42578 Advanced Business Analytics

Network Analysis

Stanislav Borysov [stabo@dtu.dk]

What is a network?

Network

=

Entities
(nodes, vertices, actors, …)

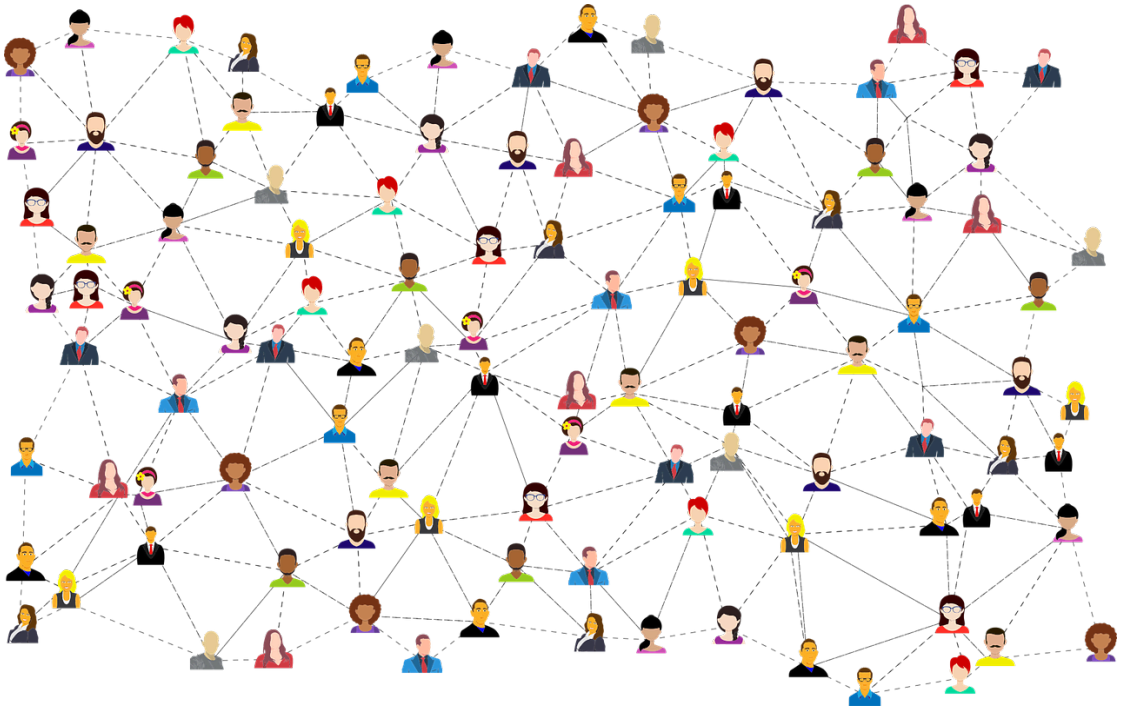
+

Relations
(links, edges, ties, connections, …)

https://www.cleanpng.com/png-complex-network-floral-design-systems-design-compu-2172128/

Networks are everywhere

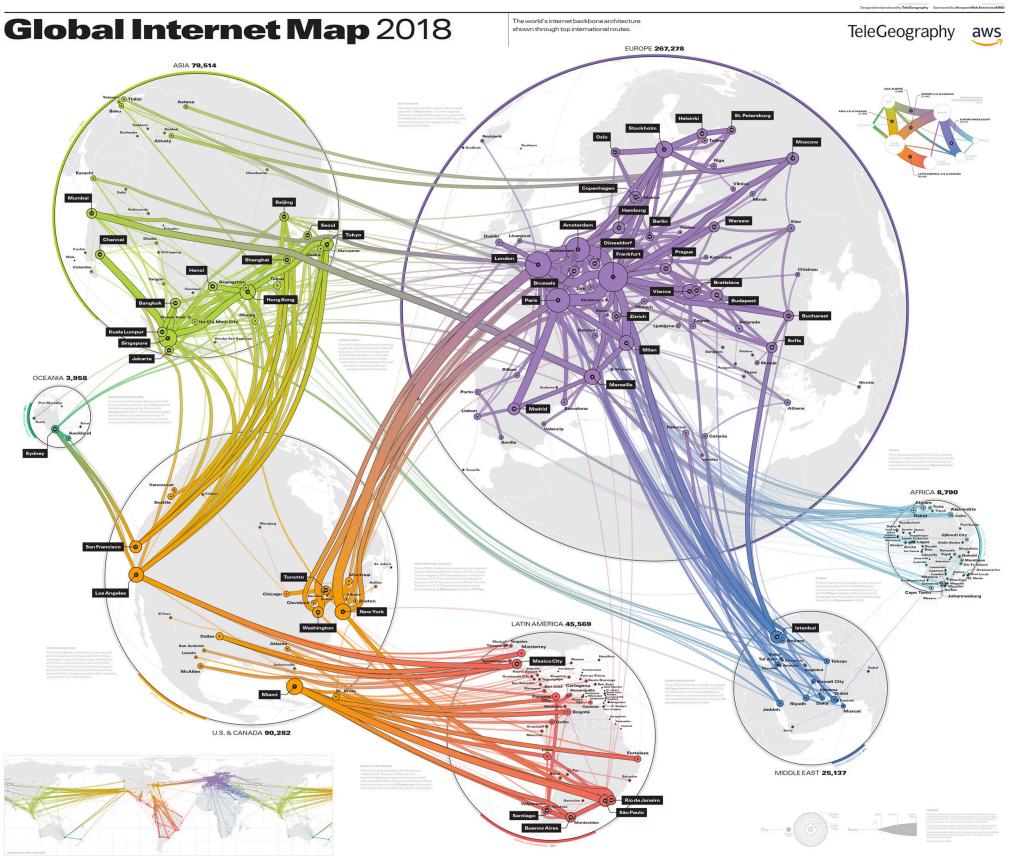
- Social networks



https://makeawebsitehub.com/social-media-sites/
https://socialmediaexplorer.com/content-sections/news-and-noise/5-up-and-coming-social-networks-you-should-know-about/

Networks are everywhere

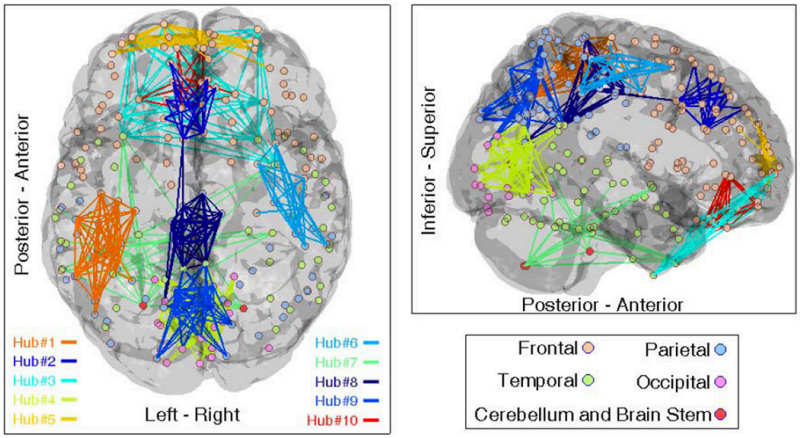
- Computer networks

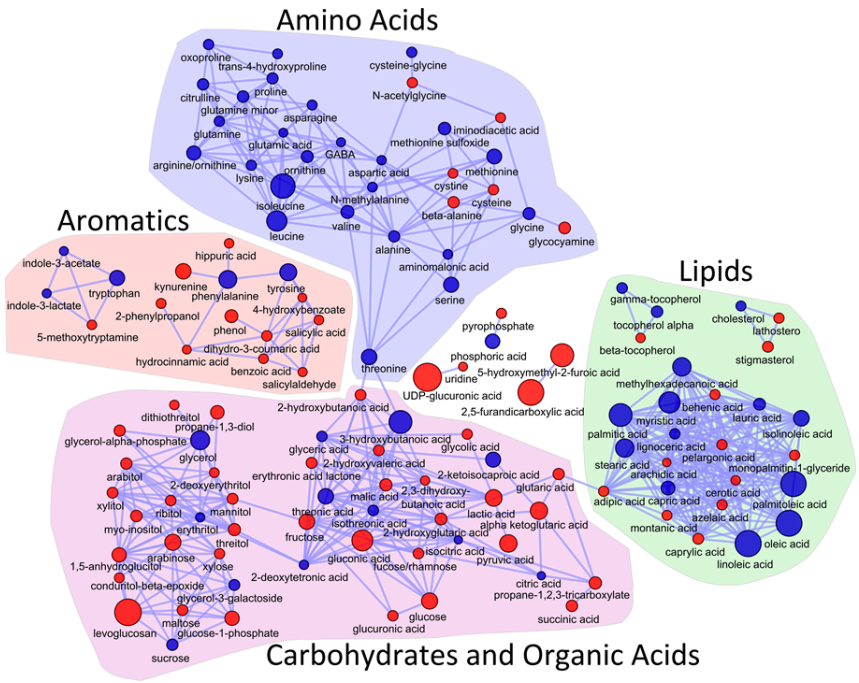


https://www2.telegeography.com/global-internet-map

Networks are everywhere

- Biological networks

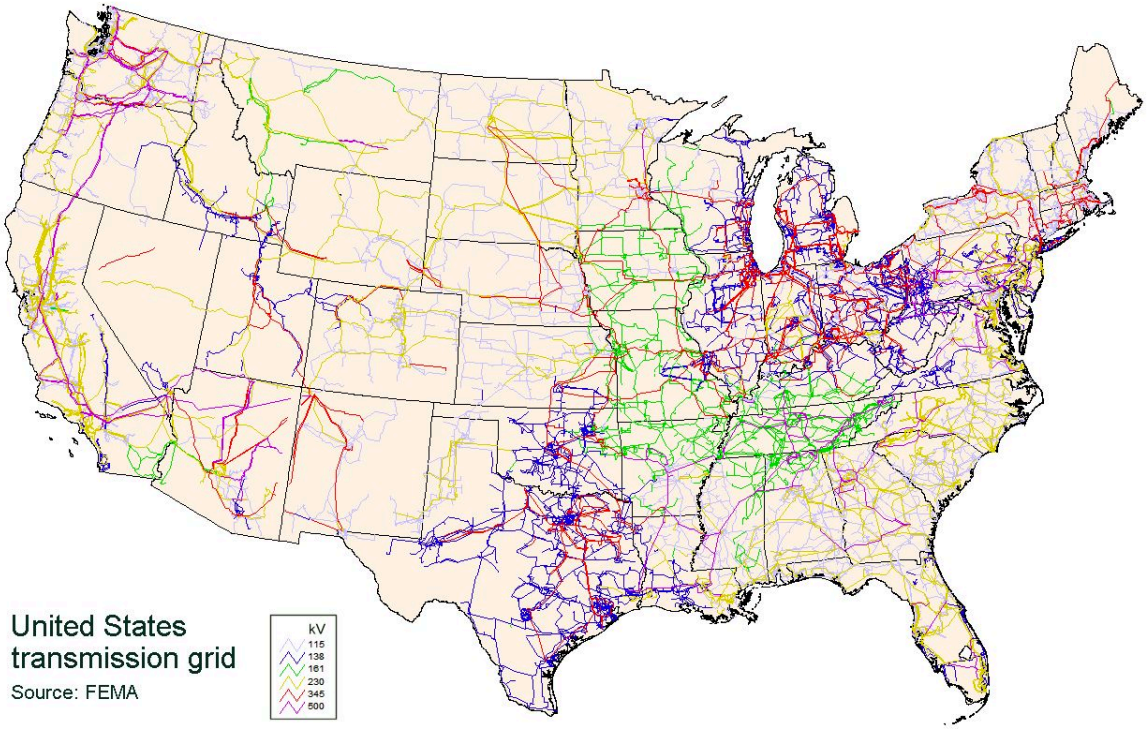




https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331879/
https://imdevsoftware.wordpress.com/2013/04/04/tutorial-building-biological-networks/

