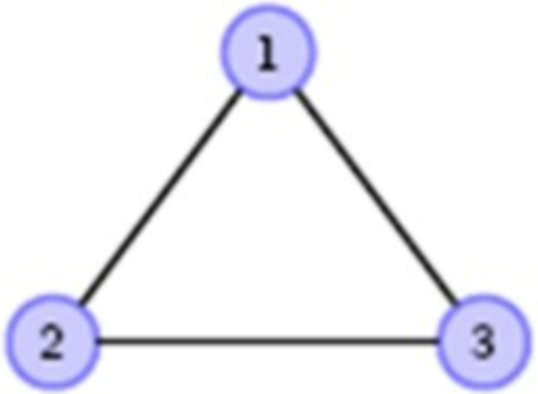


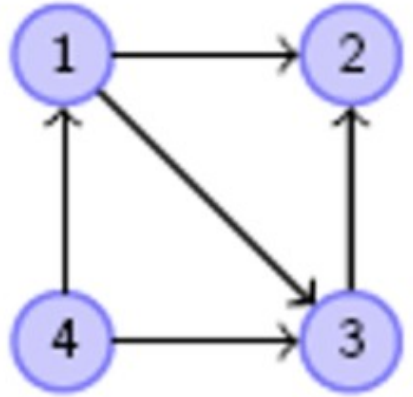
Graph walks

- **Graph walk:** a series of connected edges representing a journey through the graph from one node via other nodes to a final node following the edges. If the graph is directed, then the walk must follow the direction of the edges
- **Graph path:** a graph walk where no node is repeated except possibly the start node if it is the same as the end node

Neighbourhood and degree

- The **neighbourhood** of a node is the set of nodes it is connected to
- The **degree** of a node is the number of edges of that node
  - For undirected graphs, the **degree** of a node is the number of edges that involve that node, i.e. the size of its neighbourhood
  - For directed graphs, the **in-degree** of a node is the number of edges for which the node is the target, while the **out-degree** of a node is the number of edges for which the node is the source





Properties of graphs

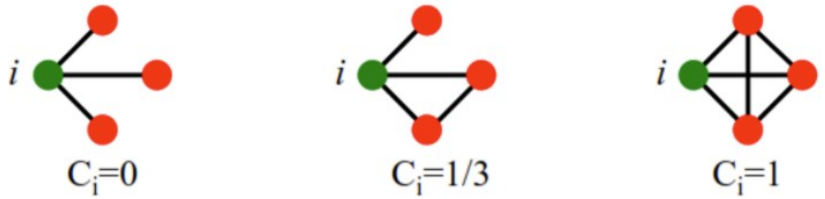
- **Degree distribution: P(k) = N(k)/n** probability of a random node having degree k, for a graph of **n** nodes and **N(k)** is the number of nodes in the graph with degree **k**
- Path **length**: number of edges that the path contains. The **shortest path** between two nodes is the path with the shortest length
- **Distance** between two nodes: length of the shortest path between them
- **Diameter:** longest distance between nodes in a connected graph.

Properties of graphs

- **Average path length:** mean average distance between any two nodes in the network

$$l_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, v_j)$$

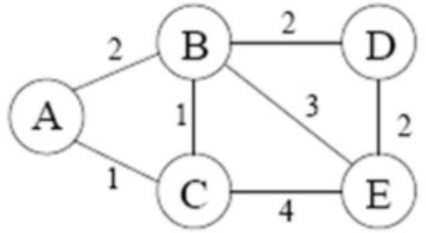
- **Clustering coefficient** of a node is the proportion of its neighbours that are connected to each other. $\mathbf{C_i = e / m}$ is the clustering coefficient of node **I** (m = total possible edges between neighbours; e = actual number of edges between neighbours)

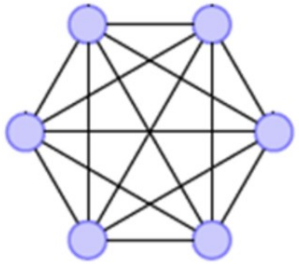


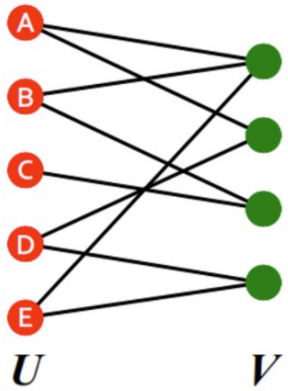
- **Average clustering**: mean average of the clustering coefficients of its nodes

