Network Science

Part 1

Network Analysis

Márton Karsai
Central European University

About me…

Márton Karsai

- Associate Professor
- Department of Network and Data Science
- Central European University
- email: karsaim@ceu.edu
- web: www.martonkarsai.com

Research:

- Network Science - temporal networks
- Computational Social Science - socioeconomic networks, social contagion phenomena
- Human dynamics - bursty human dynamics, human mobility
- Computational Epidemiology - temporal and spatial network epidemiology
- Data science - large scale data collection methods of human behaviour

Outline

- **Part 1: The network concept / metrics / properties**
- **Part 2: Network models**
- **Part 3: Application of network science**

Course Targets

The course will teach you

- how to interpret a complex system like a complex network
- to know how to analyse and characterise networks from data
- to understand the fundamental network models
- to be aware of state-of-the-art examples of real networks and their applications
- and much more…

Course Targets

The course will teach you

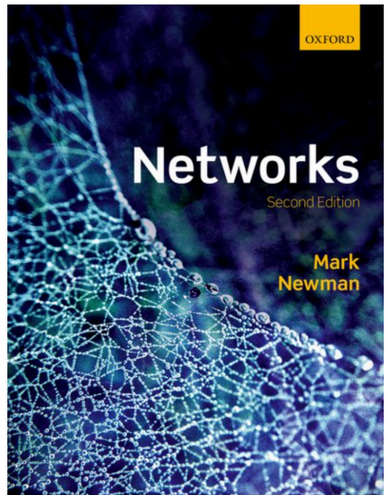
- how to interpret a complex system like a complex network
- to know how to analyse and characterise networks from data
- to understand the fundamental network models
- to be aware of state-of-the-art examples of real networks and their applications
- and much more…

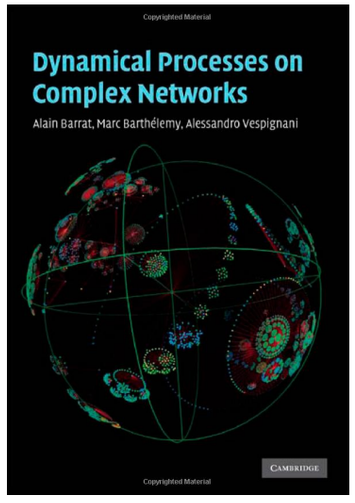
The course will NOT teach you

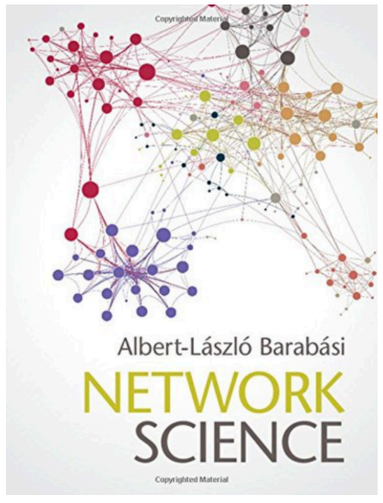
- about the implementation of methods and algorithms
- how to develop program codes
- how to visualise network data

