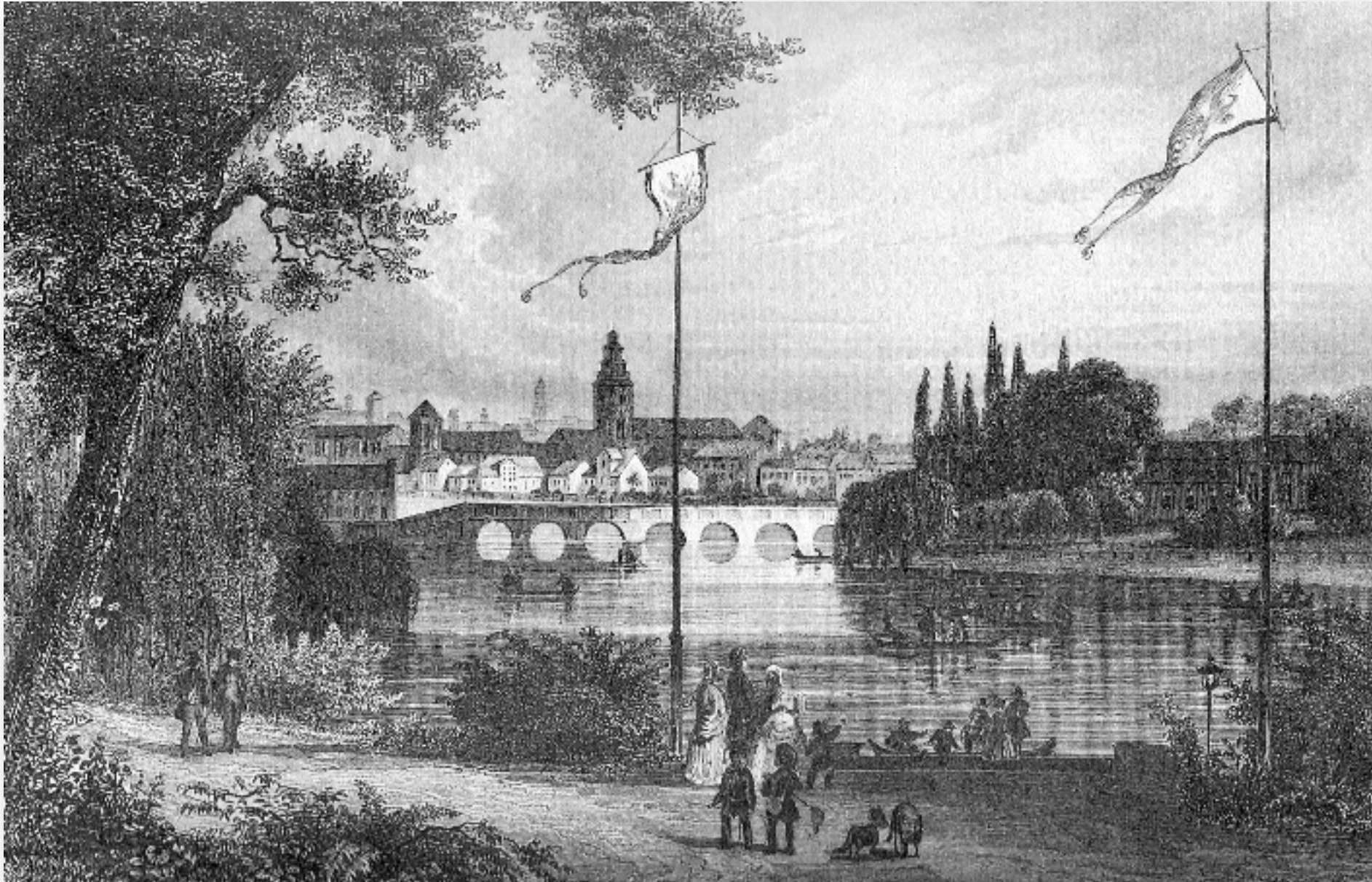


Network Analysis

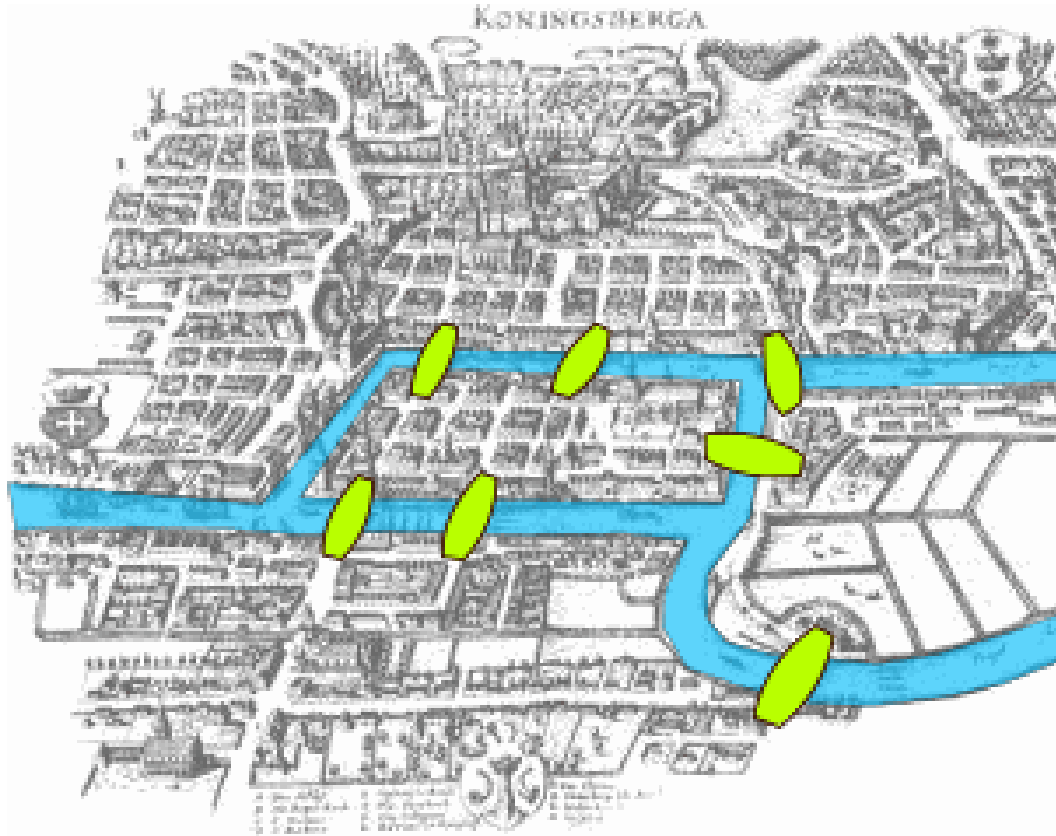
Mirco Schönfeld
University of Bayreuth

mirco.schoenfeld@uni-bayreuth.de
[@TWlY29](https://twitter.com/TWlY29)



The city of
Königsberg,
Prussia,
around 1850

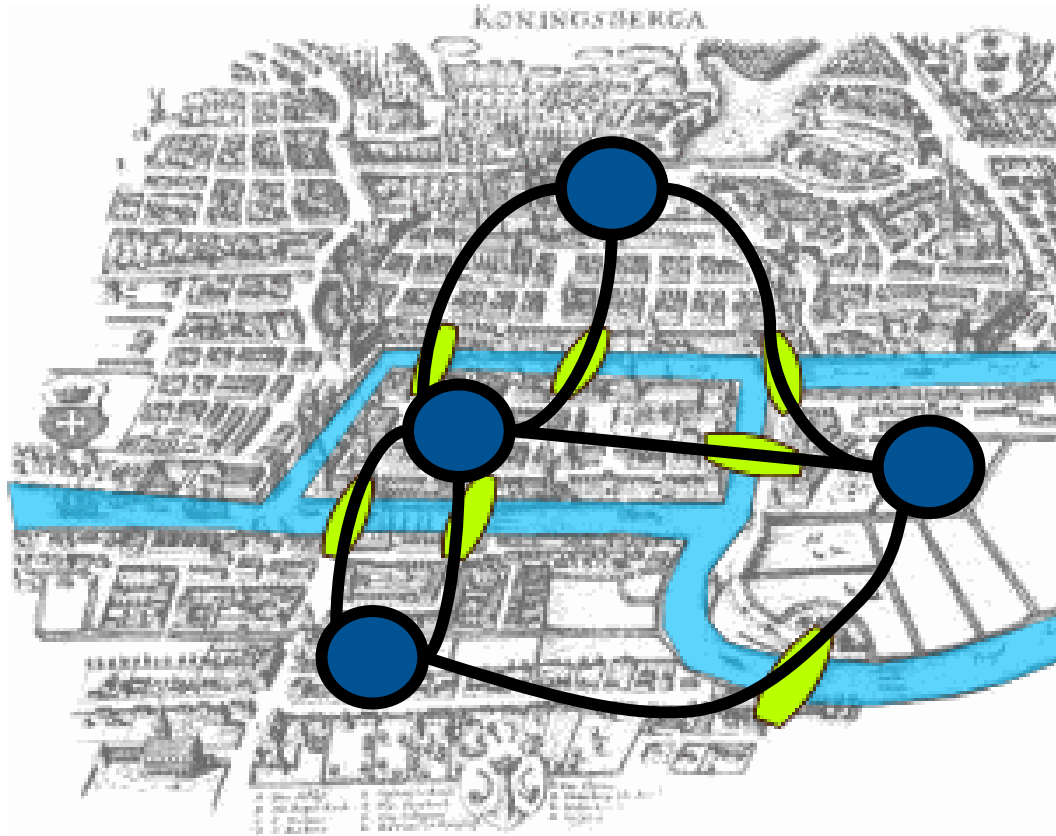
Seven bridges of Königsberg



Is there a path where you cross all seven bridges exactly once?

If so, is it also possible to do a circular route that takes you back to the starting point?