Networks are everywhere

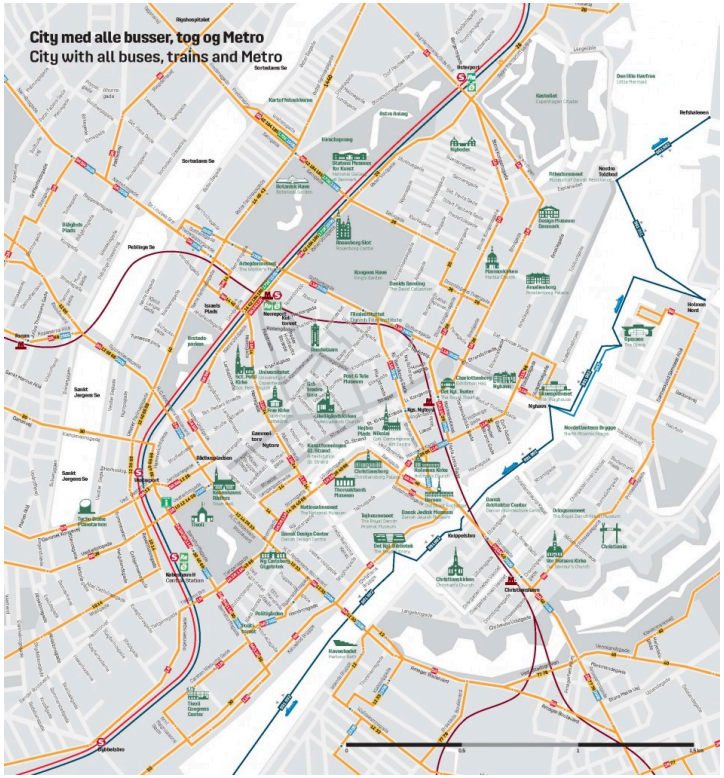
- Electrical networks



https://commons.wikimedia.org/wiki/File:UnitedStatesPowerGrid.jpg

Networks are everywhere

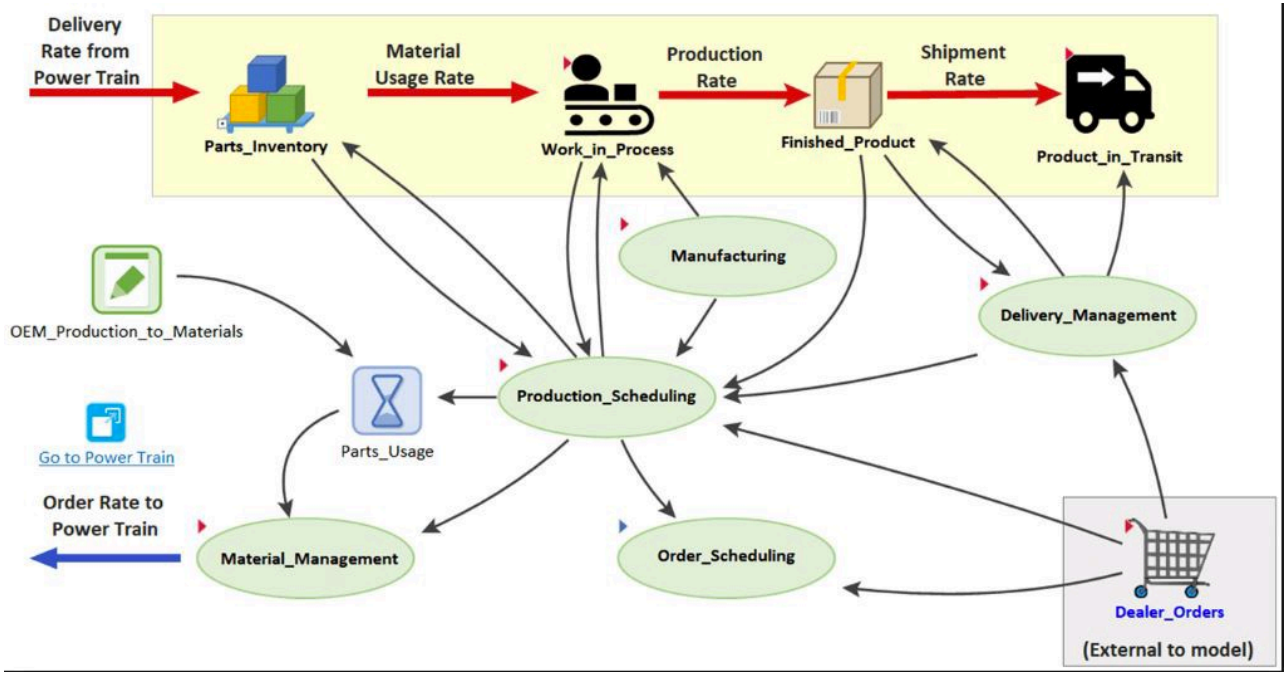
- Transport networks



http://www.scandinaviastandard.com/public-transport-in-copenhagen/

Networks are everywhere

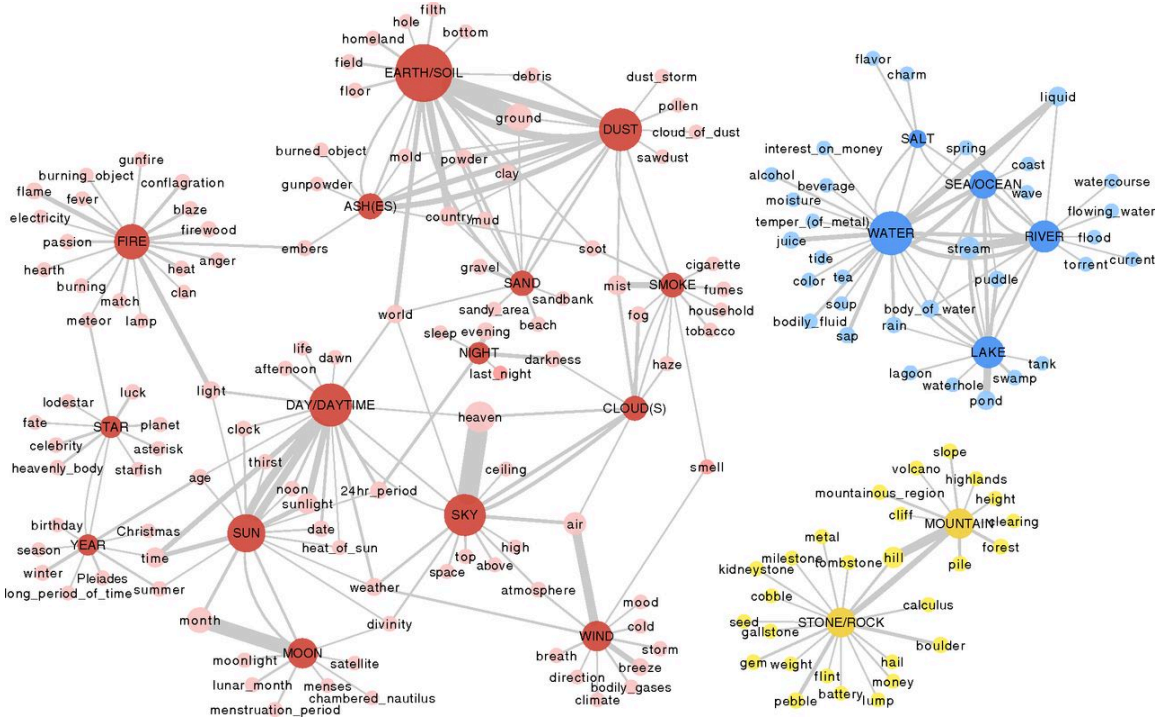
- Supply & logistic networks (spatio-temporal)



https://medium.com/analytics-vidhya/how-effectively-can-you-handle-your-supply-chain-cb4754685ca3

Networks are everywhere

- Semantic networks



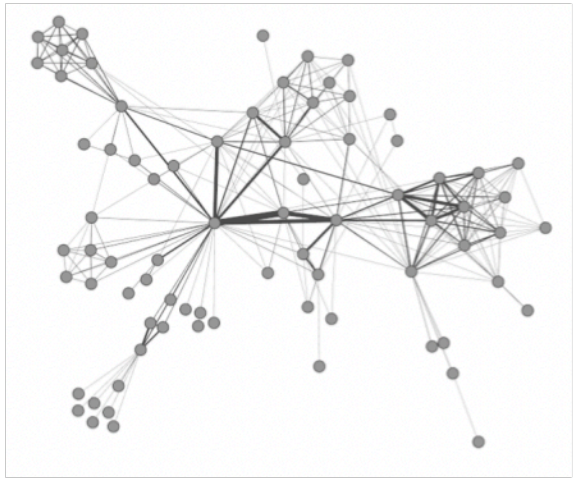
https://doi.org/10.1073/pnas.1520752113

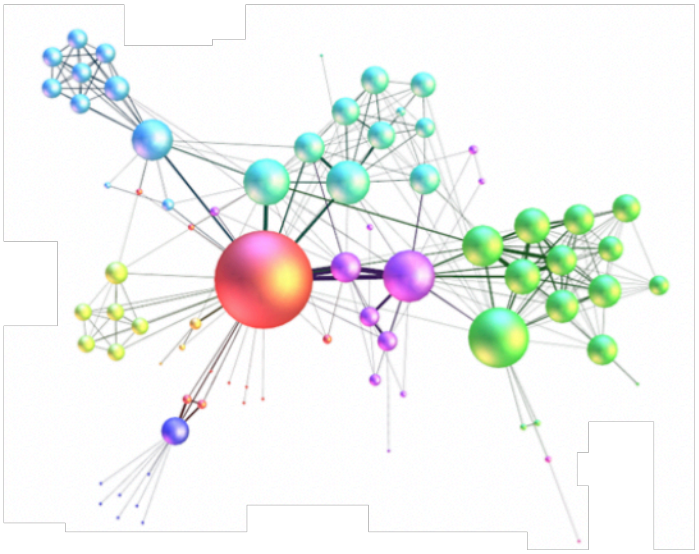
Networks are everywhere

- Physics
- Epidemiology
- Economics
- Politics
- Crime
- …

Some common tasks and application examples

- **Finding key players**
  - Target influential people for advertising
  - *"Network value of a customer is the expected profit from marketing a product to a customer, taking into account the customer's influence on the buying decisions of other customers"* (Domingos & Richardson 2001, 2002)
- **Segmentation and community detection**
  - Market, clients, organisation, complex business processes