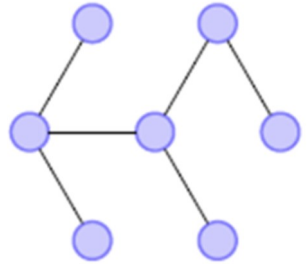
Types of graphs

- **Weighted graph:** labelled graph where edges have numeric labels, e.g. weights
- **Complete graph:** graph where every node is connected to every other node (maximum number of edges, i.e. n(n - 1) / 2)
- **Bipartite graph**: nodes can be divided into two disjoint sets U and V, where every edge connects a node in U to a node in V
- **Tree:** connected undirected graph (with n – 1 edges)

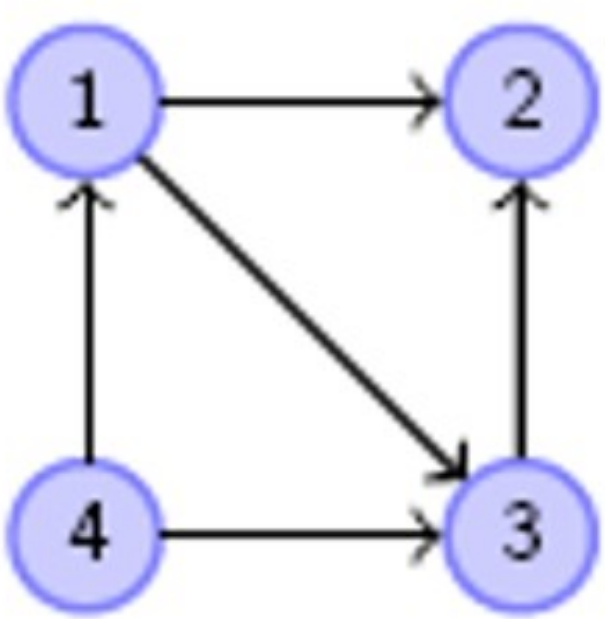






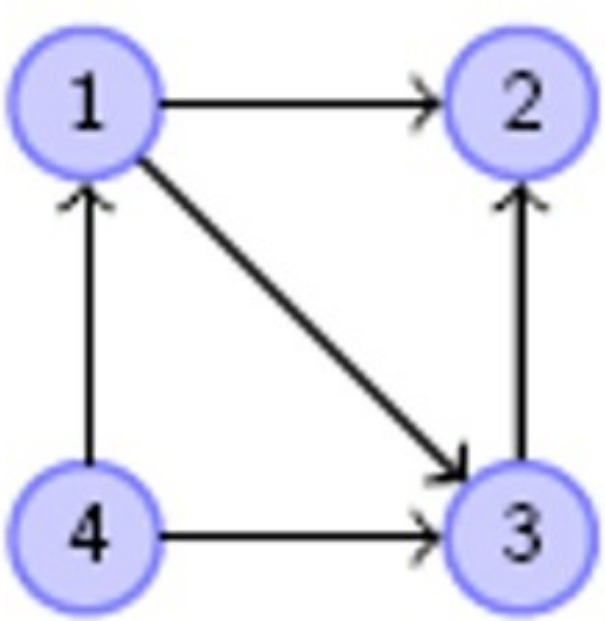


TUTORIAL EXERCISE 1



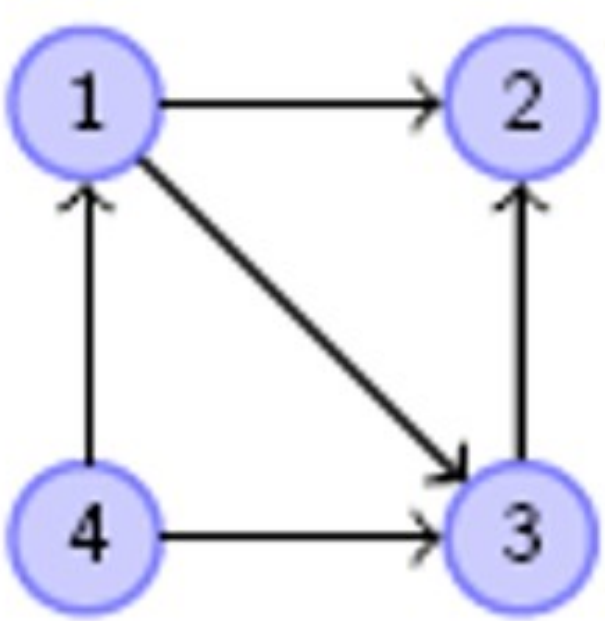
In the out-degree distribution P(k), what is the value of P(2)?