Seven bridges of Königsberg



Foundation of graph theory

by Leonhard Euler in 1736

Development of a technique of analysis:

- Replace each land mass with an abstract node
- Replace each bridge with an abstract connection

Negative resolution by Leonhard Euler in 1736

Networks, why do we care?

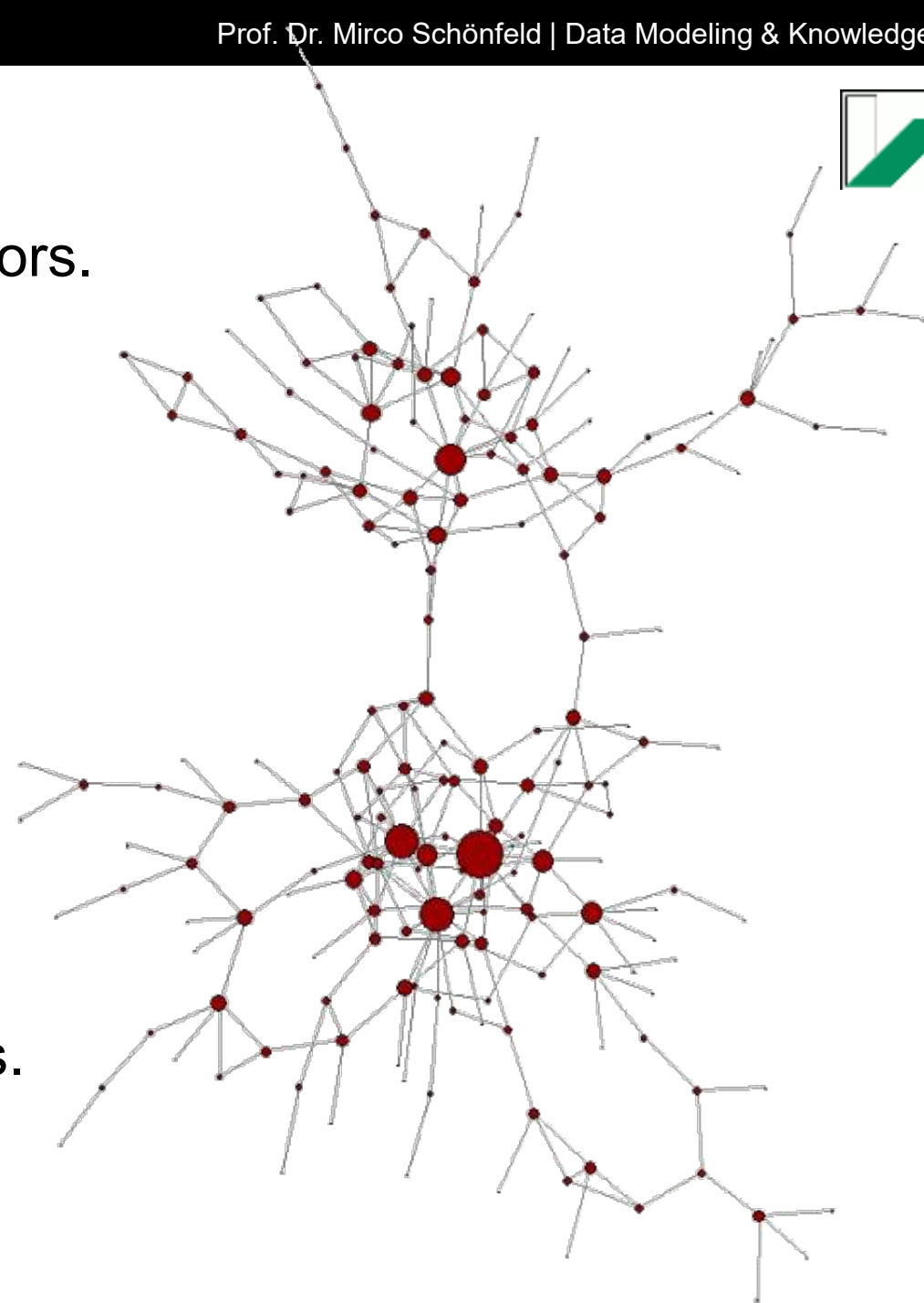
Networks are basically graphs with metaphors.

Takes context into account

More than entities and attributes:
How are the nodes related to each other?

Seemingly autonomous entities are
embedded in relations and interactions

Generic tool to modeling complex problems.



Data Questions

What are the nodes?

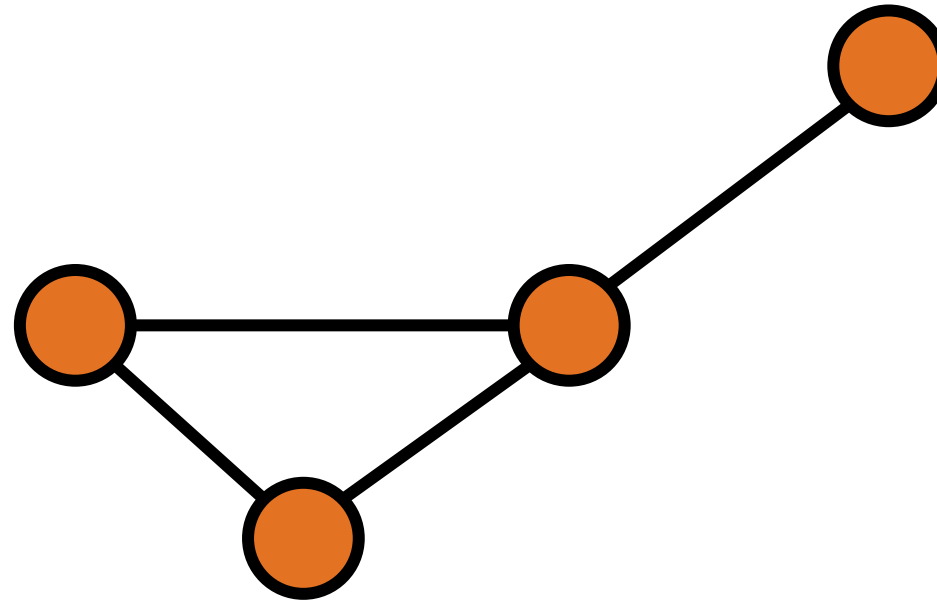
Boundary specification

What are the links?

Different types?

Is it relational data?

N : M

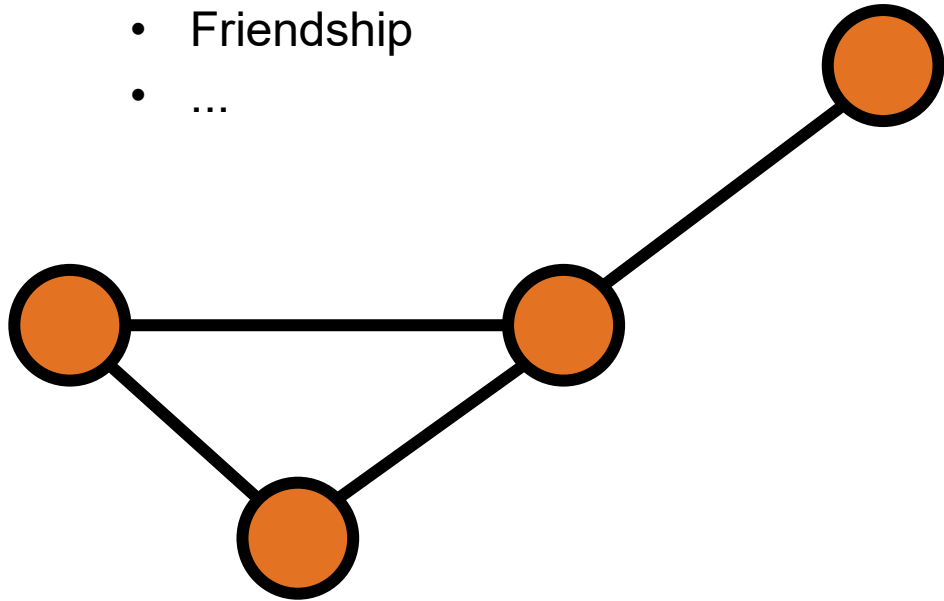


1-Mode vs. 2-Mode

1-Mode Networks

Network of people...

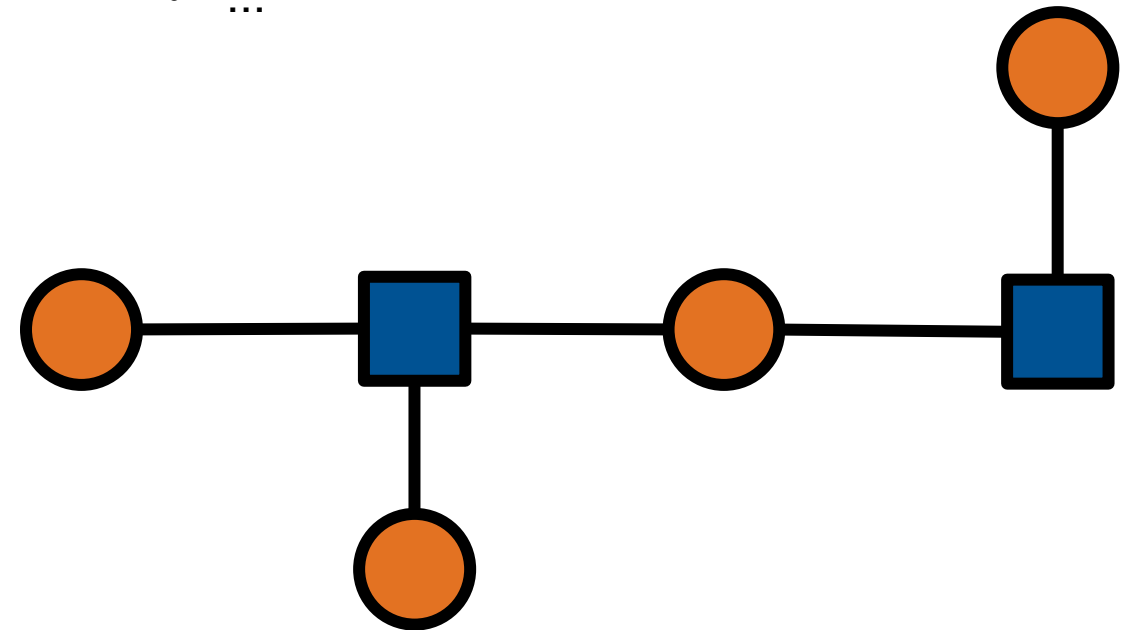
- Communication
- Friendship
- ...