Materials

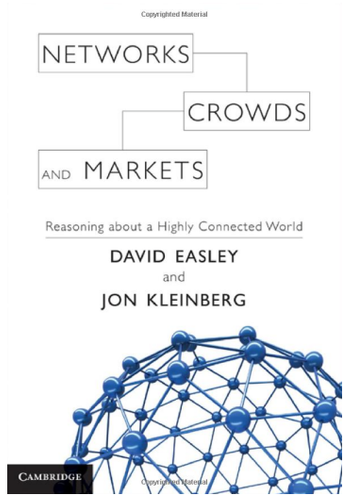
Lecture books





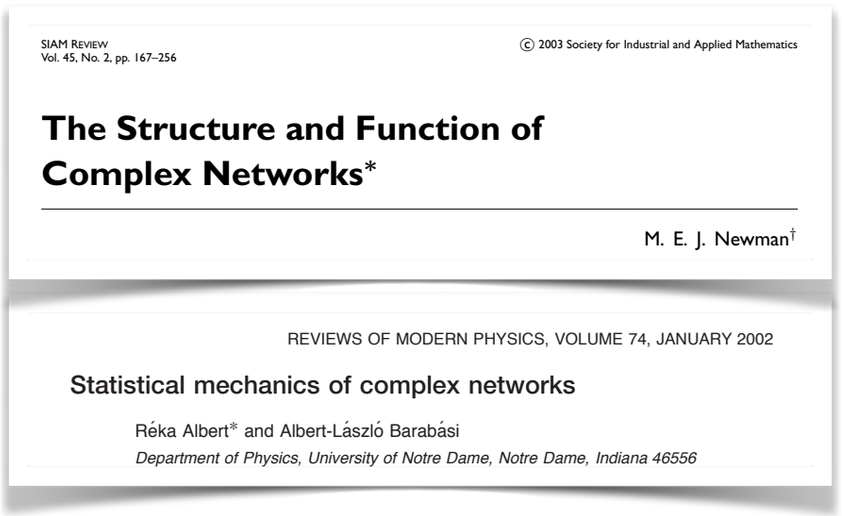


available free online



available free online

Reviews


SIAM REVIEW
Vol. 45, No. 2, pp. 167–256

© 2003 Society for Industrial and Applied Mathematics

The Structure and Function of Complex Networks*

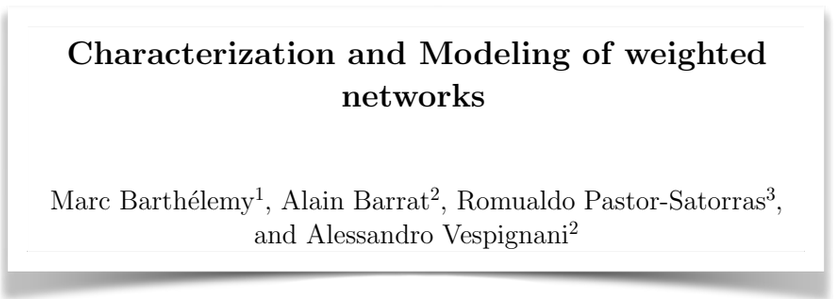
M. E. J. Newman†

REVIEWS OF MODERN PHYSICS, VOLUME 74, JANUARY 2002

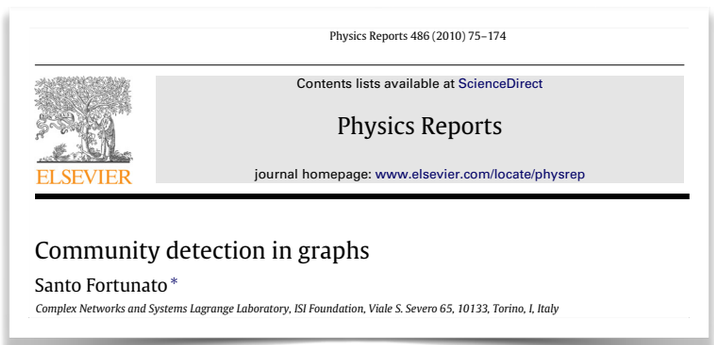
Statistical mechanics of complex networks

Réka Albert* and Albert-László Barabási

Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556


Characterization and Modeling of weighted networks

Marc Barthélemy[1], Alain Barrat[2], Romualdo Pastor-Satorras[3], and Alessandro Vespignani[2]


Physics Reports 486 (2010) 75–174

Contents lists available at ScienceDirect

Physics Reports

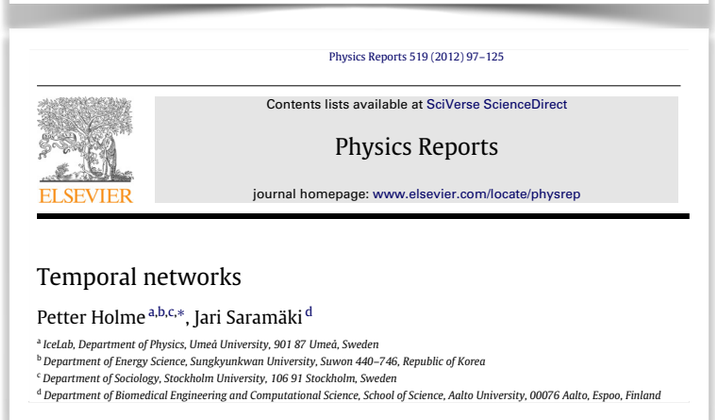
journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