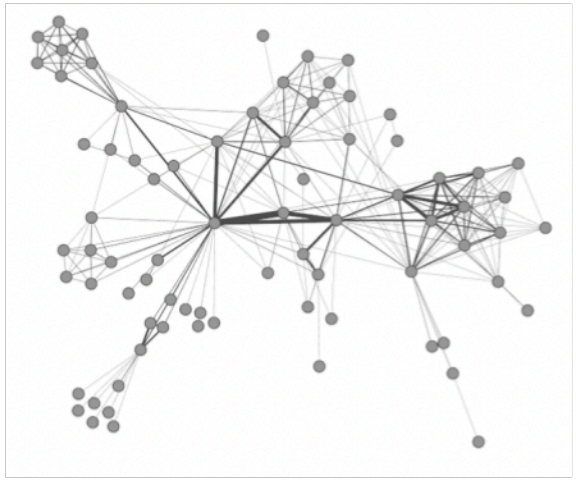




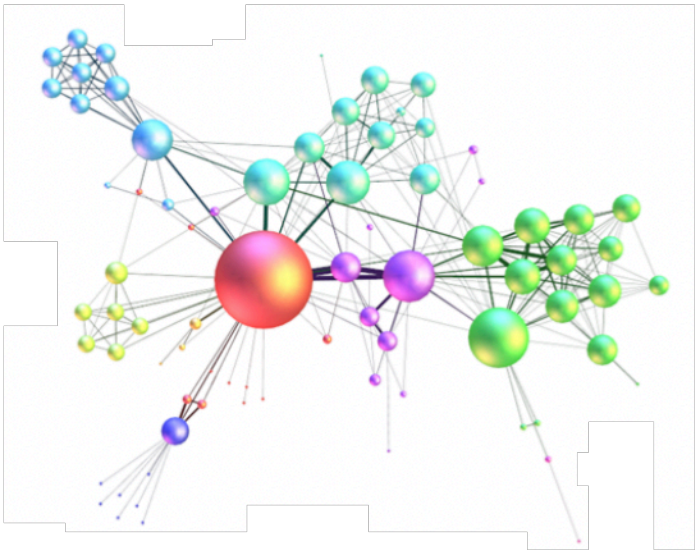
https://emlyon.github.io/mk99/generated-html/network-analysis-for-business.html

Some common tasks and application examples

- **Finding patterns**
  - Fraud detection
  - Performance analysis (e.g., https://youtu.be/KBIZoUikfwo)
- **Link prediction**
  - Recommendations
- **Robustness analysis**
  - Logistics & transportation, security



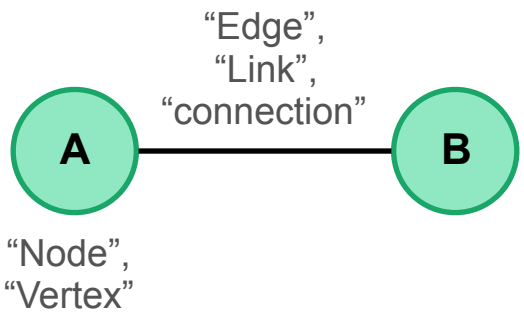
?



https://emlyon.github.io/mk99/generated-html/network-analysis-for-business.html

Basic definitions

- **Graphs** are useful to describe networks
- **G** = (**V**, **E**)
  - *V* = |**V**| - number of nodes
  - *E* = |**E**| - number of edges



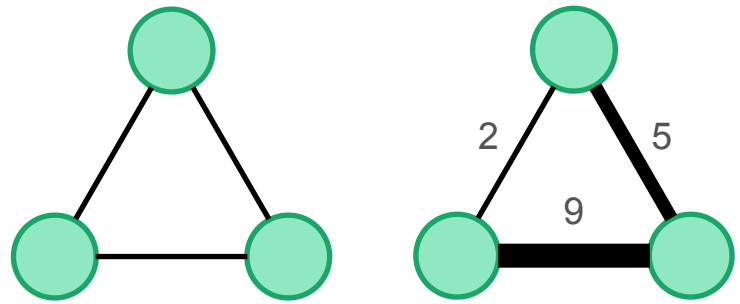
V = 2, E = 1

Basic definitions: Edges

Edge directionality

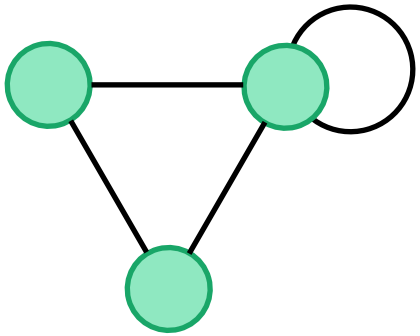
Undirected network Directed network

Connection strength



Unweighted network Weighted network

Self-loops

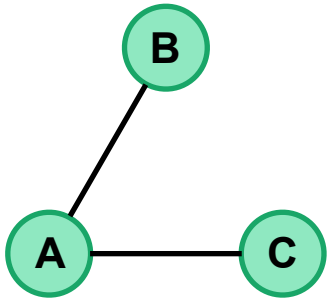


Simple graph

Undirected and no self-loops

Matrix representation (Adjacency matrix)

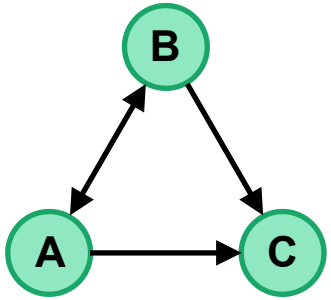
Undirected



	A	B	C
A	0	1	1
B	1	0	0
C	1	0	0

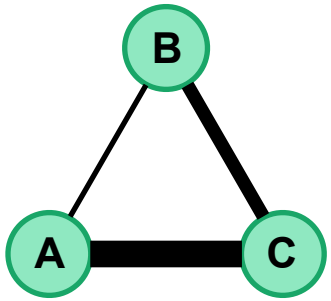
symmetric

Directed



	A	B	C
A	0	1	1
B	1	0	1
C	0	0	0

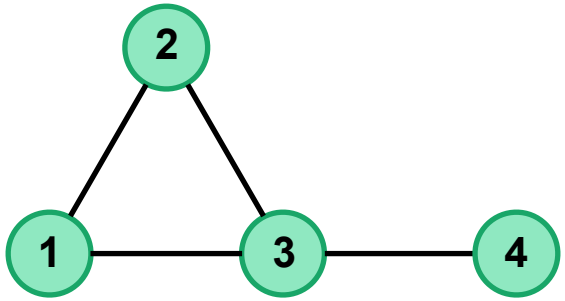
Weighted



	A	B	C
A	-	2	9
B	2	-	5
C	9	5	-

Basic definitions: Connectivity

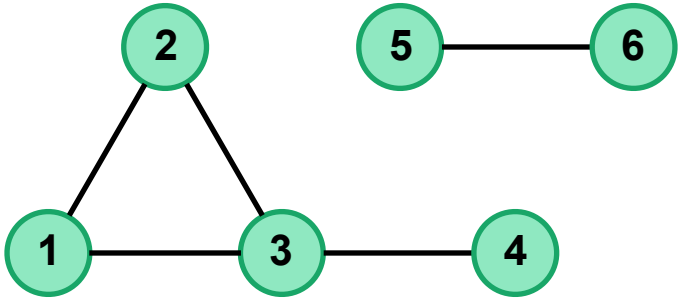
- **Path** between nodes *i* and *j* = sequence of nodes to reach *j* from *i*
- **Geodesic** = shortest path between nodes *i* and *j*
  - Also called **"degree of separation"**
- **Diameter** = maximum geodesic length in the network
- **Average shortest path length** shows how quickly information is transmitted
  - It usually grows logarithmically with the number of nodes ("Small world" phenomenon, "six handshakes")



Paths between 1 and 4: 1-3-4, 1-2-3-4
Shortest path between 1 and 4: 1-3-4
Diameter = 2
Average shortest path = 1 + 1 + 1 + 2 + 2 + 1 / 6 = 4/3

Basic definitions: Connectivity

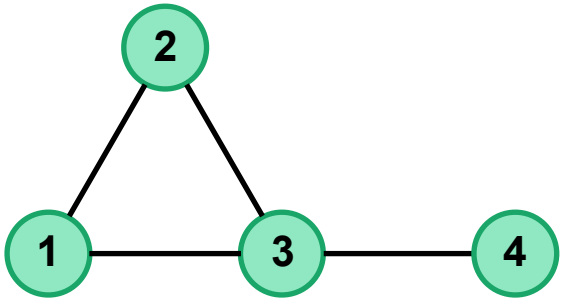
- Node *j* is **reachable** from *i* = there is a path from *i* to *j*



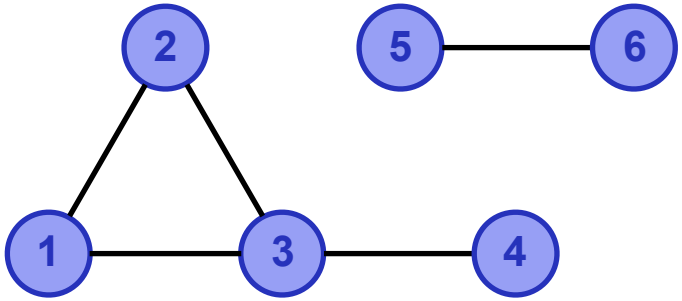
4 is reachable from 1
5 is not reachable from 1