2-Mode Networks

People and ...

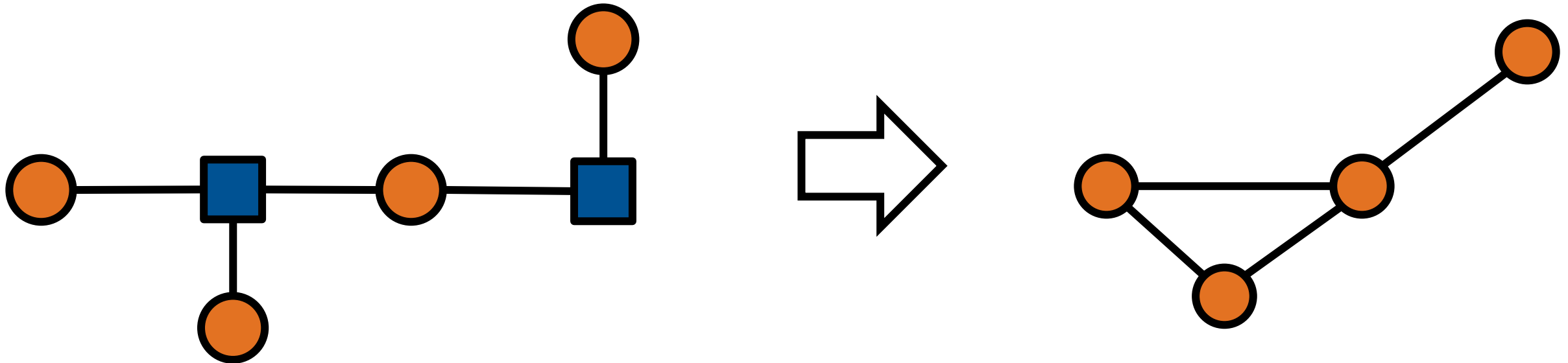
- Events
- Organizations
- Publications
- Hobbies
- ...



2-Mode to 1-Mode

Transforming 2-mode data to 1-mode data

Transform = Fold



Paths & routes

Vocabulary

Neighbors - directly connected

Path = node-link-node-link...

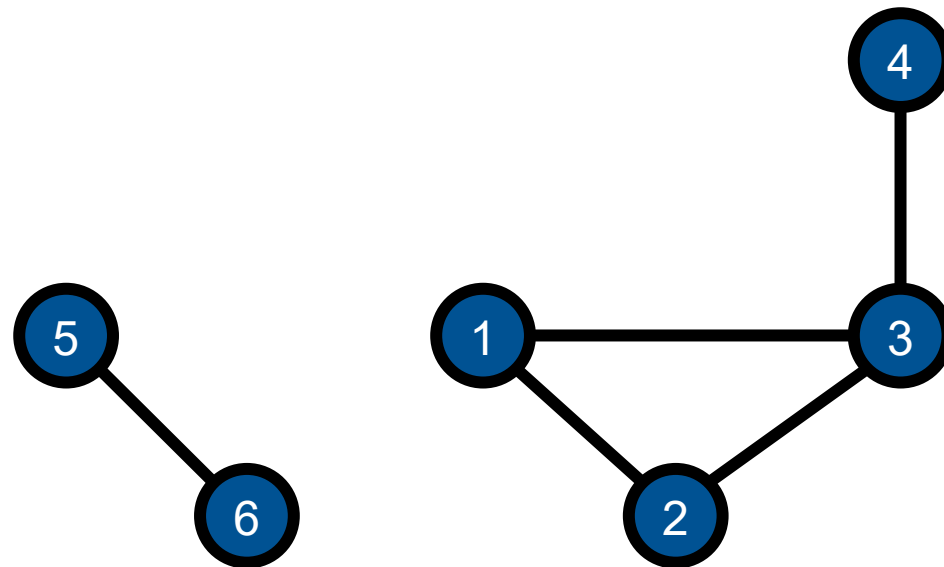
Indirectly connected - reachable, unreachable

Shortest paths

Diameter = longest shortest path (geodesic distance)

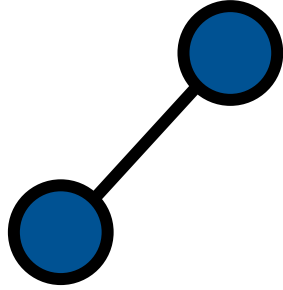
Characteristic path length = average shortest path

Component = set of reachable nodes

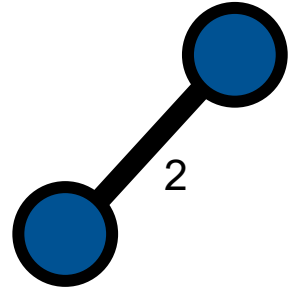


Vocabulary (cont.)

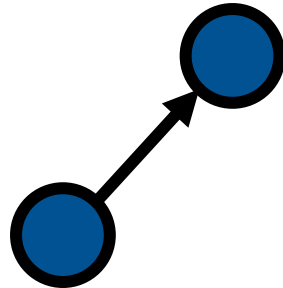
Undirected/symmetric



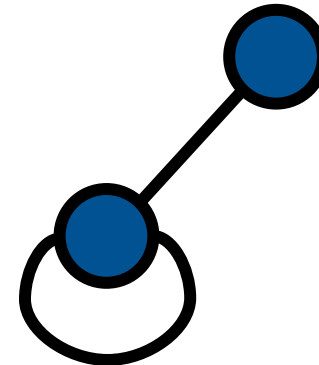
Unweighted/weighted links



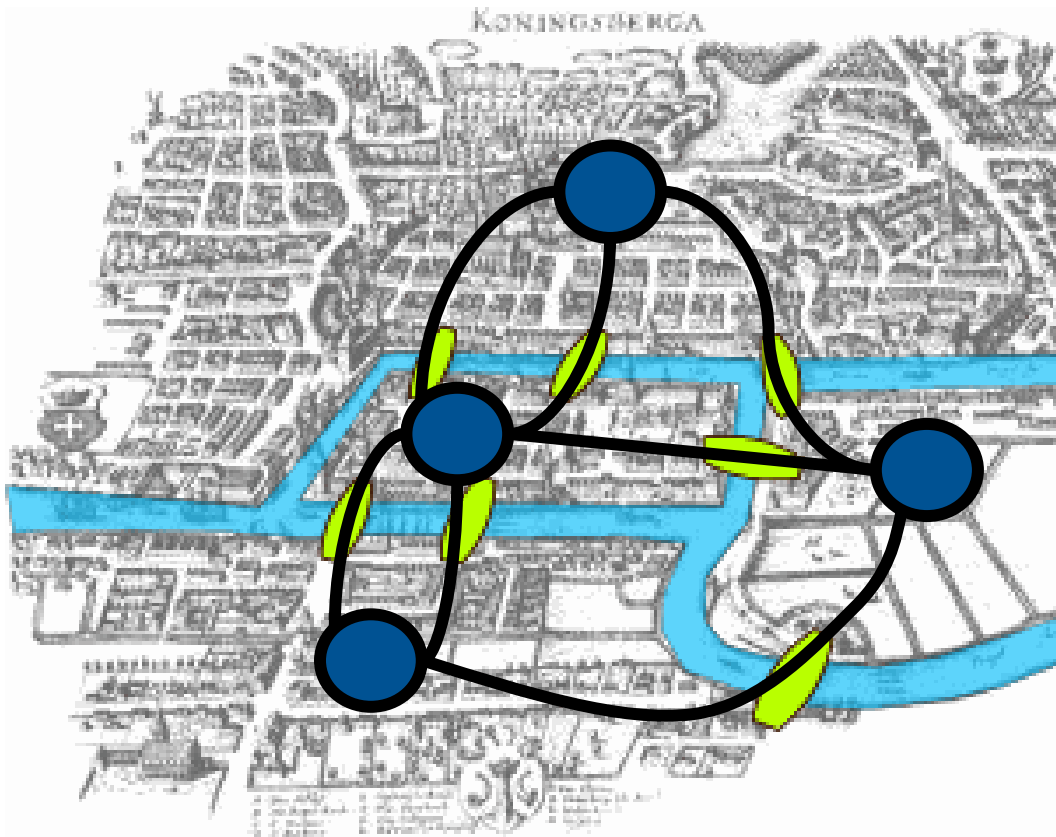
Directed links/asymmetric



Self-loops



Seven bridges of Königsberg



The Eulerian Path

by Leonhard Euler in 1736

Possibility of a walk through a graph, traversing each edge exactly once, depends on the degrees of nodes:

The graph

- ...is connected
- ...has zero or two nodes of odd degree

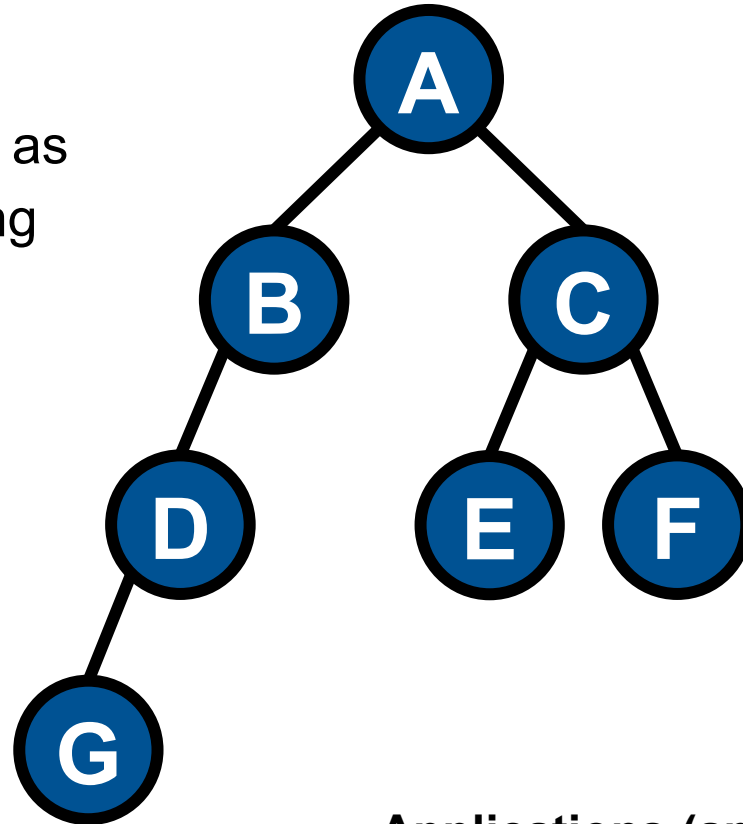
If there are nodes of odd degree, any Eulerian path will start at one of them and end at the other.