Community detection in graphs

Santo Fortunato *

Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I, Italy


Physics Reports 519 (2012) 97–125

Contents lists available at SciVerse ScienceDirect

Physics Reports

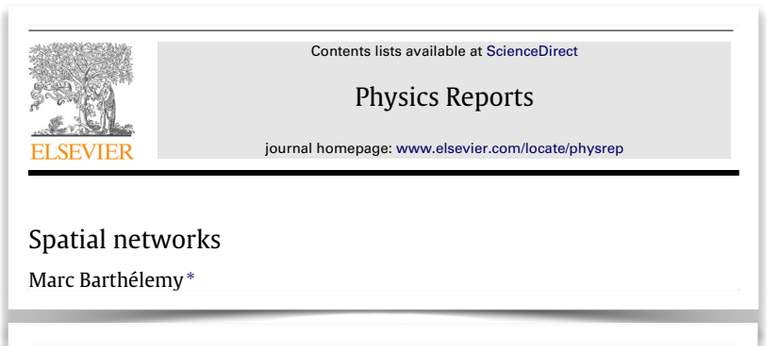
journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

Temporal networks

Petter Holme [a,b,c,*], Jari Saramäki [d]

[a] *IceLab, Department of Physics, Umeå University, 901 87 Umeå, Sweden*
[b] *Department of Energy Science, Sungkyunkwan University, Suwon 440–746, Republic of Korea*
[c] *Department of Sociology, Stockholm University, 106 91 Stockholm, Sweden*
[d] *Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, 00076 Aalto, Espoo, Finland*


Contents lists available at ScienceDirect

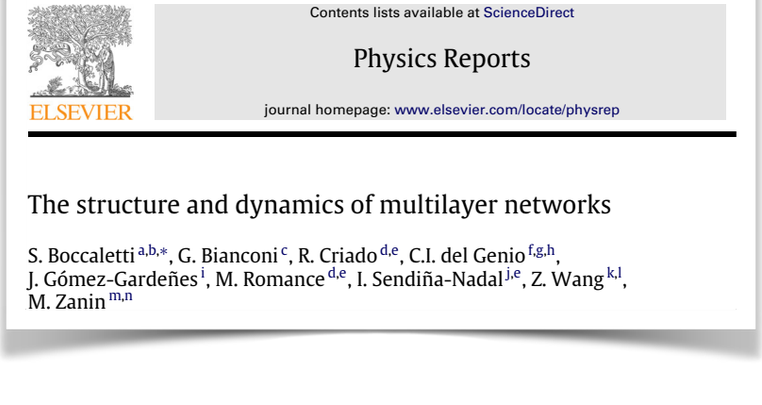
Physics Reports

journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

Spatial networks

Marc Barthélemy *


Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

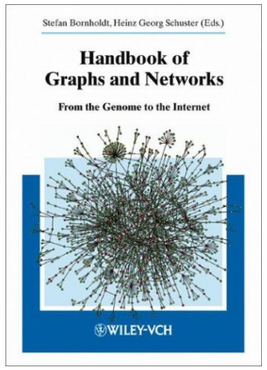
The structure and dynamics of multilayer networks

S. Boccaletti [a,b,*], G. Bianconi [c], R. Criado [d,e], C.I. del Genio [f,g,h], J. Gómez-Gardeñes [i], M. Romance [d,e], I. Sendiña-Nadal [j,e], Z. Wang [k,l], M. Zanin [m,n]

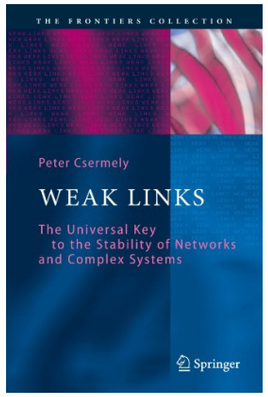
…and many more…all of them on arXiv.org!