TUTORIAL EXERCISE 2



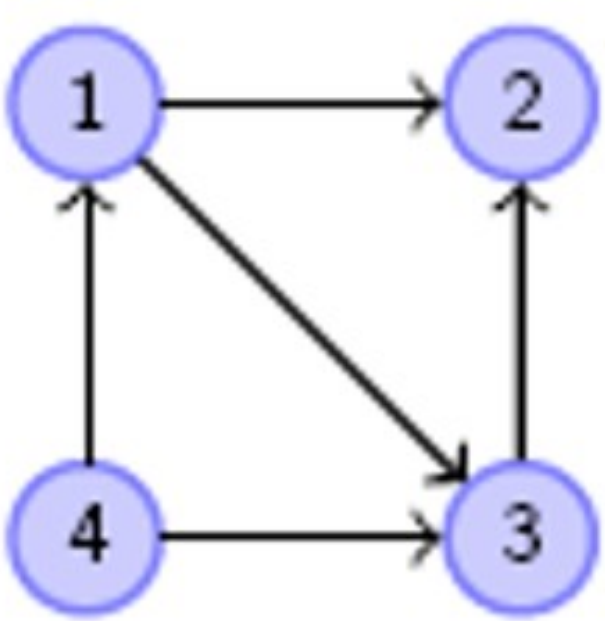
What is the diameter of the underlying undirected graph?

TUTORIAL EXERCISE 3



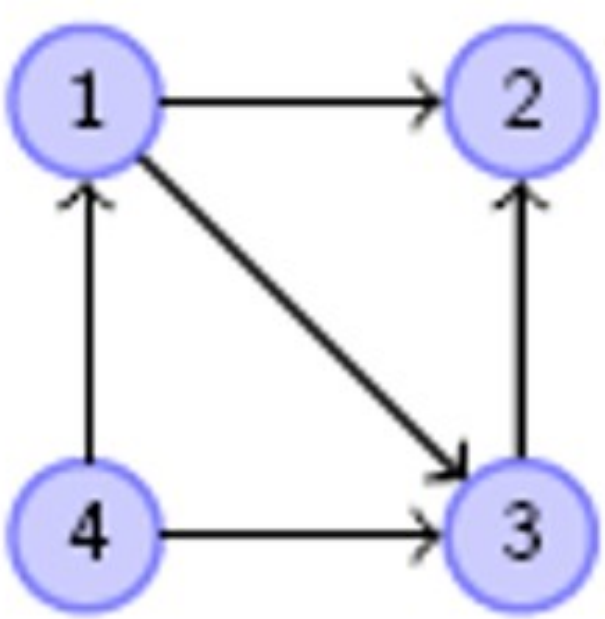
What is the average path length of the underlying undirected graph?

TUTORIAL EXERCISE 4



What is the clustering coefficient of node 1 in the underlying undirected graph?