For an Eulerian circuit, the graph needs to be connected and have zero nodes of odd degree.

Search algorithms

Depth-first search

Explore each branch as far as possible before backtracking



Order of processing the nodes:

A B D G C E F

DFS is implemented using a stack.

Traversing the entire graph takes time $O(|V| + |E|)$

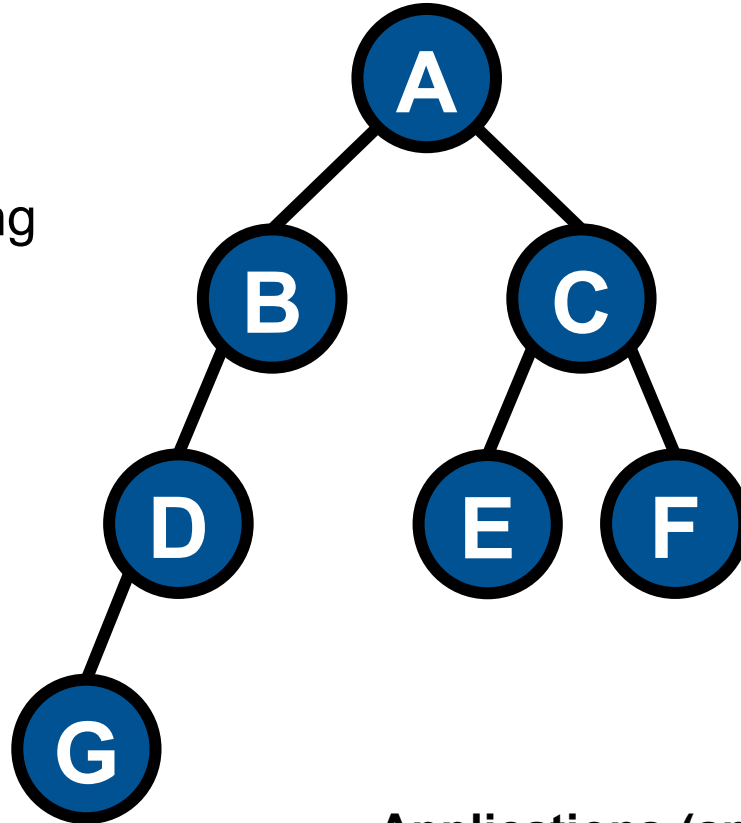
Applications (among others):

Solving puzzles with only one solution, e.g. mazes

Search algorithms

Breadth-first search

Explore all nodes at the present depth before moving on to the nodes at the next depth level.



Order of processing the nodes:

A B C D E F G

BFS is implemented using a queue.

Traversing the entire graph takes time $O(|V| + |E|)$

Applications (among others):

Finding shortest paths between two nodes

Solving games with more than one solution

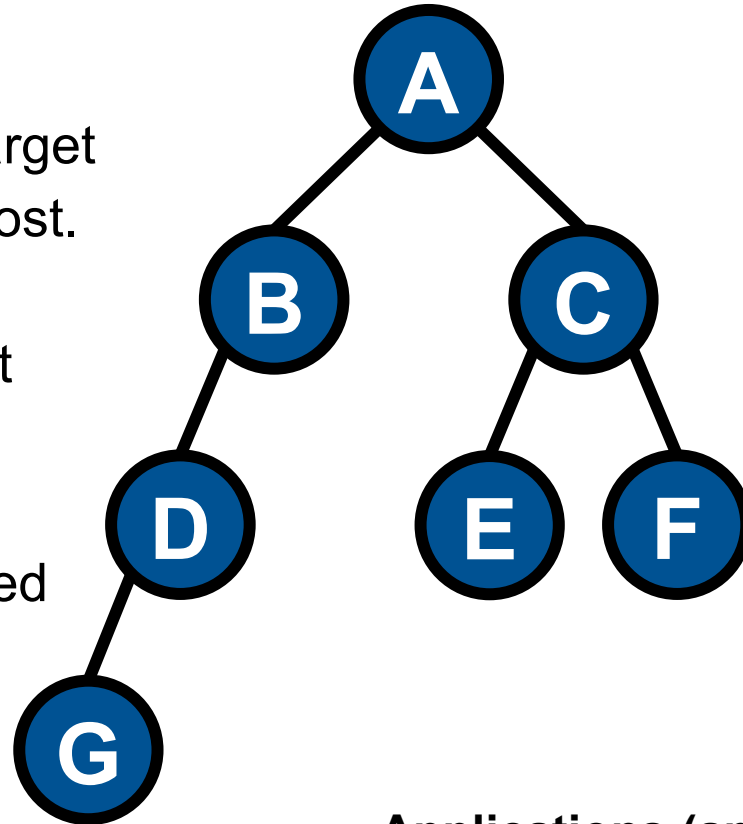
Search algorithms

A* (A star)

Finds a path to the given target node having the smallest cost.

A* is proven to find the best possible solution.

Is an example of an informed search algorithm or best-first search



A* uses a priority queue.

At each iteration, a cost function is used to select the extension of the path:

$$f(n) = g(n) + h(n)$$

$g(n)$ is the cost of the path from start to n

$h(n)$ is a heuristic function estimating the cheapest path from n to goal. Can use edge weights, for example

Applications (among others):

Finding shortest paths between two nodes

Natural Language Processing

Decision making



Applications

- Finding directions / routing

Road networks are graphs with positive weights.

Nodes represent road junctions and edges represent road segments between junctions.

Weights of edges may correspond to lengths of road segments, time needed to traverse segment, costs of traversing the segment, or else

Directed edges model one-way streets

- Improving decision making

Nodes describe states, edges describe possible transitions, e.g. single moves or turns in a Rubik cube

Shortest path represents a solution with minimal number of moves

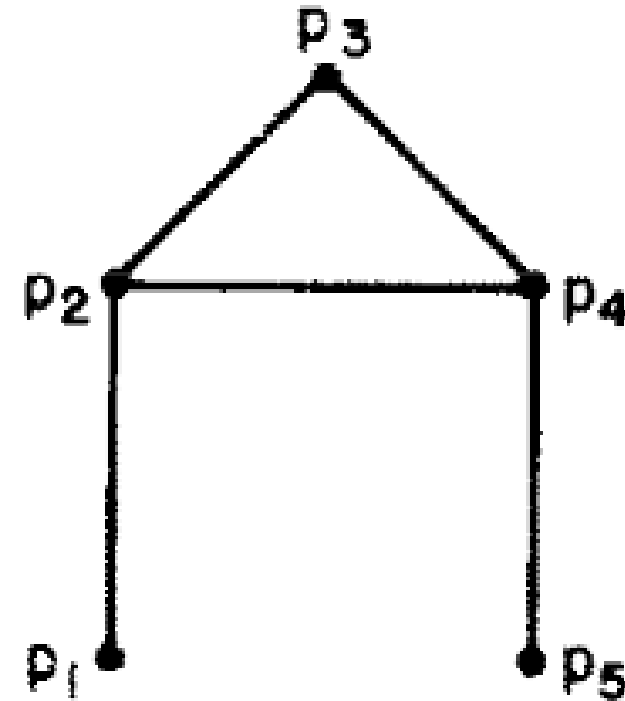
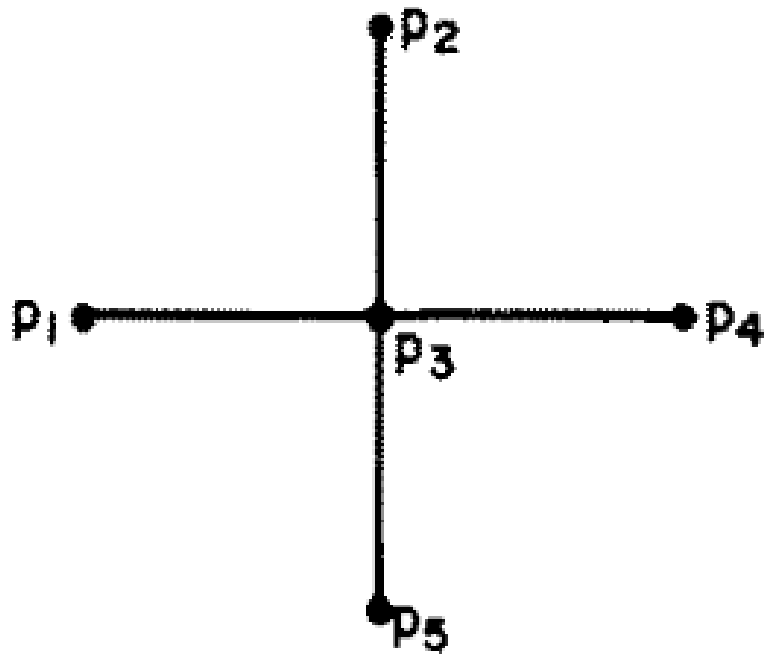
- And more

Examples: planning facility layout, logistics, strategic shortest paths (e.g. travelling salesman, Canadian traveller),...