Materials

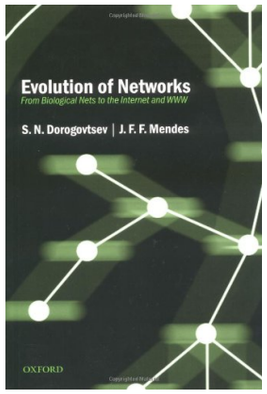
Related books



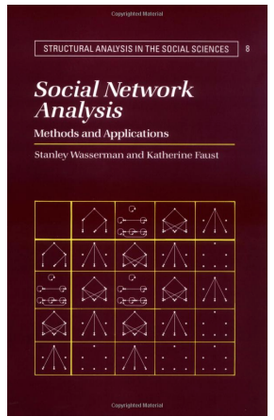
Handbook of Graphs and Networks: From the Genome to the Internet (Wiley-VCH, 2003).



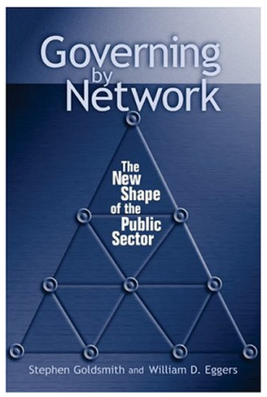
P. Csermely, Weak Links: The Universal Key to the Stability of Networks and Complex Systems (The Frontiers Collection) (Springer, 2006), rst edn.



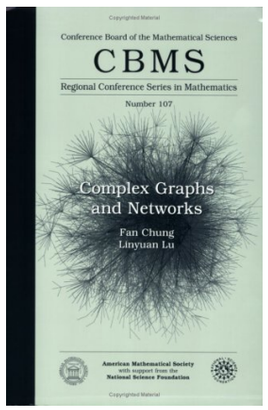
S. N. Dorogovtsev and J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW (Oxford University Press, 2003).



S. Wasserman and K. Faust
Social Network Analysis (Methods and Applications)
Cambridge University Press (1994)



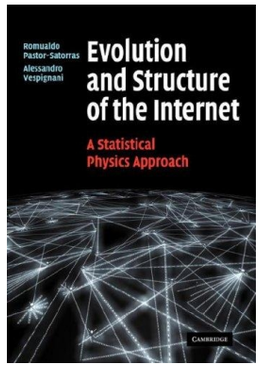
S. Goldsmith, W. D. Eggers, Governing by Network: The New Shape of the Public Sector (Brookings Institution Press, 2004).



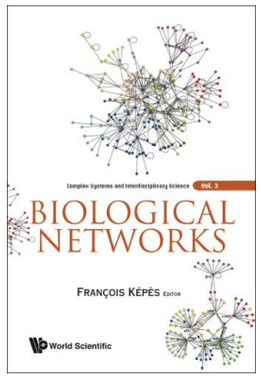
L. L. F. Chung, Complex Graphs and Networks (CBMS Regional Conference Series in Mathematics) (American Mathematical Society, 2006).

Materials

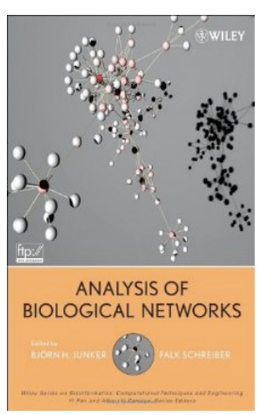
Related books



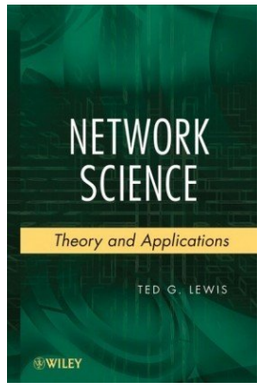
R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet: A Statistical Physics Approach (Cambridge University Press, 2007), rst edn.