Basic definitions: Connectivity

- An *undirected* graph is **connected** if any node is reachable from any other node



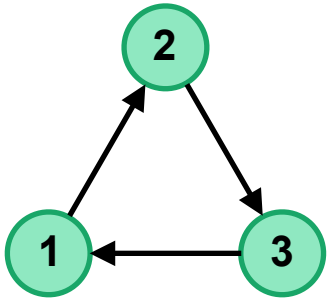
Connected graph



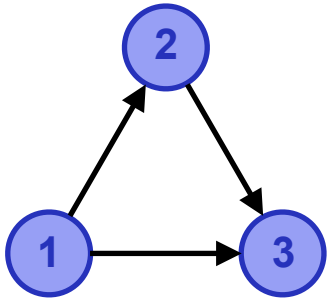
Disconnected graph

Basic definitions: Connectivity

- A *directed* graph is **strongly connected** if any node is reachable from any other node
- A *directed* graph is **weakly connected** if any node is reachable from any other node all edges being converted to undirected



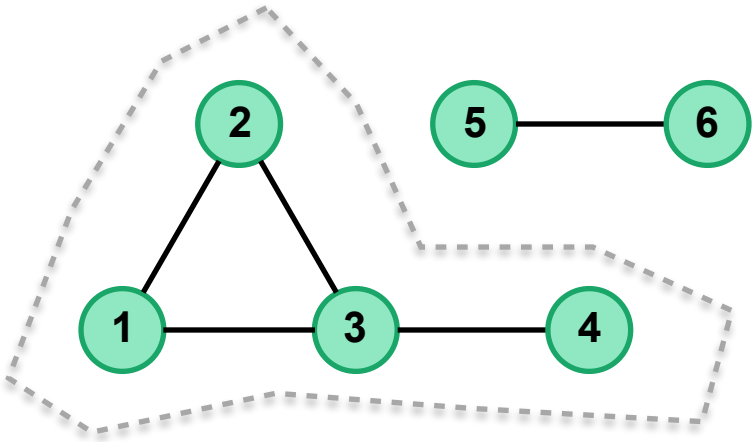
Strongly connected graph



Weakly connected graph

Basic definitions: Connectivity

- A **giant component** of a graph is the largest connected subgraph



Giant component

Basic definitions: Connectivity

- **Complete graph** has all possible edges (every node is connected to all other nodes)
  - Undirected graph: E_{max} = *?*
  - Directed graph: E_{max} = *?*

	A	B	C
A			
B	1		
C	1	1	

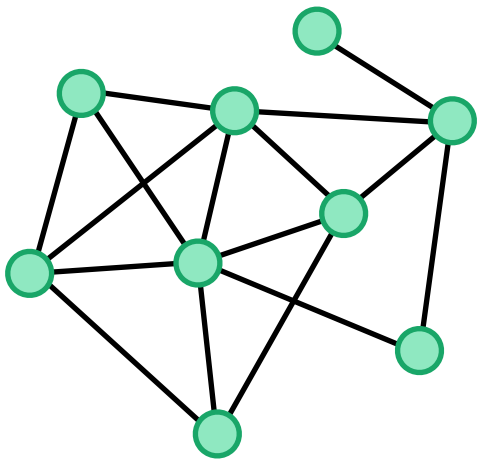
	A	B	C
A		1	1
B	1		1
C	1	1	

Basic definitions: Connectivity

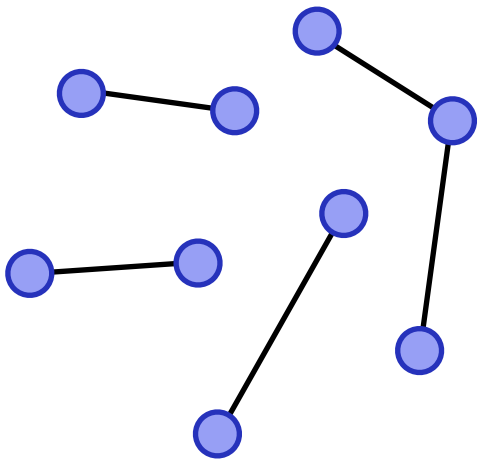
- **Complete graph** has all possible edges (every node is connected to all other nodes)
  - Undirected graph: $E_{max} = V * (V - 1) / 2$
  - Directed graph: $E_{max} = V * (V - 1)$

Basic definitions: Connectivity

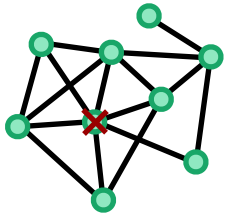
- **Density** is the proportion of edges w.r.t. the maximum number of edges possible
  - Unweighted graph: $D = E / E_{max}$
  - Weighted graph: D = average edge weight



High density (D=0.44)



Low density (D=0.14)



Basic definitions: Connectivity

- **Robustness** of a network measures sensitivity of its connectivity properties to nodes / edges removal (e.g., “faults”, “breaks”, “attacks”)
- Robustness is task-specific and can be defined in many ways, for example,
  - Minimum number of nodes / edges to remove so the graph becomes disconnected
  - Size of the giant component after nodes / edges removal

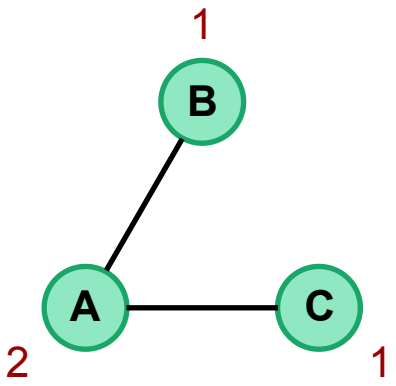
- *“The study of robustness in complex networks is important for many fields. In ecology, robustness is an important attribute of ecosystems, and can give insight into the reaction to disturbances such as the extinction of species. For biologists, network robustness can help the study of diseases and mutations, and how to recover from some mutations. In economics, network robustness principles can help understanding of the stability and risks of banking systems. And in engineering, network robustness can help to evaluate the resilience of infrastructure networks such as the Internet or power grids.”*

https://en.wikipedia.org/wiki/Robustness_of_complex_networks

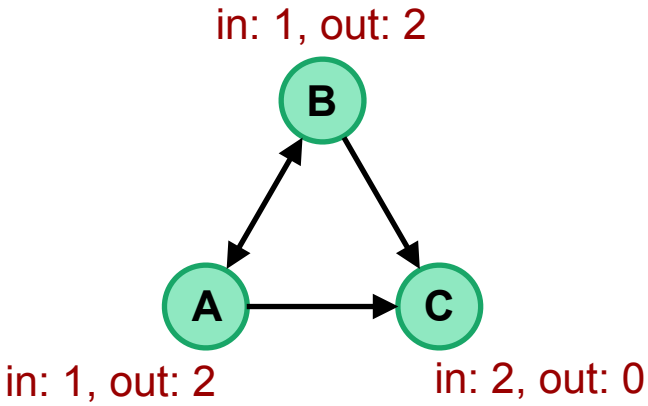
Basic definitions: Node degree

Node degree = number of other nodes connected to the node

Undirected



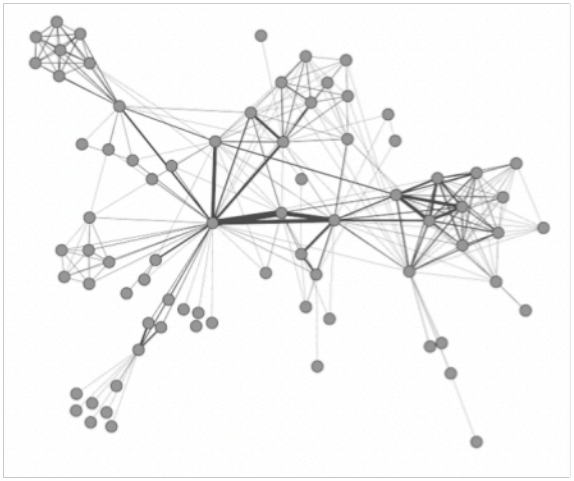
Directed



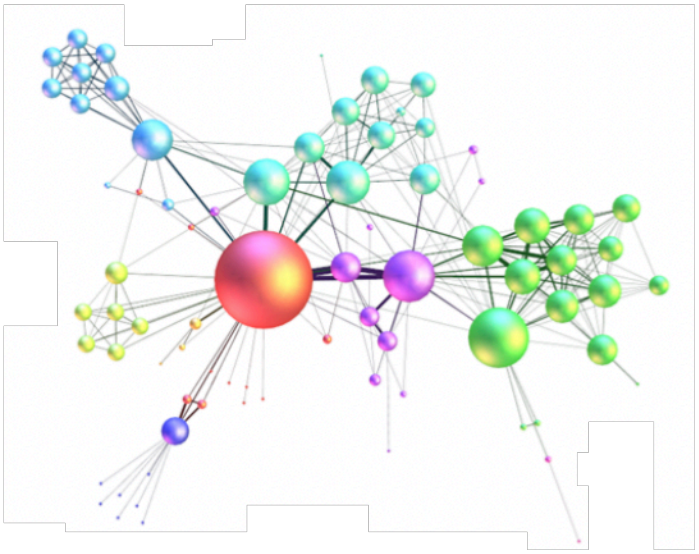
degree = in-degree + out-degree

Node centrality measures

- Which nodes are most “important”?