Centralities

What is Central in a Network?

“The point at the center of a star or wheel [...] is the most central position”



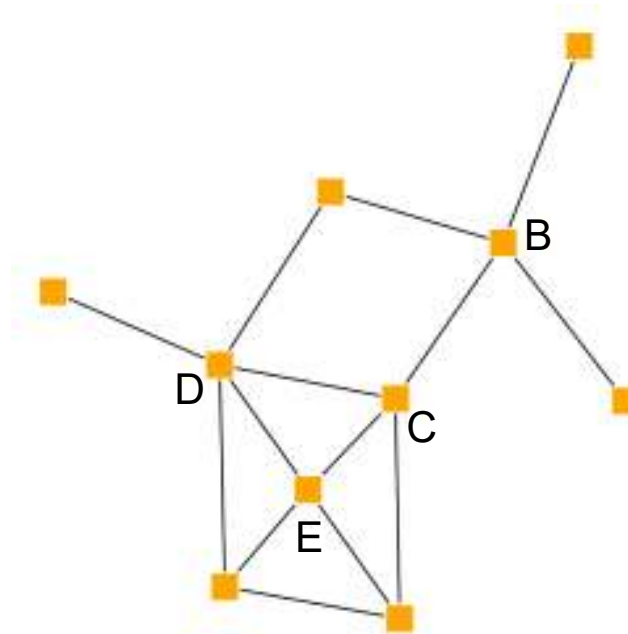
What is Central in a Network?

The central point has the largest degree
Well connected to many other nodes

It is on the most geodesic paths
Involved in many inter-network communications

It is as close to all points as possible
Short path lengths

It is connected to other important nodes
Recursive definition of importance



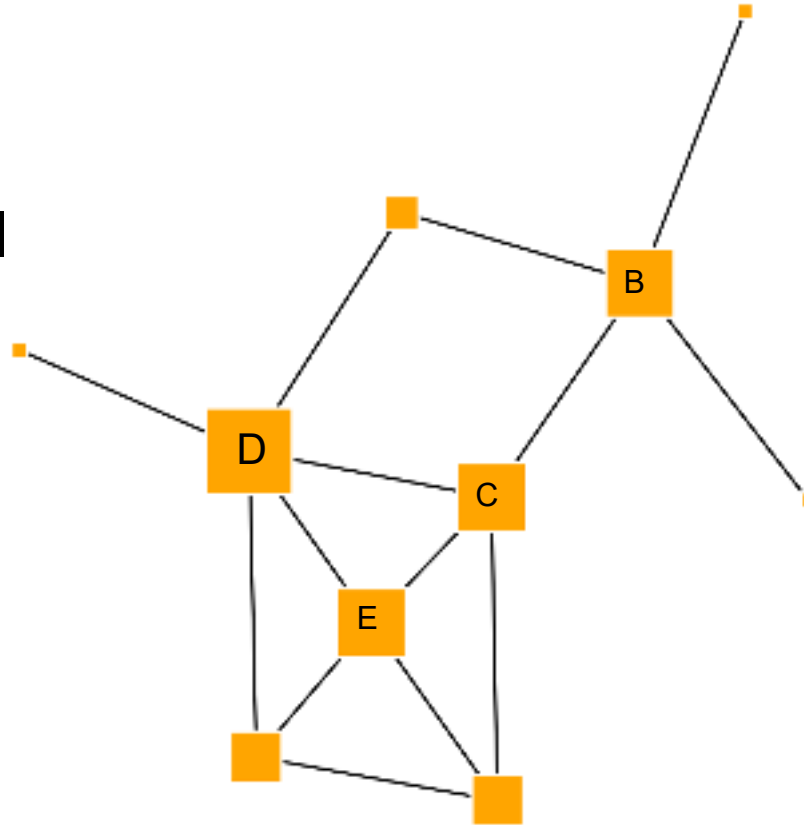
Definition 1: Largest Degree

Motivation

- Most contacts
- highest activity

Higher degree -> more central

Degree centrality



$$C_D(n_k) = \sum_{i=1}^n a(n_i, n_k)$$

$$a(n_i, n_k) = \begin{cases} 1 & \text{if } i \text{ and } k \text{ are connected} \\ 0 & \text{else} \end{cases}$$

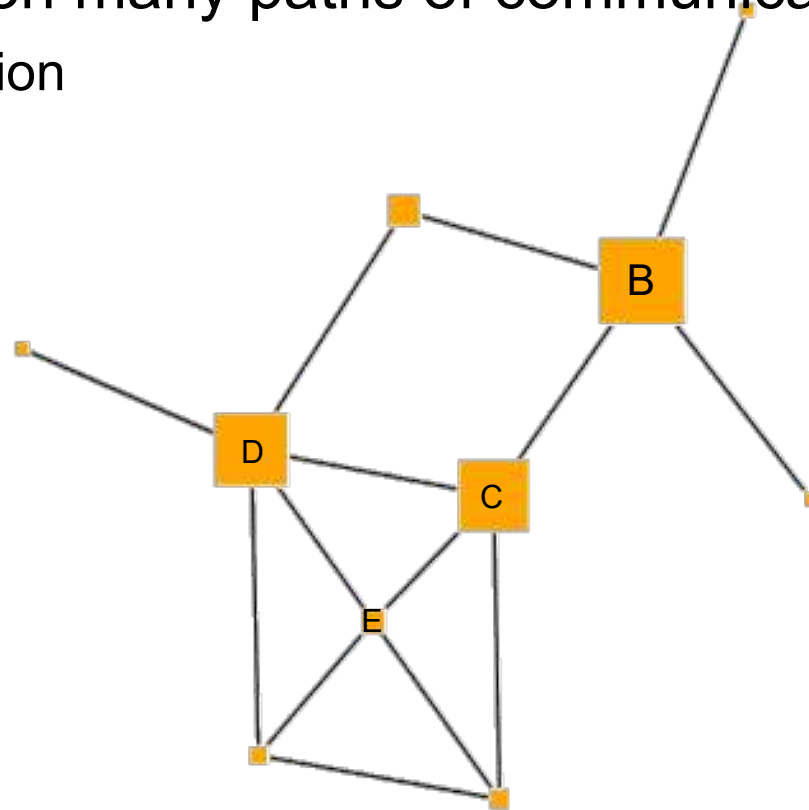
Definition 2: Most geodesics (shortest paths)

Motivation:

A person is central if she lies on many paths of communication

- Can withhold or distort information
- Control

Betweenness Centrality



Number of shortest paths
from i to j including n

Number of shortest paths
from i to j

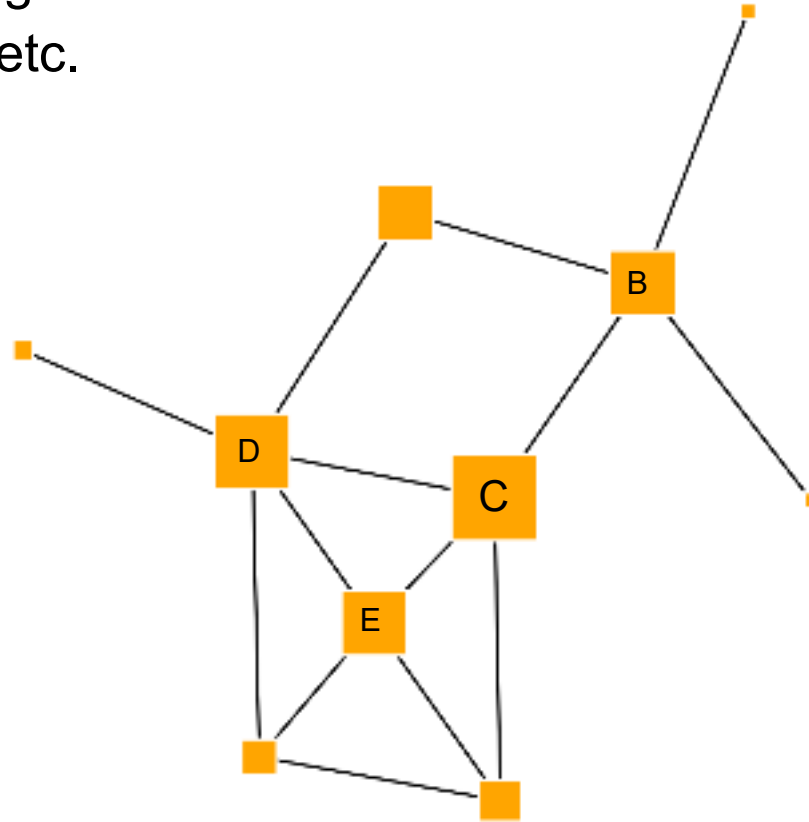
$$C_B(n_k) = \sum_{i < j} \frac{g_{ij}(n_k)}{g_{ij}}$$

Definition 3: Shortest distance to all other points

Motivation:

- Being close to everybody else is good
- Close to information, resources, etc.