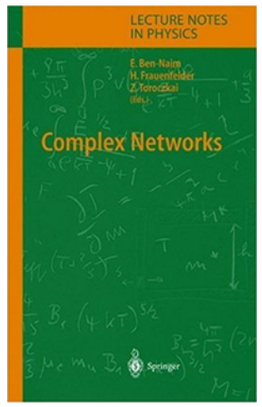
F. Kopos, Biological Networks (Complex Systems and Interdisciplinary Science) (World Scientic Publishing Company, 2007), rst edn.



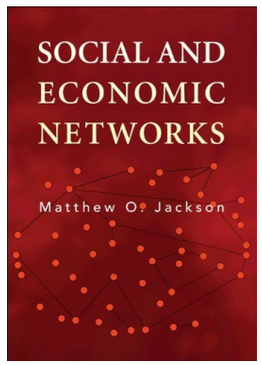
B. H. Junker, F. Schreiber, Analysis of Biological Networks (Wiley Series in Bioinformatics) (Wiley-Interscience, 2008).



T. G. Lewis, Network Science: Theory and Applications (Wiley, 2009).



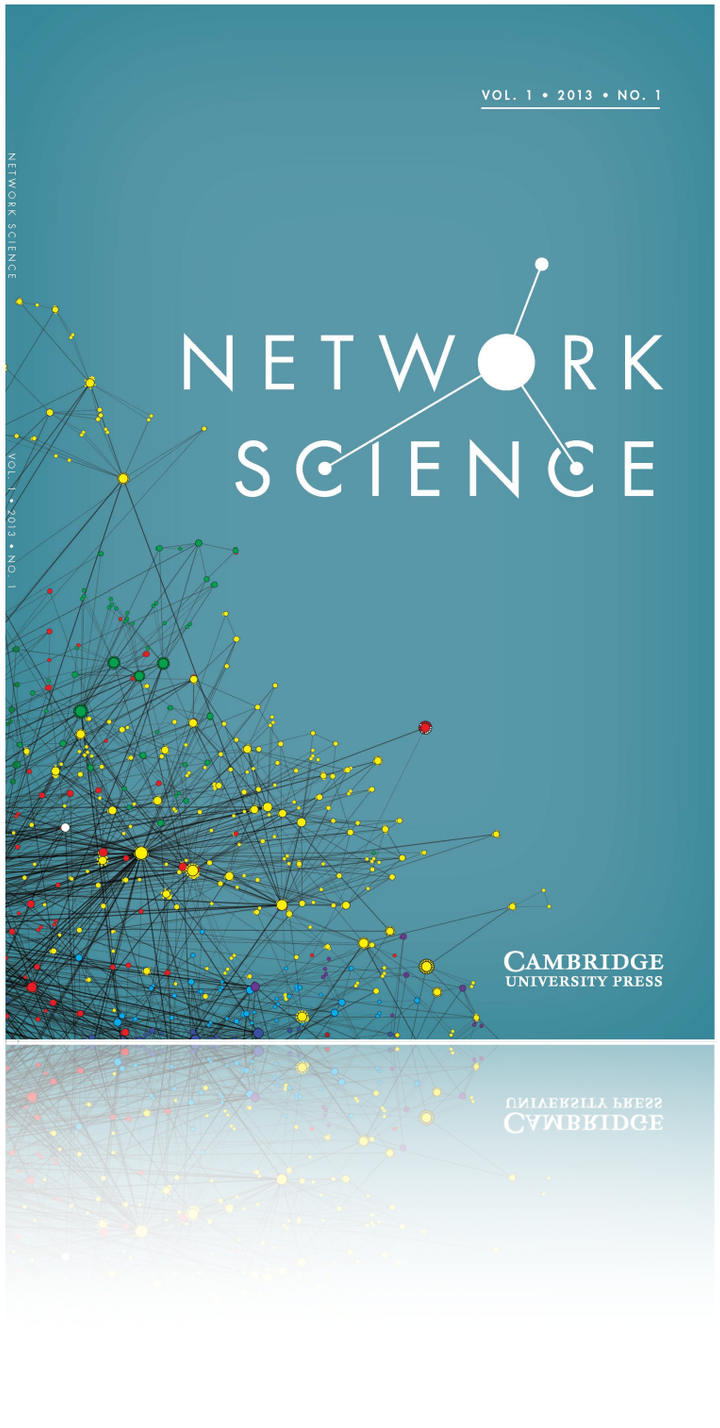
E. Ben Naim, H. Frauenfelder, Z.Torotzai, Complex Networks (Lecture Notes in Physics) (Springer, 2010), rst edn.