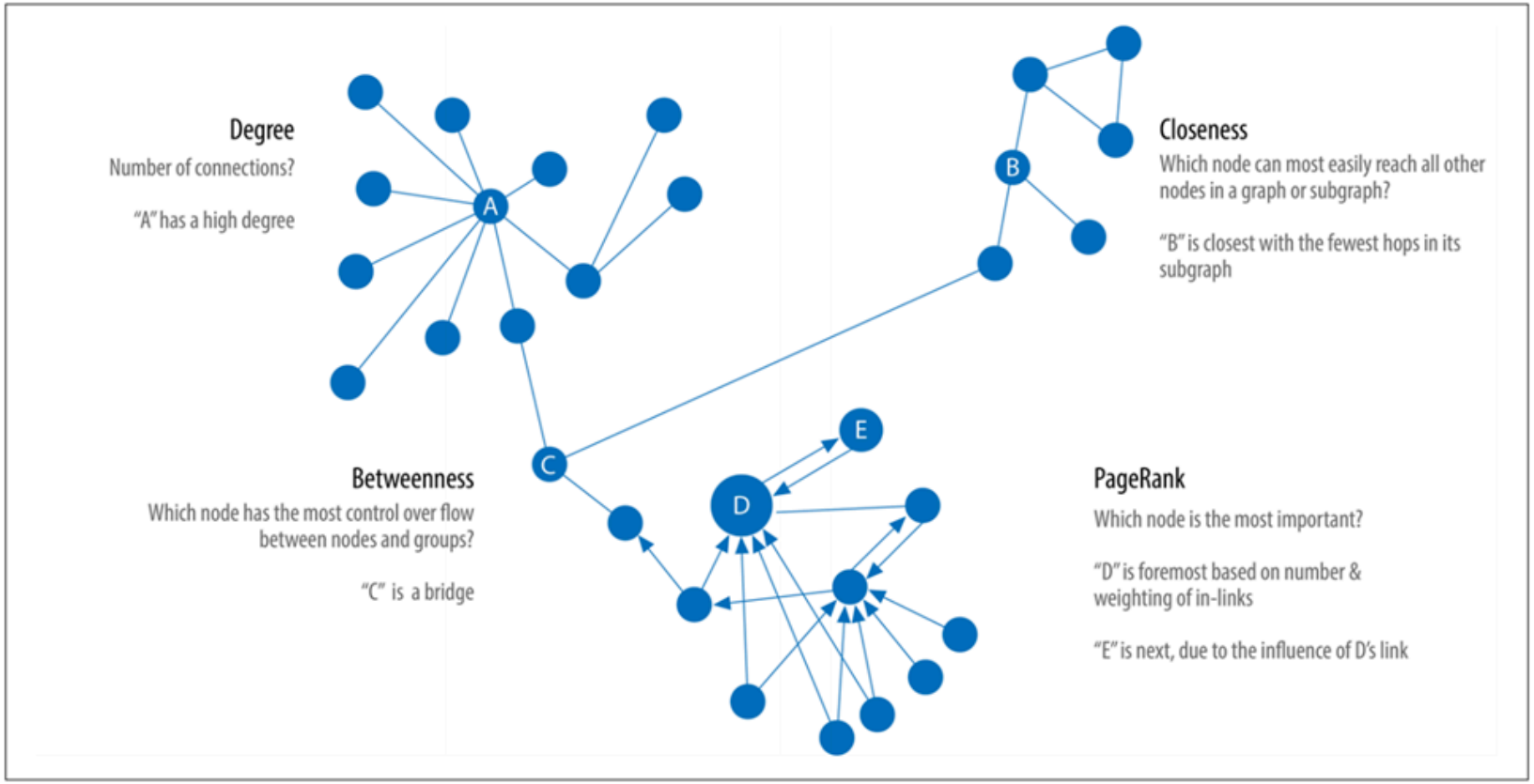


?



https://emlyon.github.io/mk99/generated-html/network-analysis-for-business.html

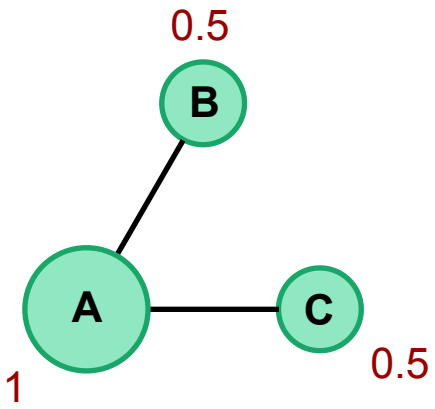
Node centrality measures



https://neo4j.com/graph-algorithms-book/

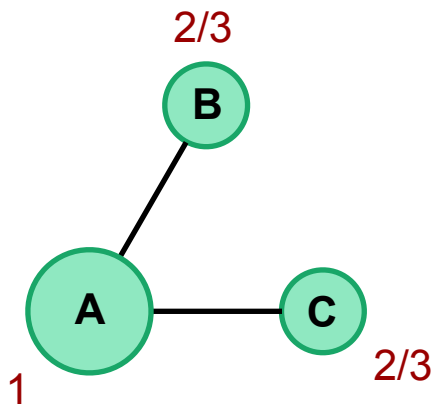
Degree centrality

- *"Who knows the most people?"*
- **C_{DC}(node) = node degree**
  (often divided by theoretical max (in-/out-/)degree possible: *V*-1)
  - Undirected: Number of "friends"
  - Directed: Number of "followers" (in-degree) / "following" (out-degree)
    - In-degree: Popularity status, prestige
    - Out-degree: Expansiveness, gregariousness



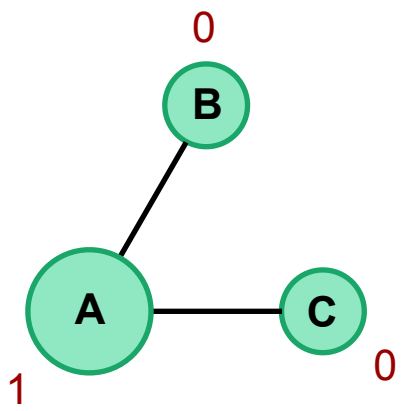
Closeness centrality

- *"Who has the shortest distance to other people?"*
- **C_{CC}(node) = 1 / [sum of shortest paths from the node to all other nodes]**
  (often multiplied by theoretical max degree possible)
- Problem: Defined only for connected graphs



Betweenness centrality

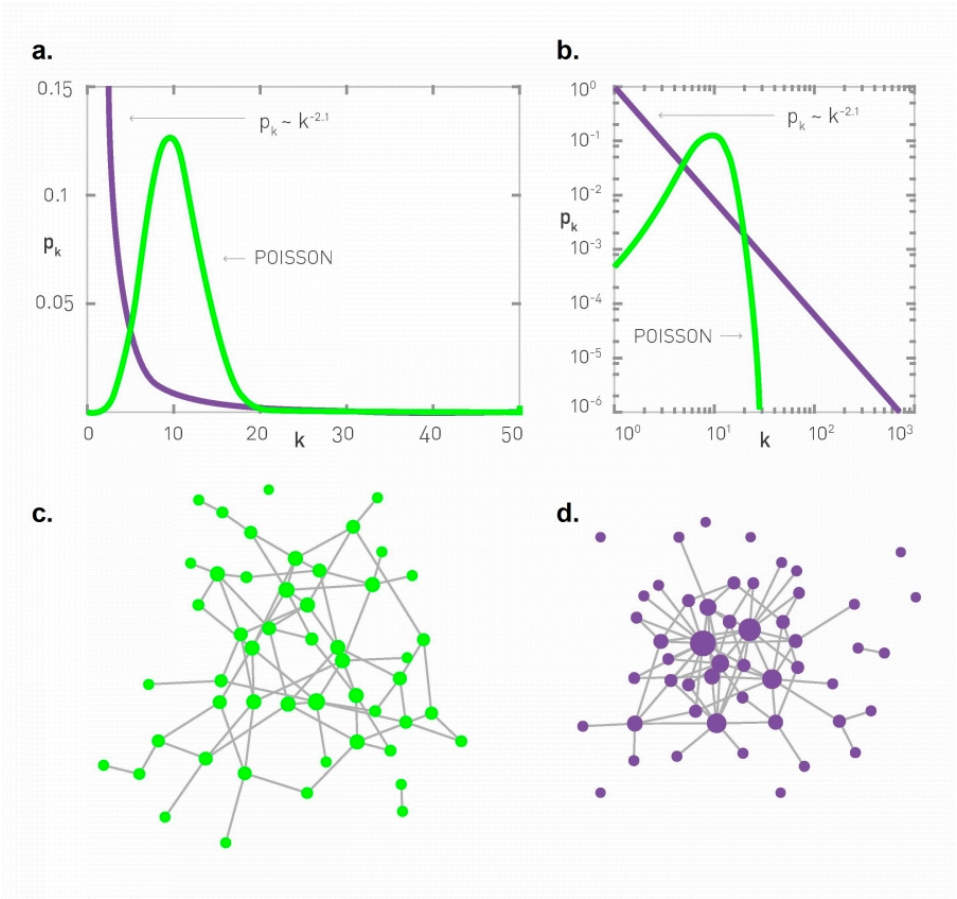
- *"Who is in the middle of information flows?"*
- **C_{BC}(node) = sum over all node pairs [# of shortest paths between a pair that go through the node / # of shortest paths between the pair]** (often divided by the total number of pairs of vertices - 1: (V-1)(V-2)/2 (undirected) or (V-1)(V-2) (directed))
- Problems:
  - Defined only for connected graphs
  - Computationally demanding



Other centrality measures

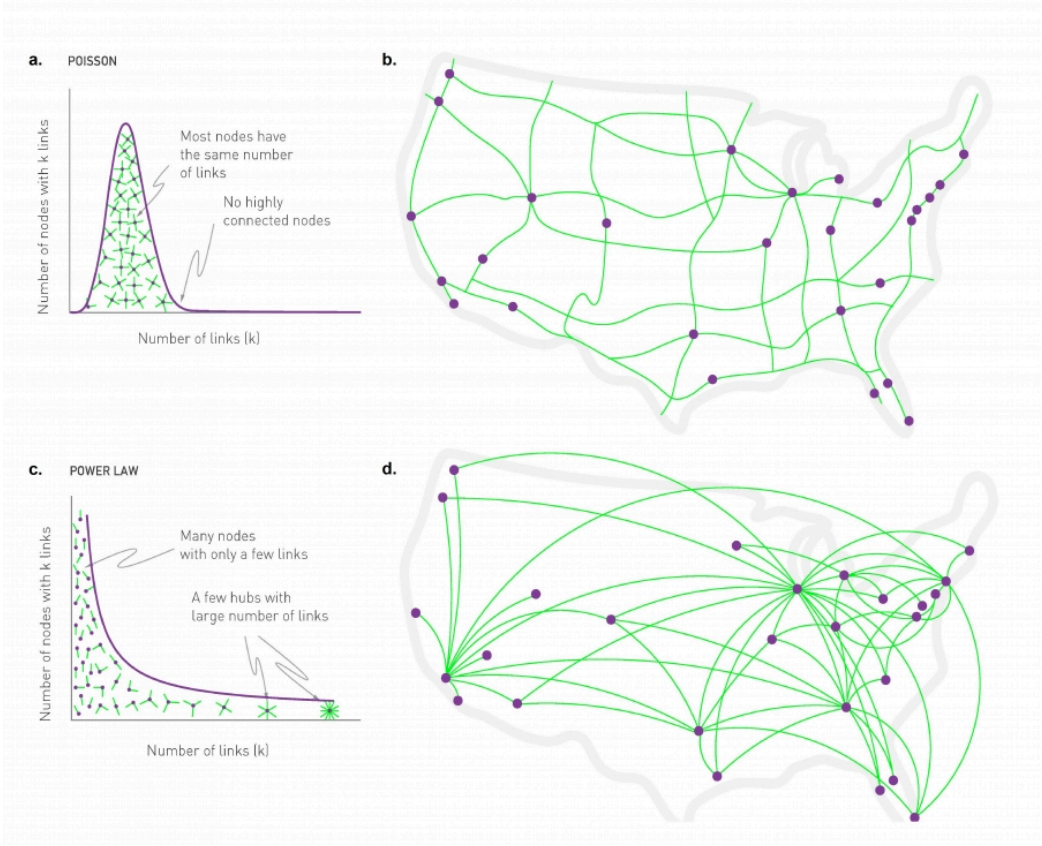
- Eigenvector centrality
  - PageRank
- Information centrality
- Proximity prestige
- Contribution index
- …