Closeness Centrality



Sum up all geodesic paths from the point (k) to all other points

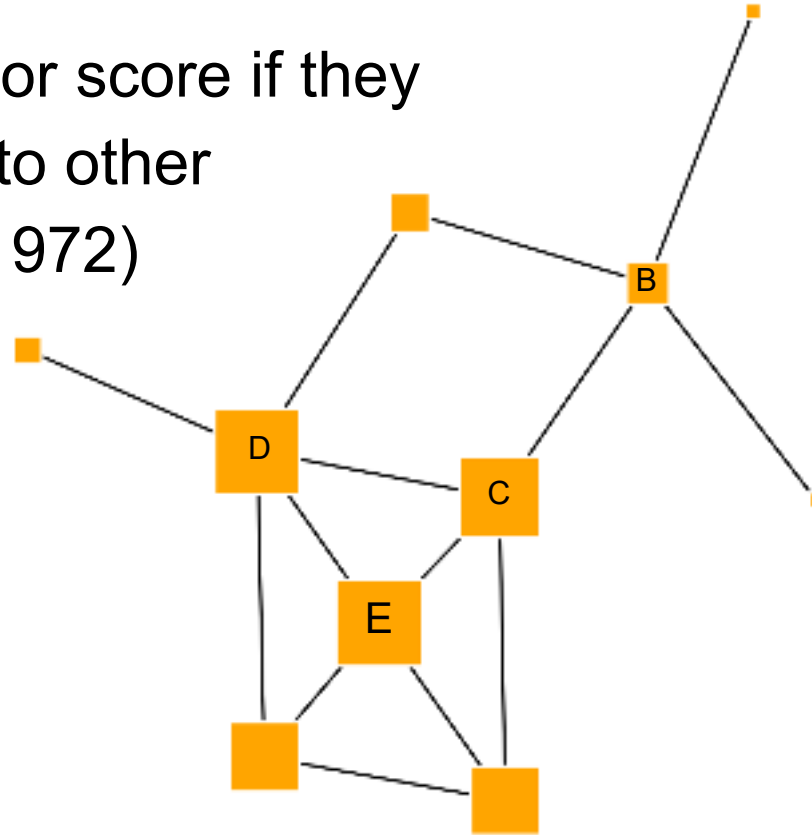
$$C_C(n_k) = \frac{1}{\sum_{i \neq k} d(n_k, n_i)}$$

Definition 4: Being linked to by other important nodes

Eigenvector centrality is based on eigenvector calculation in linear algebra.

Agents have a high eigenvector score if they are important and connected to other important agents (Bonacich, 1972)

Eigenvector centrality



$$C_E(u) = \frac{1}{\lambda} \sum_{v=1}^{|V|} w_{uv} C_E(v)$$

where λ is a constant. We can rewrite the equation as:

$$\lambda C_E = W \cdot C_E$$

Applications and Limitations

Limitations:

- Choice of centrality metric is highly dependent on the context
- Centrality indices indicate most important nodes
 - No relative importance (not quantification of difference in importance)
 - Features identifying most important nodes may be meaningless for other nodes

Applications:

- Identify most influential person(s) in social networks
- Identify key nodes in infrastructure networks (Internet, urban, disease networks...)
- Identify important web sites (e.g. a variant of Eigenvector centrality called *PageRank* was used to rank search results in Google)

Groups

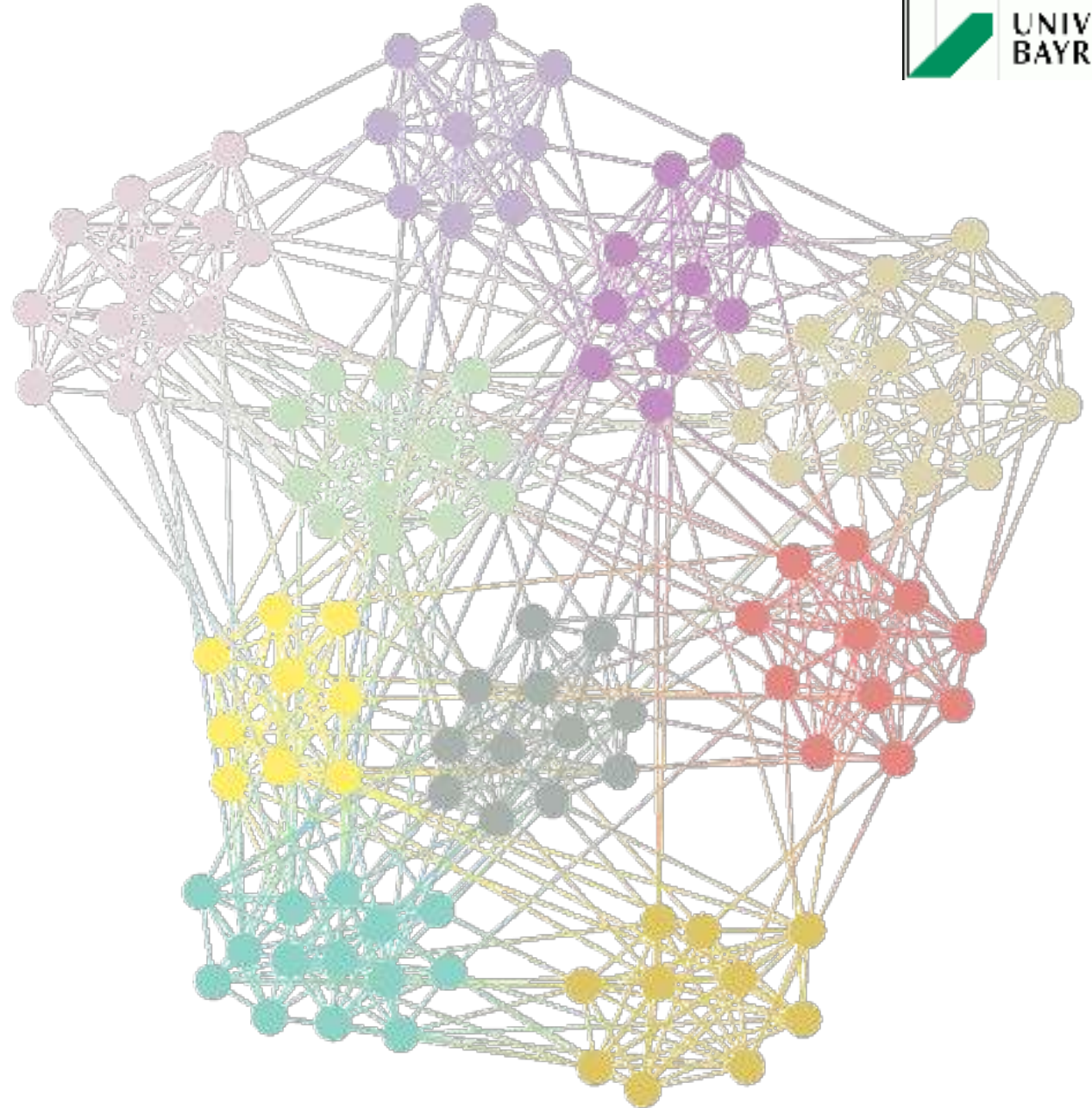
Groups & Communities

Different definitions of groups

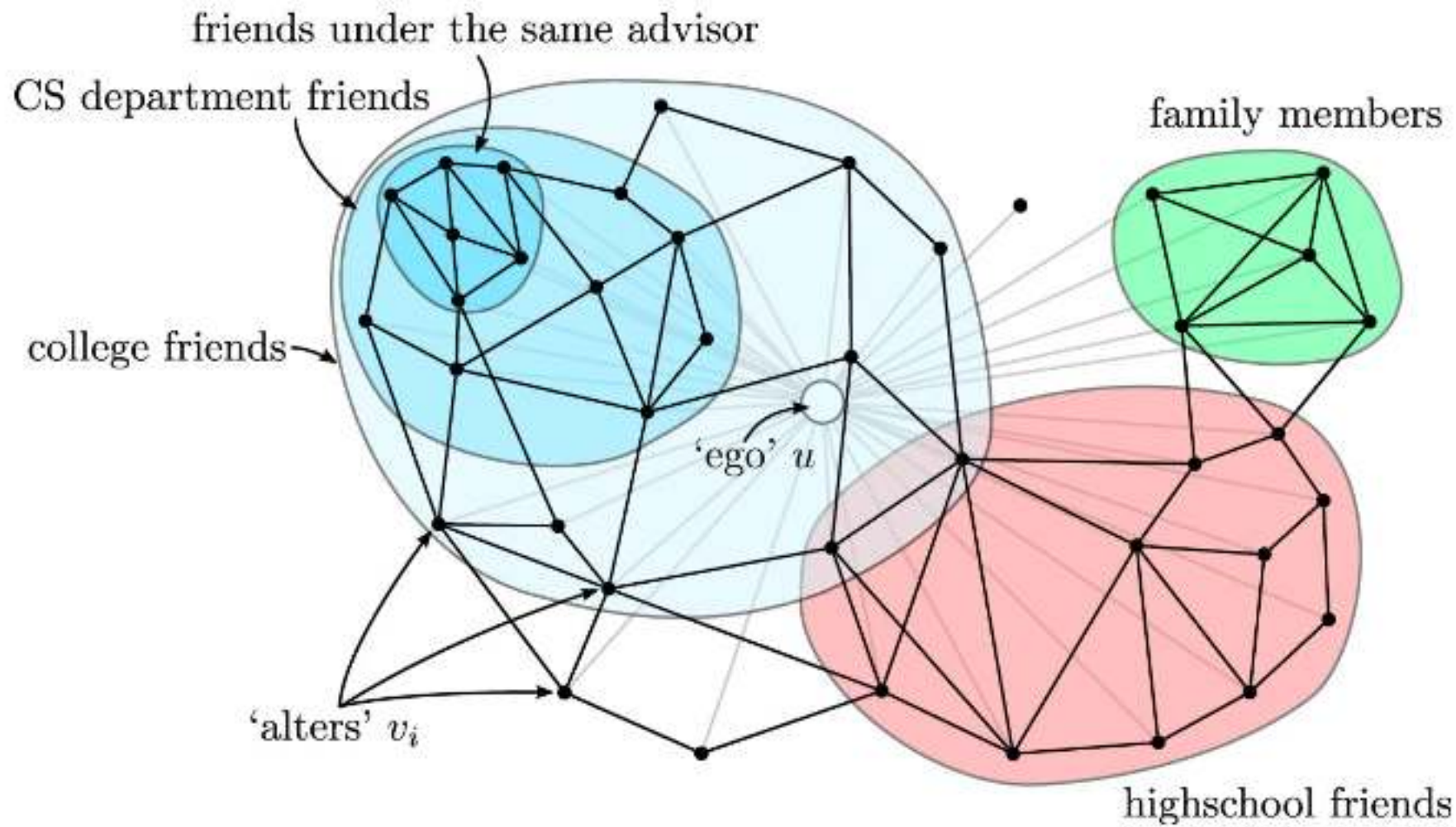
How to detect communities?