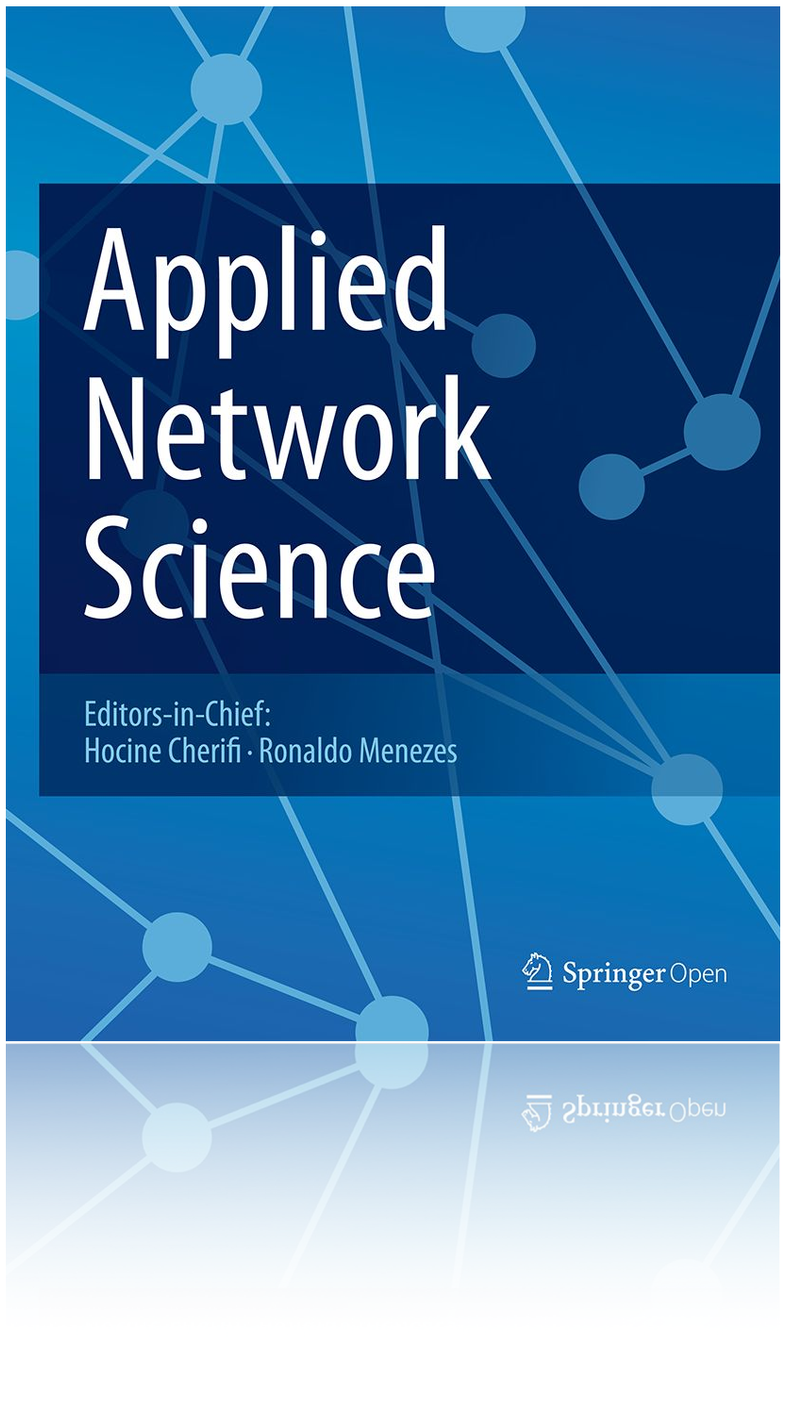


M. O. Jackson, Social and Economic Networks (Princeton University Press, 2010).

Materials

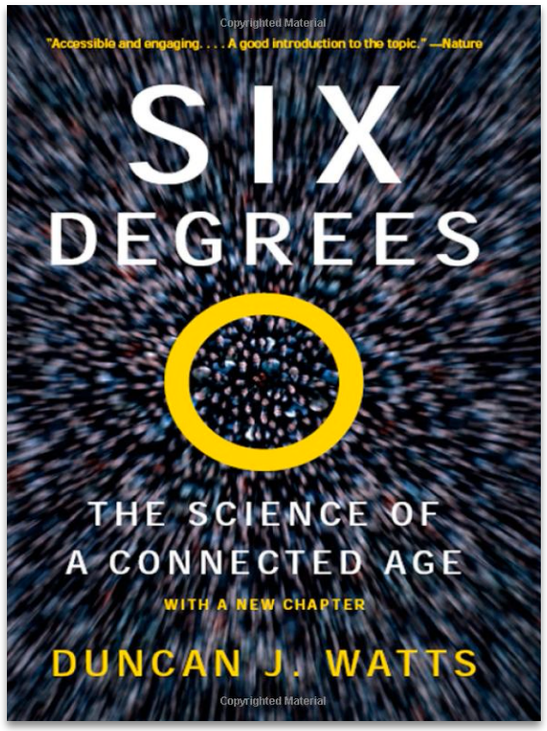
Journals

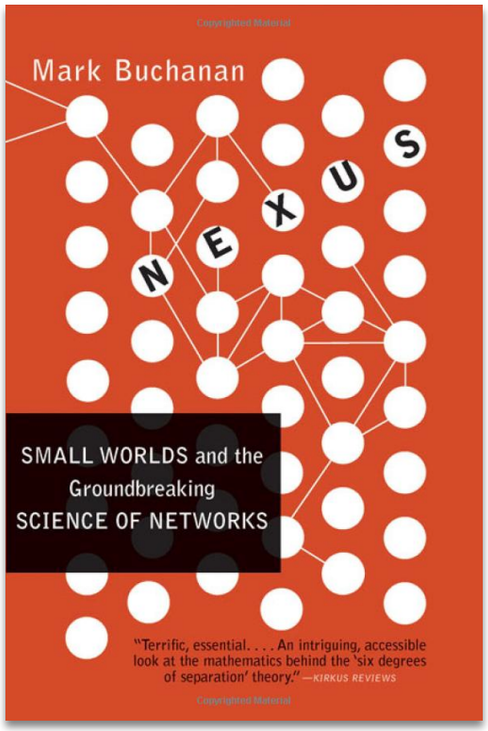


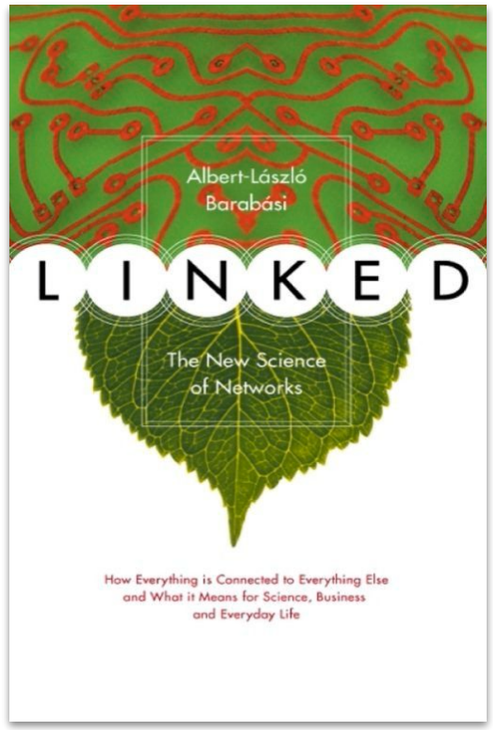


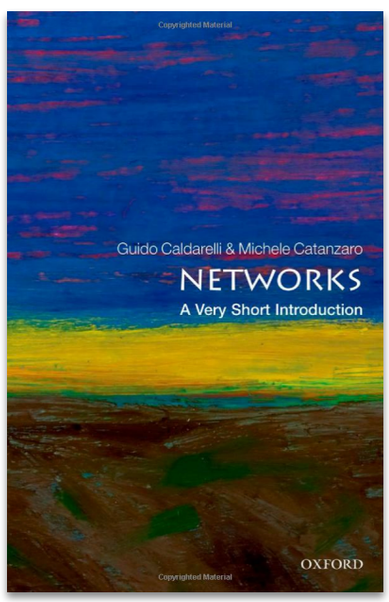
Materials

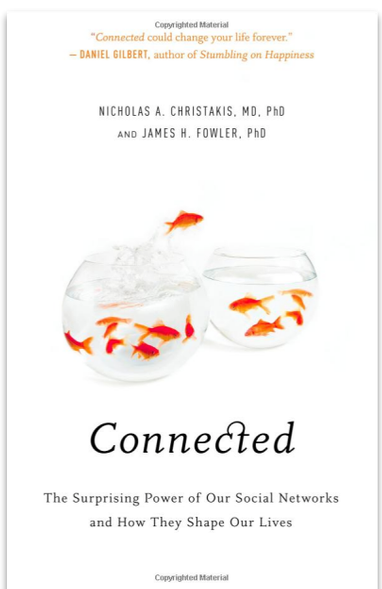
Pop-science books











Part 1 schedule

1. Complexity and complex system

2. The network approach

3. Types of networks

4. Network characteristics

Complexity
and
complex networks

Stephen Hawking, who in an interview at the turn of the millennium was asked the following question:

- Some say that while the 20th century was the century of physics, we are now entering the century of biology. What do you think of this?

To which he responded:

- I think the next century will be the century of complexity.

Excerpt from the book of Geoffrey West “Scale.” Penguin Publishing Group (2017).

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
–adjective
1.