Scale-free vs random networks



http://networksciencebook.com/chapter/4

Scale-free vs random networks

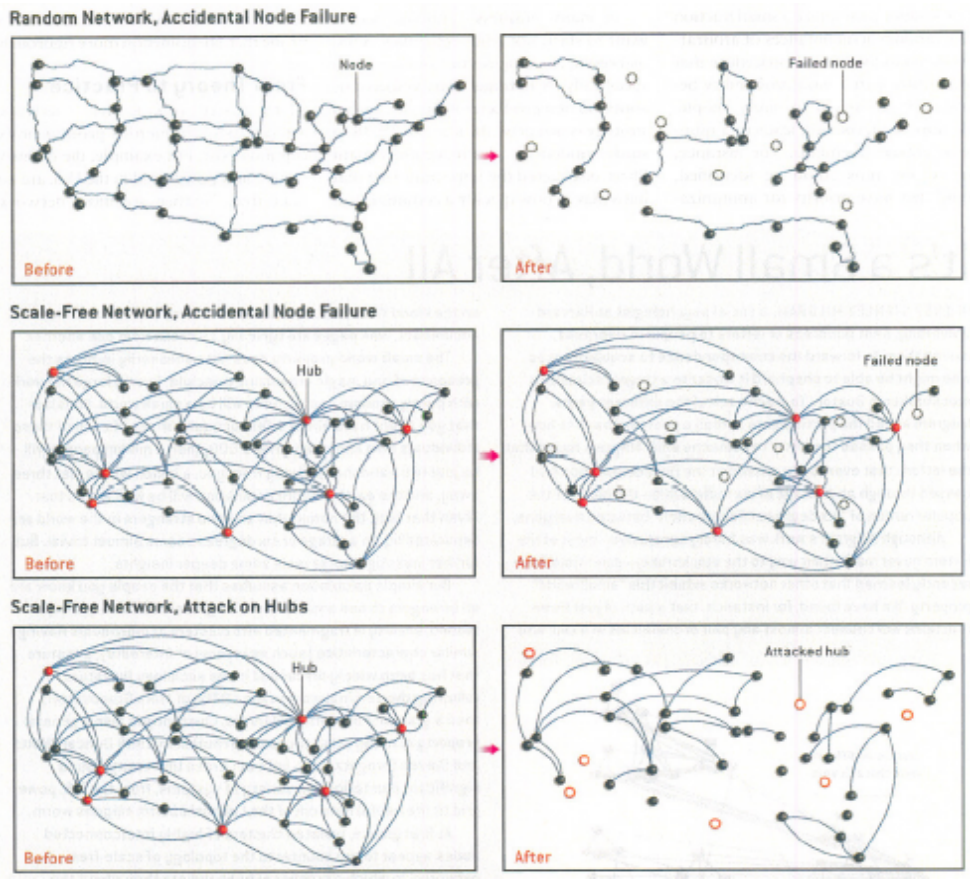


A **random network** looks a bit like the US national highway network in which nodes are cities and links are the major highways. There are no cities with hundreds of highways and no city is disconnected from the highway system.

A **scale-free network** looks like the US air-traffic network, whose nodes are airports and links are the direct flights between them. Most airports are tiny, with only a few flights. Yet, we have a few very large airports, like Chicago or Los Angeles, that act as major hubs, connecting many smaller airports.

http://networksciencebook.com/chapter/4

Scale-free vs random networks: Robustness

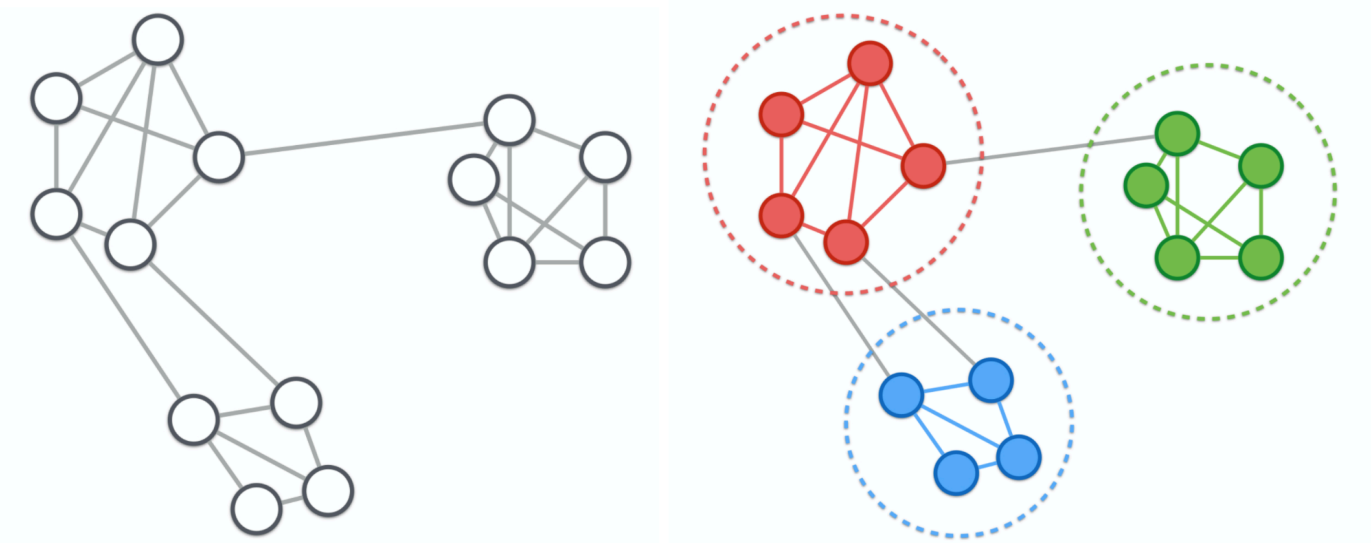


- The accidental failure of a number of nodes in a random network can fracture the system into non-communicating islands.
- Scale-free networks are more robust in the face of such failures.
- Scale-free networks are highly vulnerable to a coordinated attack against their hubs.

http://didawiki.cli.di.unipi.it/lib/exe/fetch.php/dm/han.ppt

Community detection (Segmentation)

- **Community detection** aims to partition a graph into clusters of densely connected nodes, with the nodes belonging to different communities being only sparsely connected.

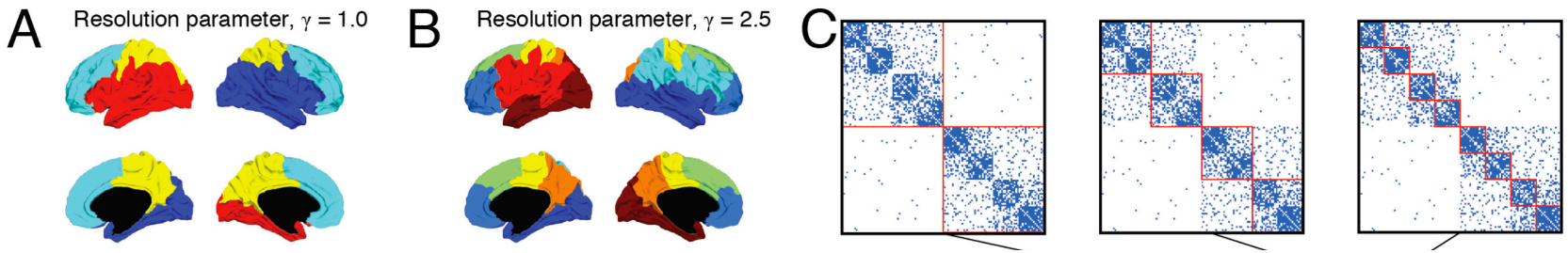


https://blog.insightdatascience.com/graph-based-machine-learning-6e2bd8926a0

Community detection

- **Examples:**
  - Social groups
  - Functional brain regions

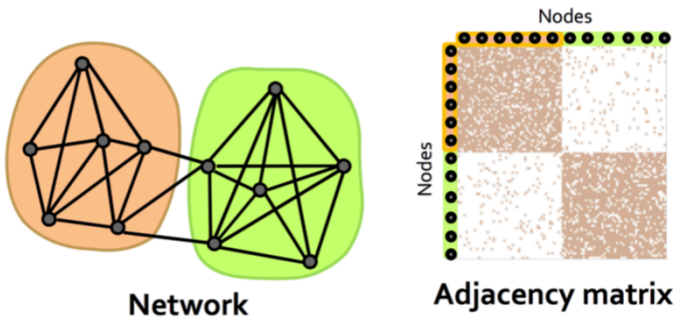
- **Applications:**
  - Understanding a system's structure, segmentation
  - Visualising and simplifying huge networks
  - Recommendations
  - …



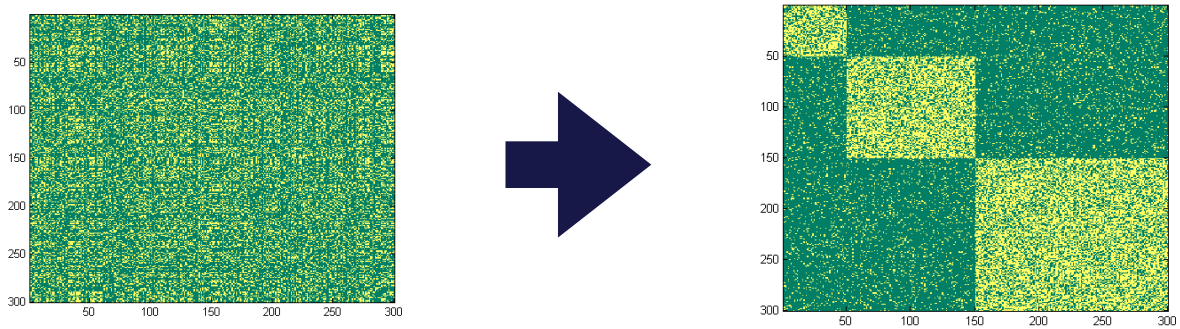
https://www.sciencedirect.com/science/article/pii/S1053811916306152