Different algorithms for different
community definitions

To summarize / predict the high level
structure of the graph

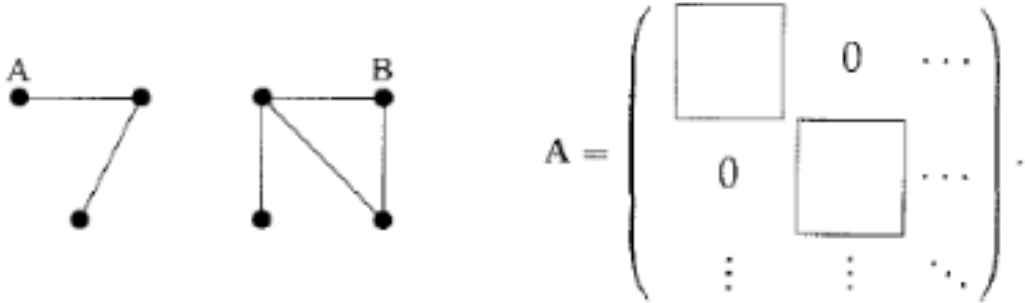


Groups & Communities



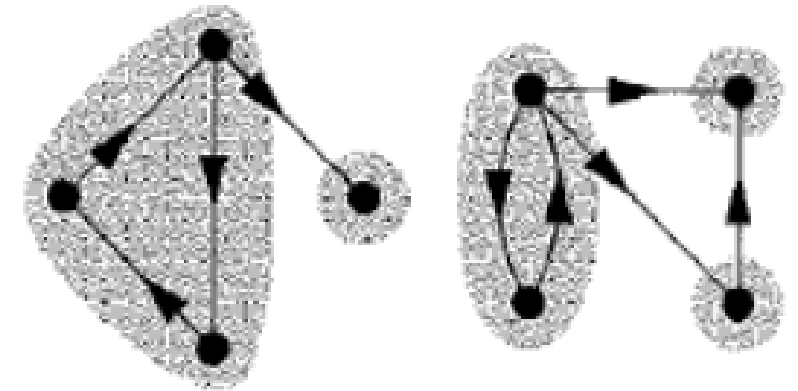
Components

Disconnected network



What about directed networks (e.g. www)?

- 2 undirected (weak) components
- 5 directed (strong) components
- Contain cycles for (every) node



Acyclic directed networks have no strong components

Out-Component

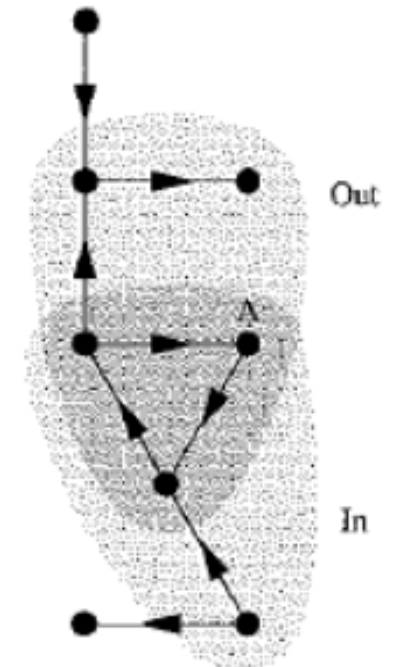
How far can you get from one point (no walk back)?

Reachability of nodes

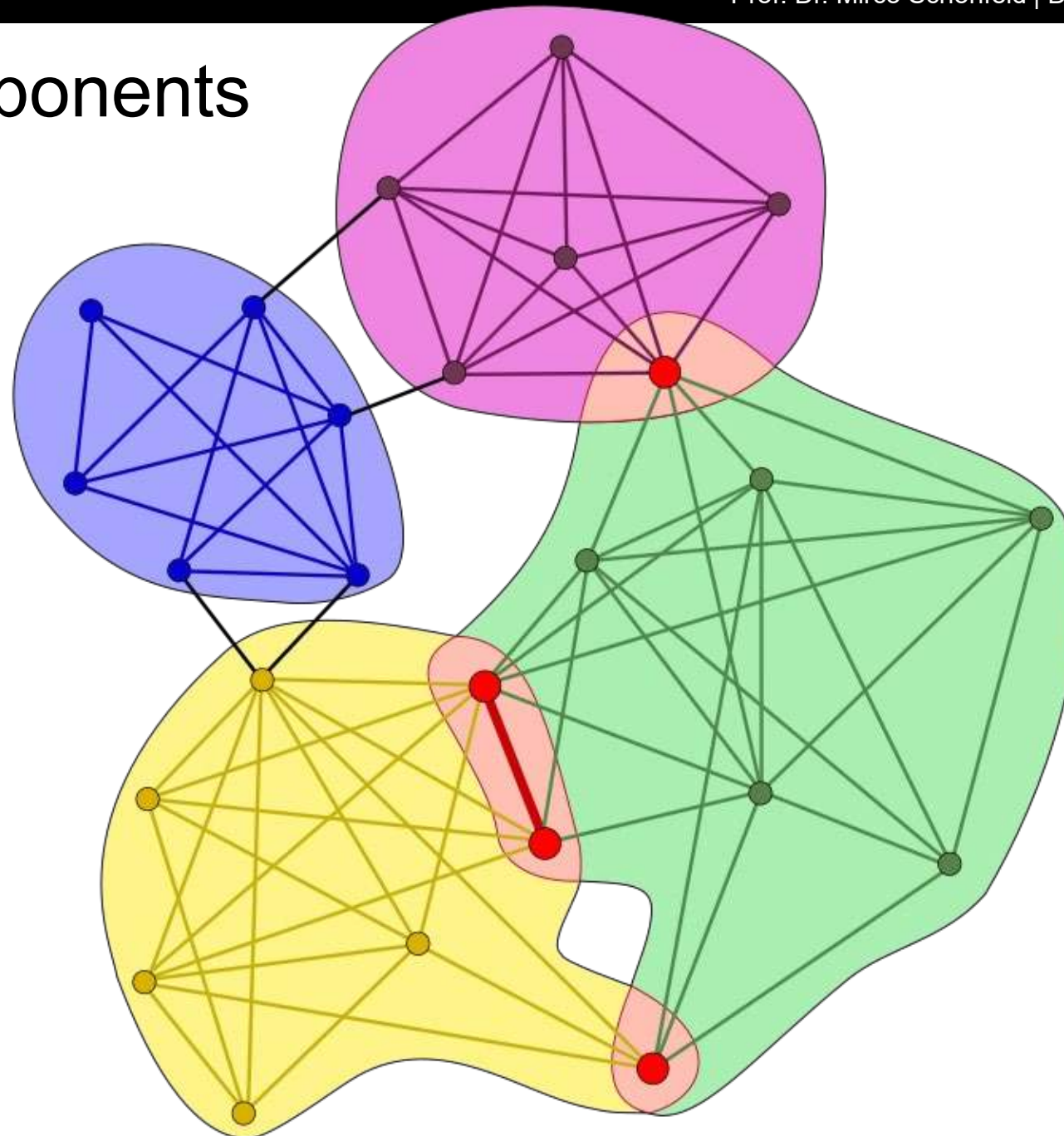


Strongly connected components have identical out-components

The intersection of in- and out-components are strong components



Visualize Components



Community Detection

Goals:

- Separate graph into groups of nodes
- Minimum connections among the groups
- No fixed number of groups or group size

Understanding large-scale structure of networks

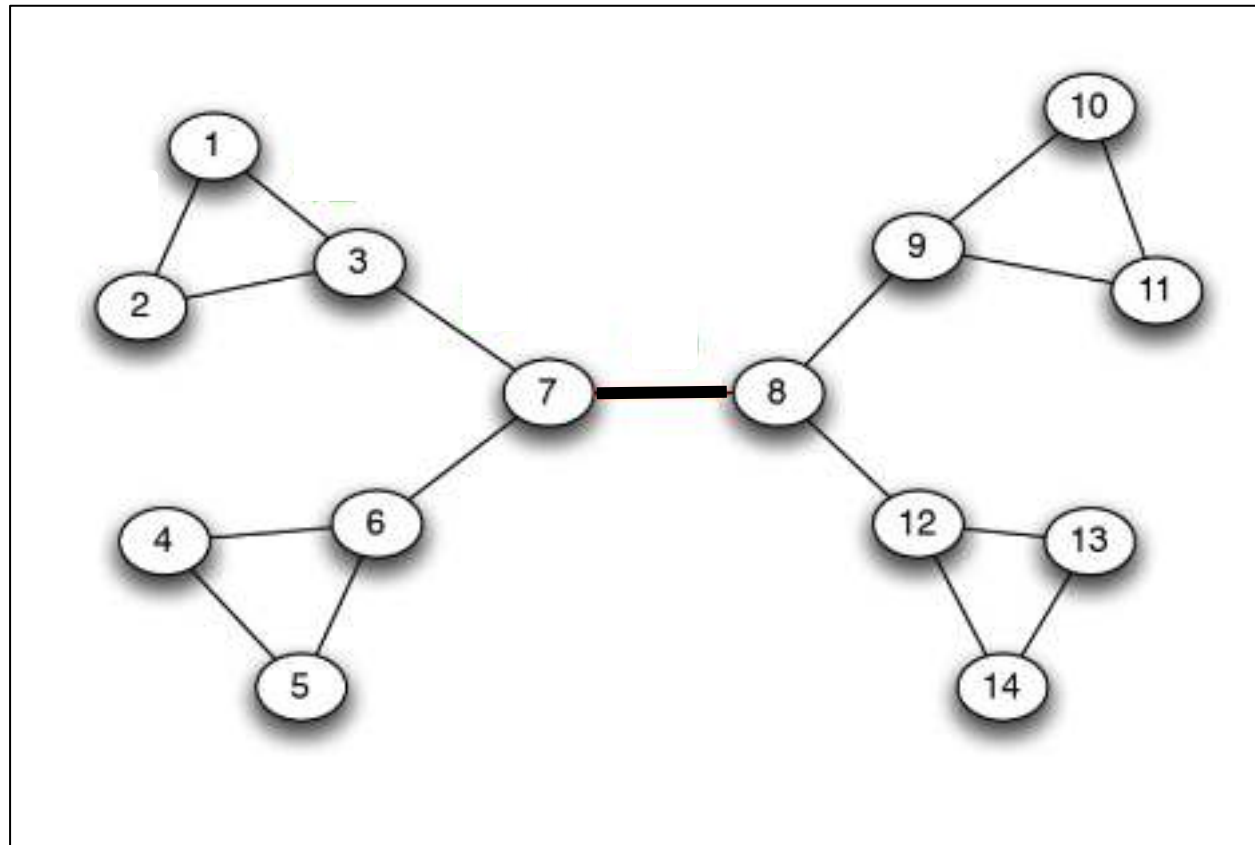
Newman/Girvan Grouping

Newman & Girvan [2004]