TUTORIAL EXERCISE 5



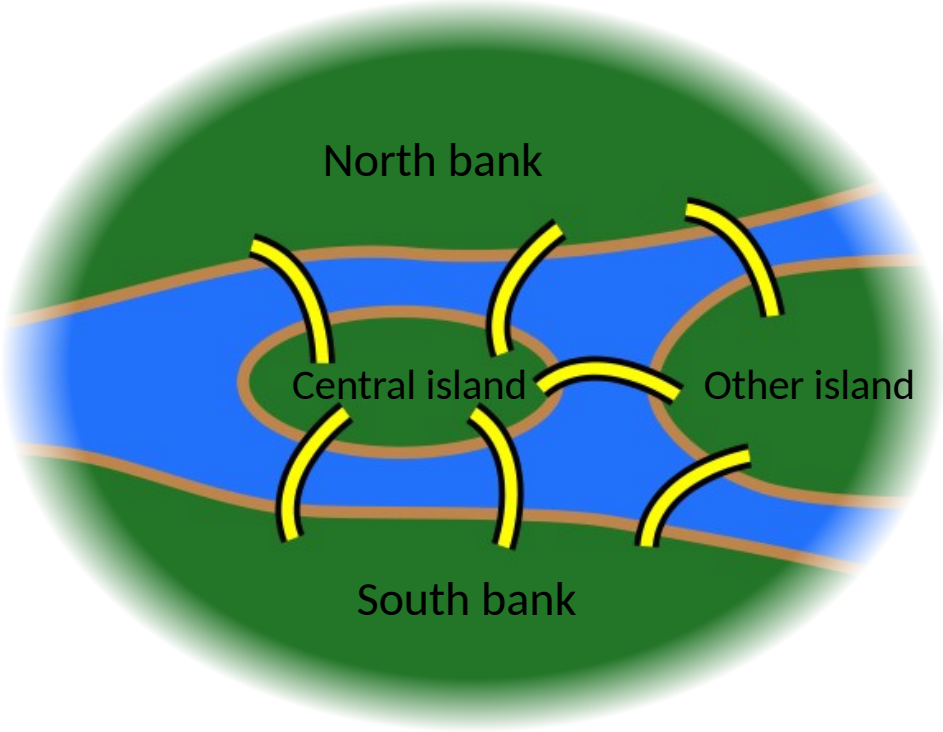
How many of the following sequences could be returned by a depth-first search from node 4?

[4, 1, 2, 3]	[4, 1, 3, 2]
[4, 2, 1, 3]	[4, 2, 3, 1]
[4, 3, 1, 2]	[4, 3, 2, 1]

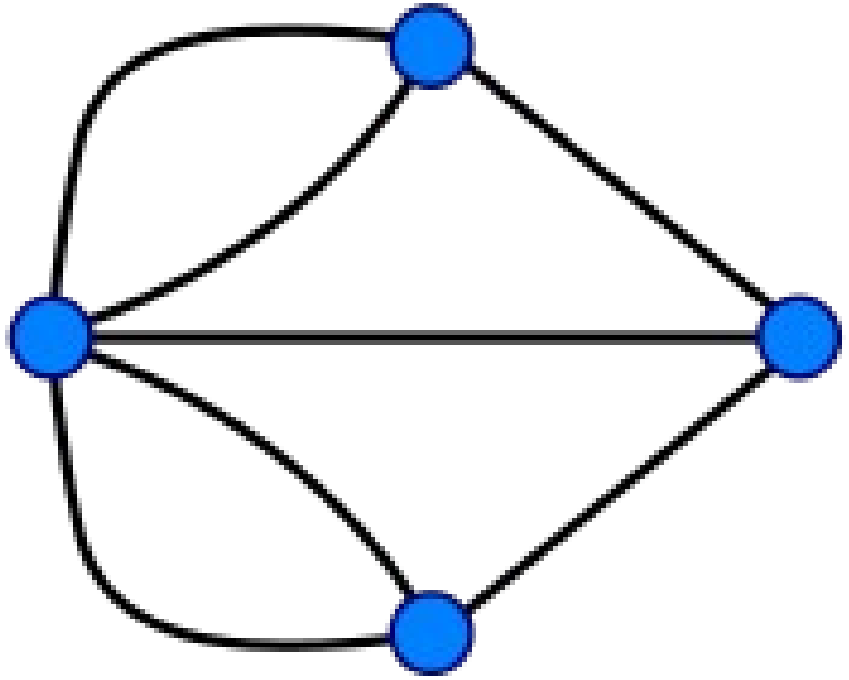
Origins of graph theory

Königsberg bridge problem

- The city of *Königsberg* in *Prussia* (now *Kaliningrad*, *Russia*) was set on both sides of the *Pregel* River
  - two large islands which were connected to each other and the mainland by seven bridges.
- The problem was to find a walk through the city that would cross each bridge once and only once.
  - The islands could not be reached other than by the bridges, and every bridge must have been crossed completely every time.
- Euler proved that the problem has **no** solution