Community detection

- Communities correspond to dense blocks in the adjacency matrix



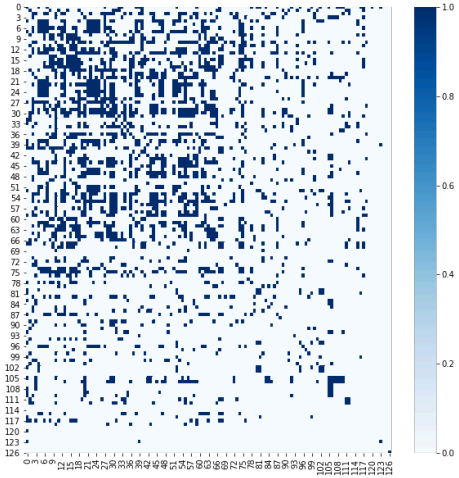
- Re-ordering

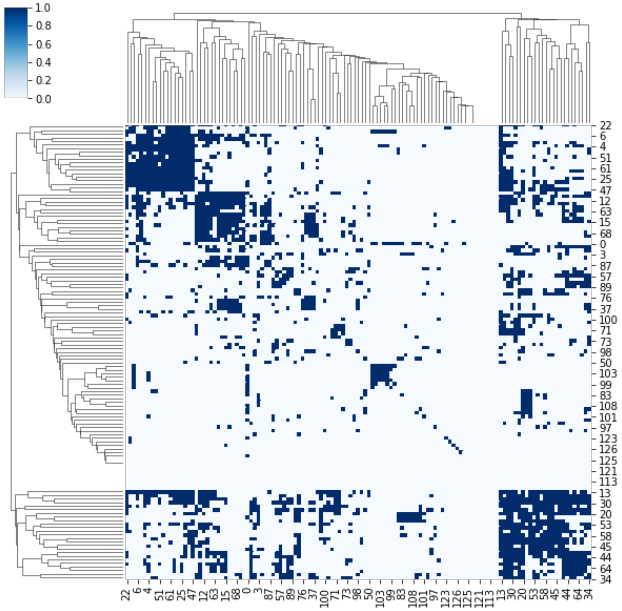


Jure Leskovec, Anand Rajaraman, Jeff Ullman. Mining of Massive Datasets
Huan Liu, Lei Tang and Nitin Agarwal (SocialCom'09), 2009.

Community detection algorithms

- **Problem is NP-hard**
- **Many methods exist (no “silver bullet” though):**
  - Hierarchical clustering
  - Spectral clustering
  - Girvan-Newman Method
  - Modularity
  - …



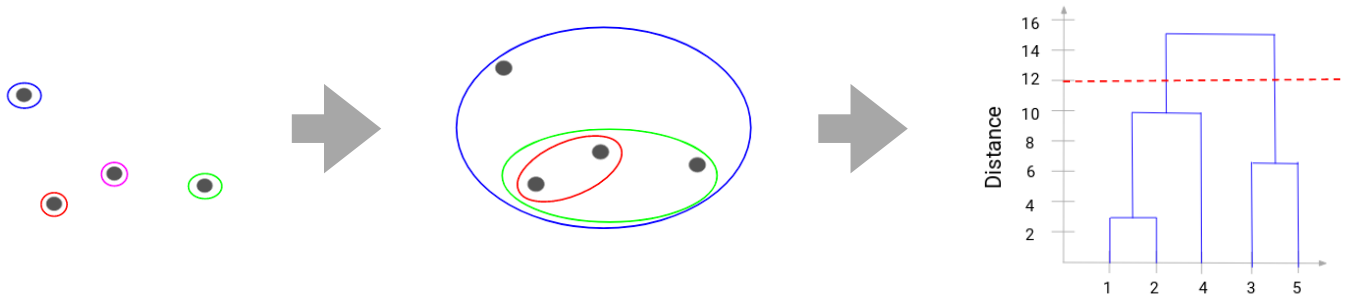


Agglomerative hierarchical clustering (variation)

1. Initialise each node as a community (i.e., cluster)
2. Calculate distance matrix between the communities
3. Repeat until one community left
    1. Merge to closest communities
    2. Recalculate distance matrix between the communities
4. Select number of clusters using some criteria (e.g., max modularity)

Example of the distance measures between two

- **nodes:** Euclidean distance between the rows in the adjacency matrix
- **communities:** Average distance between all the nodes in the communities



https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/
http://ccicada.org/wp-content/uploads/2017/06/Community-Detection-with-Hierarchical-Clustering-Algorithms-Feb-3-2017.pdf

Managing complexity: Design structure matrix

- **Design Structure Matrix** (DSM) techniques support the management of complexity by focusing attention on the elements of a complex system and how they relate to each other. DSM−based techniques have proven to be very valuable in understanding, designing, and optimizing complex system architectures such as those of products, organizations, and processes.

- **dsmweb.org**

DSM examples

- Component-Based DSM

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Radiator	A		X														
Engine Fan	B	X				X											
Heater Core	C																X
Heater Hoses	D																
Condenser	E						X		X								
Compressor	F					X			X	X							
Evaporator Case	G																X
Evaporator Core	H					X	X			X							X
Accumulator	I						X		X								
Refrigeration Controls	J																
Air Controls	K																
Sensors	L																
Command Distribution	M																
Actuators	N																
Blower Controls	O																X
Blower Monitor	P			X				X	X							X	

https://dsmweb.org/different-dsm-types/

DSM examples

- Component-Based DSM

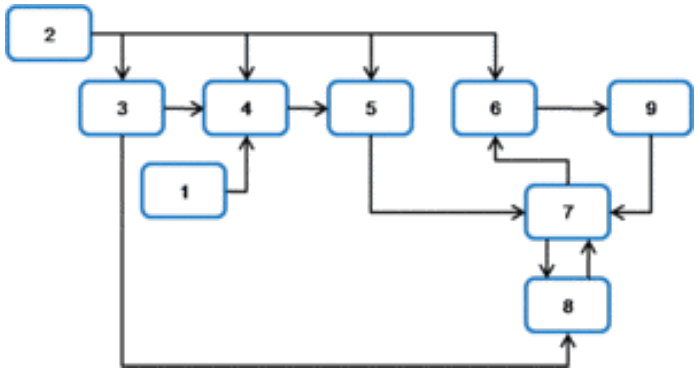
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Radiator	A		X														
Engine Fan	B	X				X											
Heater Core	C																X
Heater Hoses	D																
Condenser	E						X		X								
Compressor	F					X			X	X							
Evaporator Case	G																X
Evaporator Core	H					X	X			X							X
Accumulator	I						X		X								
Refrigeration Controls	J																
Air Controls	K																
Sensors	L																
Command Distribution	M																
Actuators	N																
Blower Controls	O																X
Blower Monitor	P			X				X	X							X	

		D	J	K	L	M	N	A	B	E	F	I	H	C	P	O	G
Heater Hoses	D																
Refrigeration Controls	J																
Air Controls	K																
Sensors	L																
Command Distribution	M																
Actuators	N																
Radiator	A								X								
Engine Fan	B							X		X							
Condenser	E								X		X		X				
Compressor	F									X		X	X				
Accumulator	I										X		X				
Evaporator Core	H									X	X	X			X		
Heater Core	C														X		
Blower Monitor	P												X	X		X	X
Blower Controls	O														X		
Evaporator Case	G														X		

https://dsmweb.org/different-dsm-types/

DSM examples

- Activity-Based (Task-Based) DSM



	1	2	3	4	5	6	7	8	9
1				X					
2			X	X	X	X			
3				X				X	
4					X				
5							X		
6									X
7						X		X	
8							X		
9							X		

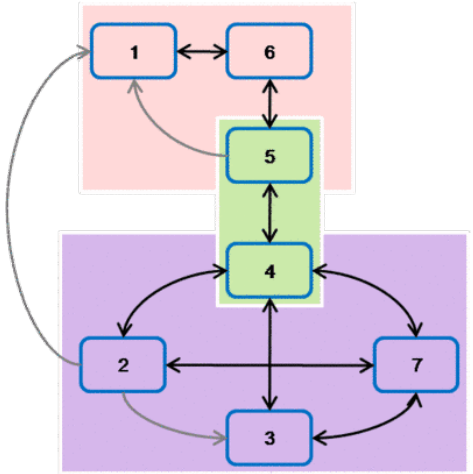
https://dsmweb.org/different-dsm-types/

DSM examples

- Activity-Based (Task-Based) DSM

	1	2	3	4	5	6	7
1						X	
2	X		X	X			X
3				X			X
4		X	X		X		X
5	X			X		X	
6	X				X		
7		X	X	X			

	1	6	5	4	2	3	7
1		X					
6	X		X				
5	X	X		X			
4			X		X	X	X
2	X			X		X	X
3				X			X
7				X	X	X	



https://dsmweb.org/clustering-a-dsm/

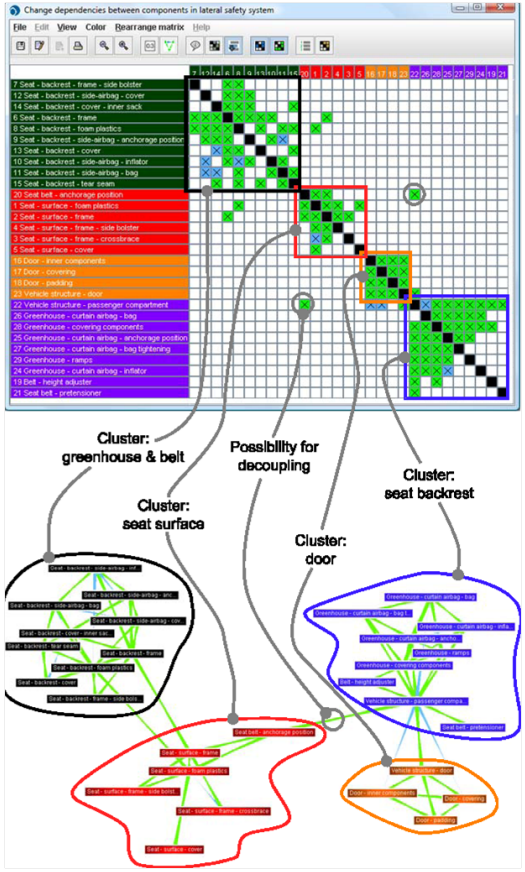
Industrial use of DSM

- **https://dsmweb.org/industrial-use-of-dsm/**
  - Market structure analysis
  - Structuring automotive safety development
  - Setting up teams for integrated automotive development

https://dsmweb.org/industrial-use-of-dsm/

DSM Application example

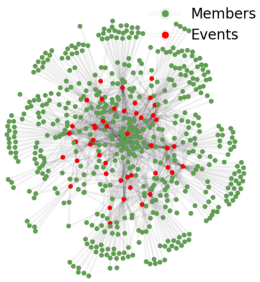
- Structuring automotive safety development