- Calculate edge betweenness centrality
- Remove edge with highest betweenness centrality
- Repeat process

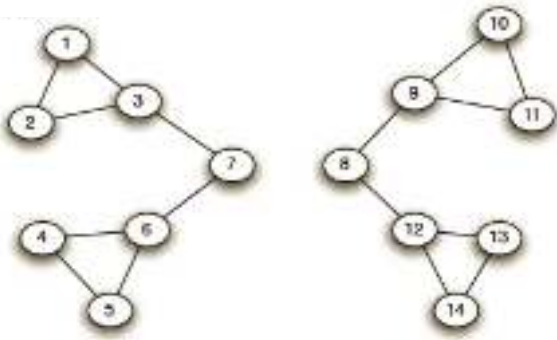
When to stop?

→ K-groups or modularity

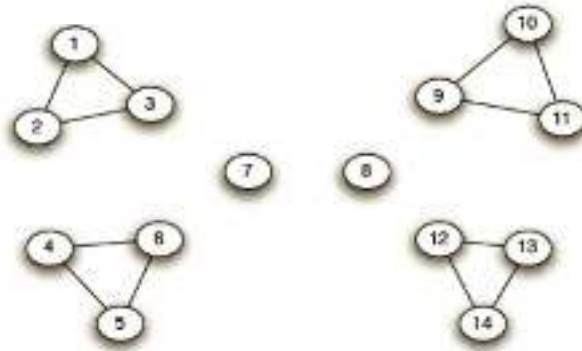


Newman/Girvan Grouping

Step 1



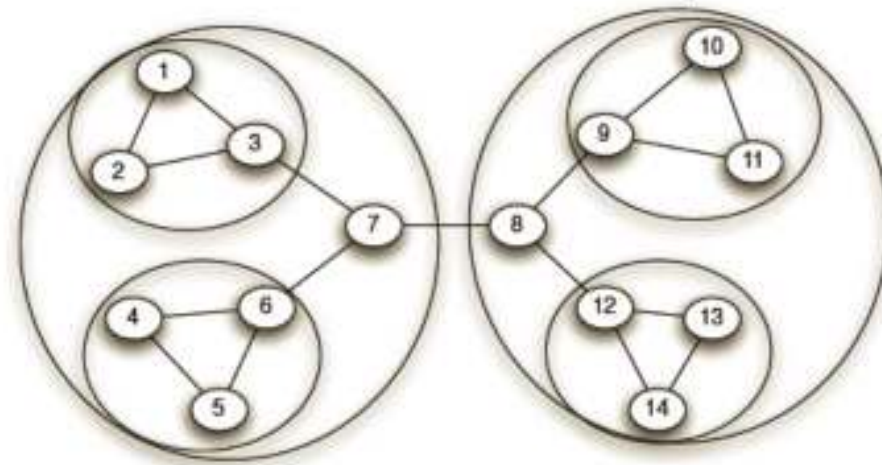
Step 2



Step 3



Hierarchical network decomposition





How to Evaluate Grouping?

Fewer links between groups “than expected”

Count links within and between groups

All = within + between

→ Goal: Optimize links within groups compared to what is expected

Modularity maximization: most commonly used

Perfect solution = exponential time complexity

Efficient heuristic optimization algorithms

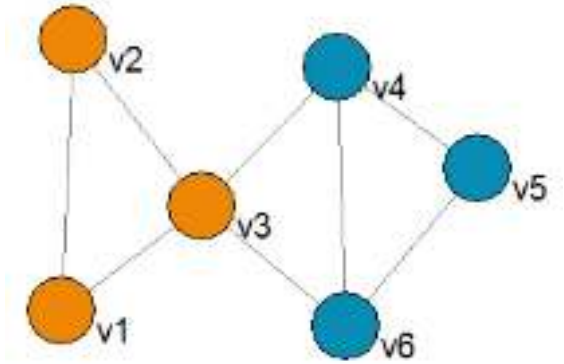
Modularity

Fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random

Degrees k_i and k_j

$s_i = 1$ for group 1, $s_j = -1$ for group 2

$2m =$ number of ends of edges



$$\frac{1}{4m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j$$



Simple Modularity Maximization

Two random communities of equal size

Algorithm:

- For every node:
 - How much would modularity change if node would move
 - Move best node
- Repeat until no improvement

No constraint on group size

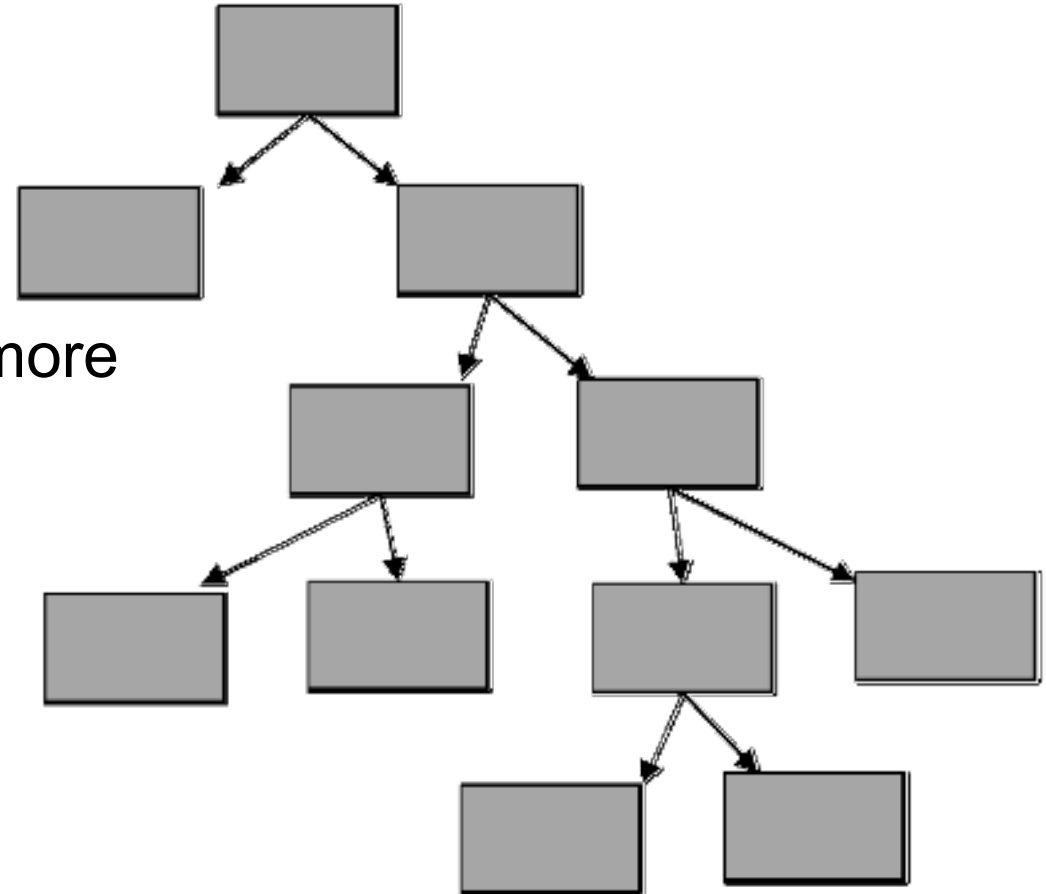
Quite fast $O(nm)$

More Than Two Groups

Modularity maximization works

Repeatedly bisecting the network

Stop when modularity does not increase anymore





Important to Know!

Community/Partitioning algorithm find a best solution
Regardless whether a good solution exists!

Modularity value serves as a kind of a significance level for clustering

Attributed networks

Attributed networks

Nodes and/or edges have attributes assigned

Examples for node attributes:

- user content in social media
- reviews in co-purchasing networks
- paper abstracts in citation networks
- content on linked web pages
- ...

Attributes are a rich set of data describing unique characteristics of nodes and edges

Customer reviews

★★★★☆ 4.7 out of 5

2,829 global ratings



~ How are ratings calculated?



Erika

★★★★★ Exactly what you want

Reviewed in the United States on 9 December 2018

Colour Name: Rot Verified Purchase



Every home should have one of these pillow cases.



167 people found this helpful



You want this. You really do. I got it for my friend as a joke but it's so good I'm considering getting one for myself too.

If I were to make one complaint, it's that the sequins aren't fully colored, and the area near where it's sewed on is white, creating many white specks that look like Nick's image. But let's be real. This isn't \$100. For the price that it is, it is GLORIOUS.

It's so much fun to play with and you can make so many looks with these.



Analyzing attributed networks

Utilize structural + attribute information

Attributes may be used in various ways:

- Guide community detection by optimizing for attribute homogeneity
- Mine patterns correlated in structural dimension and attribute dimension
- Predict links based on estimation of homophily

Thanks.

mirco.schoenfeld@uni-bayreuth.de