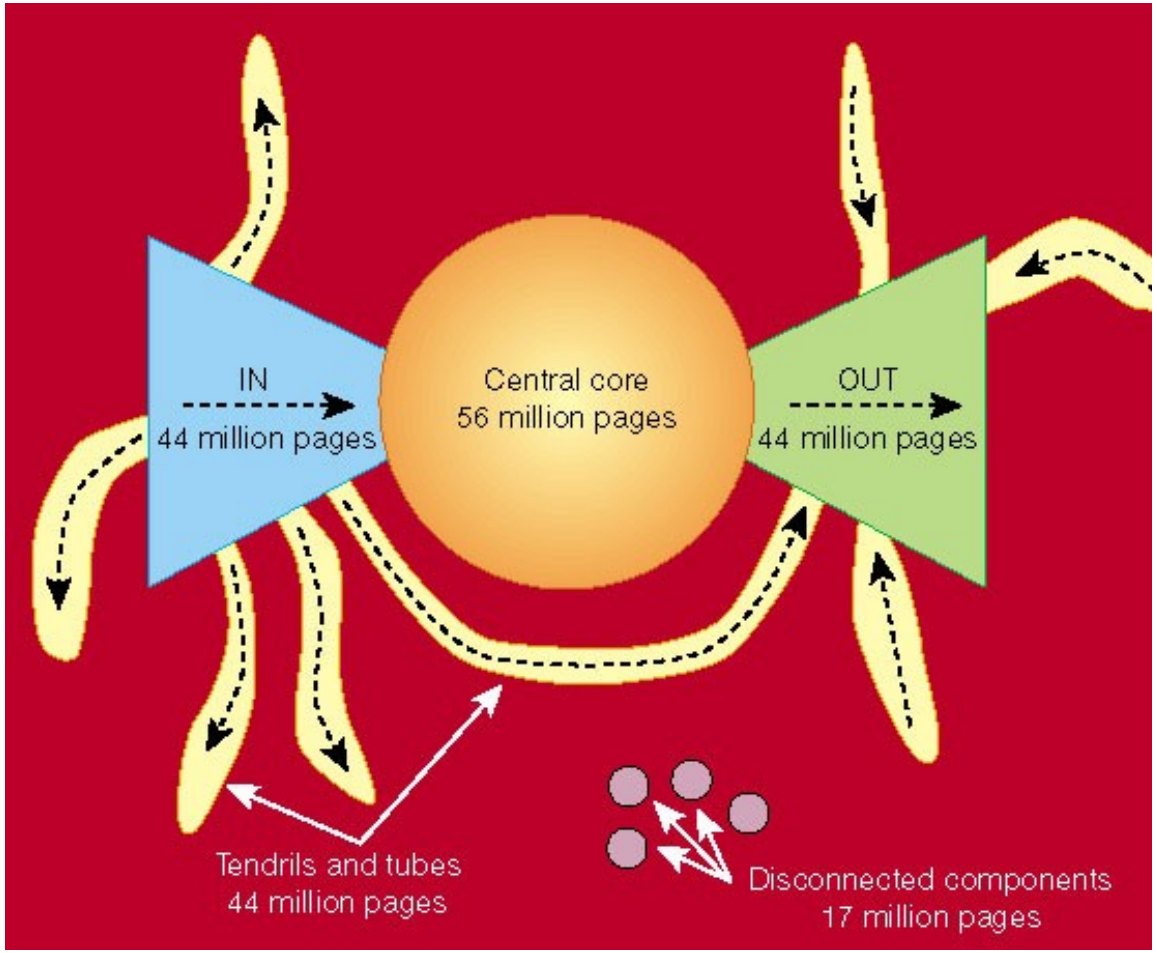
Königsberg bridge problem



Web as a Graph

- Web as a directed graph:
  - Nodes = web pages & Edges = hyperlinks
- Directed version of the Web graph:
  - Largest SCC: 28% of the nodes (56 M)
  - Taking a random node v
    - Out(v) ≈ 50%
    - In(v) ≈ 50%

"The web is a bowtie", Nature, vol. 405, May 2000



Weekly discussion topic

All students: post your thoughts on the module's discussion forum in KEATS

Imagine we are going to analyse how the housing market in a city functions.

- What networks could be relevant to consider in analysing this topic?
- What kinds of graph, as studied in this lesson, represent these networks?
- What might the properties studied in this lesson, such as diameter or cluster coefficient, applied to these networks tell you about the city?
- What public data sets can you find that provide you with data on the networks you've identified?
- Who might hold private network data sets that would also be valuable in this analysis?