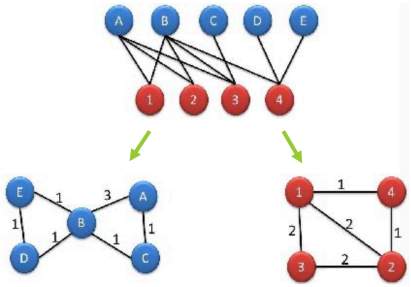
Heterogeneous networks

- **Two kinds of networks:**
  - Homogeneous vs. heterogeneous
- **Homogeneous networks**
  - Single object type and single link type
  - Single model social networks (e.g., friends)
  - WWW: a collection of linked Web pages
- **Heterogeneous networks**
  - Multiple object and link types
  - Medical network: patients, doctors, disease, contacts, treatments
  - Bibliographic network: publications, authors, venues

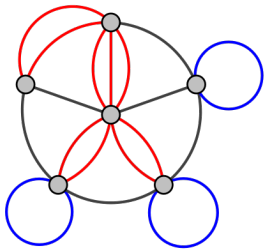
Heterogeneous network



Bipartite graph

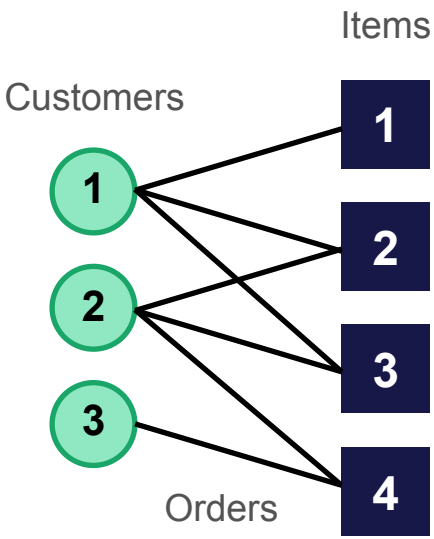


Multigraph

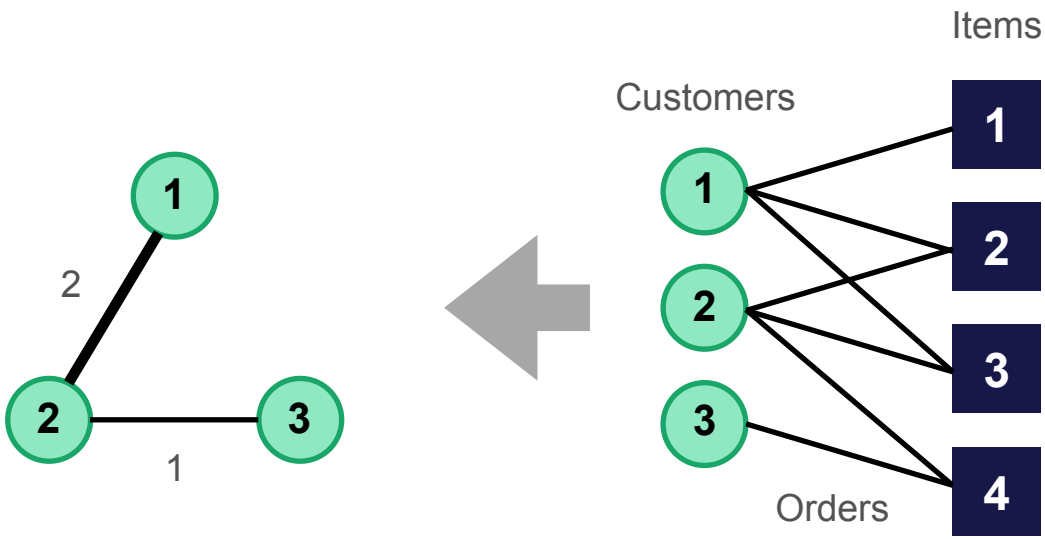


https://github.com/stkbailey/nashnetx/
https://en.wikipedia.org/wiki/Multigraph

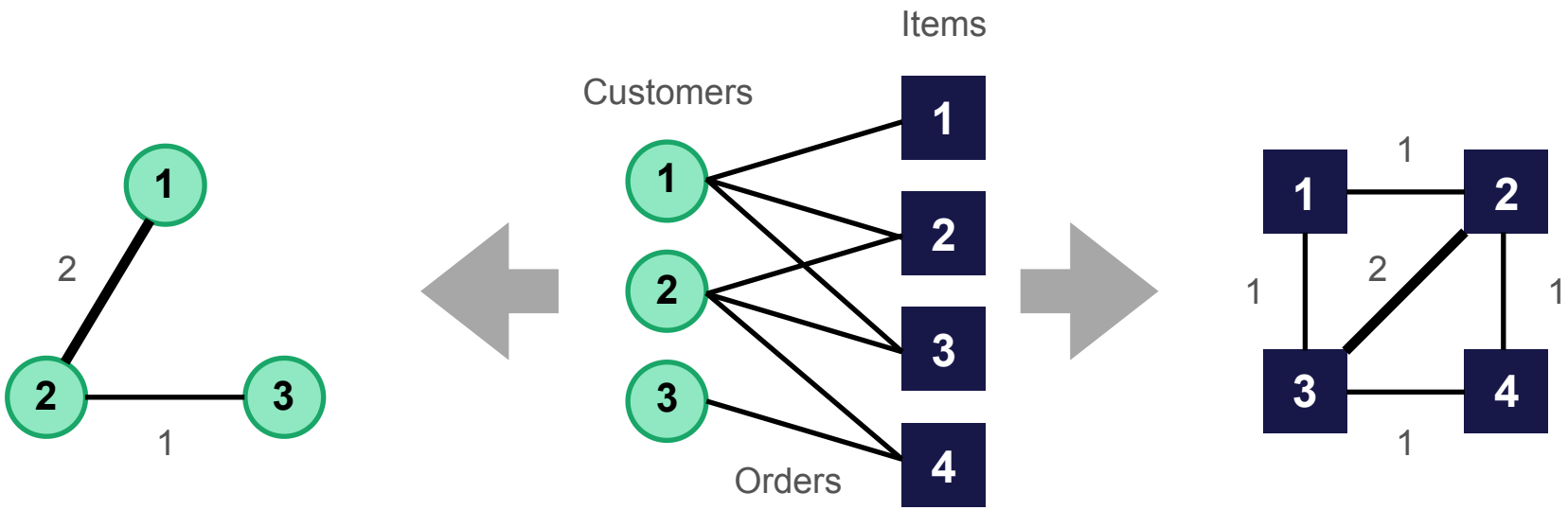
Bipartite networks



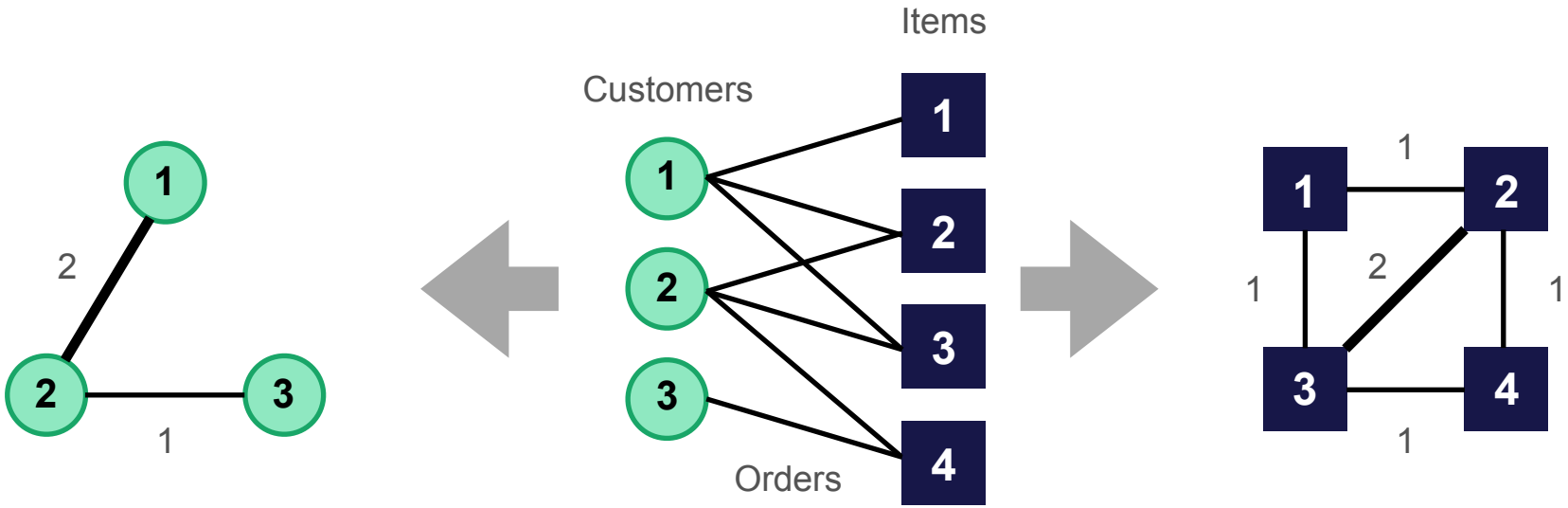
Bipartite networks



Bipartite networks



Bipartite networks



- Customer-Customer (User-User) similarity metrics
- Centrality analysis
- …

Combined analysis

- Item-Item similarity metrics
- Centrality analysis
- …

Bipartite networks applications

- **Recommender systems**
  - "Bipartite network projection and personal recommendation", Phys Rev. E 76, 046115 2007 *[optional reading]*
  - "Network-based recommendation algorithms: A review", Physica A 452 (2016) 192–208 *[optional reading]*

- **Web advertising** (e.g., Google AdWords algorithm)
  - "Mining of Massive Datasets", Chapter 8 (http://www.mmds.org/) *[optional reading]*

- **Dating apps**

- …

Exercise

- “Network Analysis.ipynb”

Homework

- Working with OpenStreetMap data:
  Basic Network Analysis Tutorial
  https://www.kaggle.com/usui113yst/basic-network-analysis-tutorial