composed of many interconnected parts; compound; composite: a complex highway system.
2.
characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.
3.
so complicated or intricate as to be hard to understand or deal with: a complex problem.

Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

Complexity

Complex Systems

- Self-organised
- Evolving
- Adaptive

- No central organising mind
- No conventional way of description

Complex Systems: how to approach

Statistical description

- Systems with random features
- One sample does not characterise the typical behaviour
- Statistical averages of quantities

Empirical data analysis

- How to detect patterns and structure in information?
- How to characterize the system instead of its building blocks?
- Multivariate methods etc

Analytical approach

- Write down (coupled) differential equations for interactions
- Attempt to solve
- Usually no closed-form solutions; numerical solutions, phase space analysis, etc

Simulations

- Postulate rules (e.g. the ant raids)
- Simulate and observe system behaviour
- Try to match empirical observations

OR

Complex Networks

…a way of mapping complexity

Each complex system can be interpreted as a complex network, which identifies the interactions between the interconnected components

The network approach

- Combines the elements of all the other approach
- Disregards (unnecessary) details of the system
- Focuses on the structure of interactions
- Statistical characterisation of system

The network approach

1. **Measuring** - make observations on Nature
2. **Modelling** - attempt to explain observations:
   2.1. Choose the right level of coarse-graining
   - Units: Vertices or nodes ⇔ interacting elements
   - Edges or links ⇔ interactions

   2.2. Strip the problem to its simplest form
   - Interaction structure ⇔ evolution and behaviour of system

   2.3. Formulate the problem in mathematical terms
   - Statistical analysis of network structure
   - Dynamics of processes taking place on networks
3. **Validating** - check if calculations or simulations can
   - reproduce the observations
   - explain the observations
4. Go back to 1. & 2. and rethink

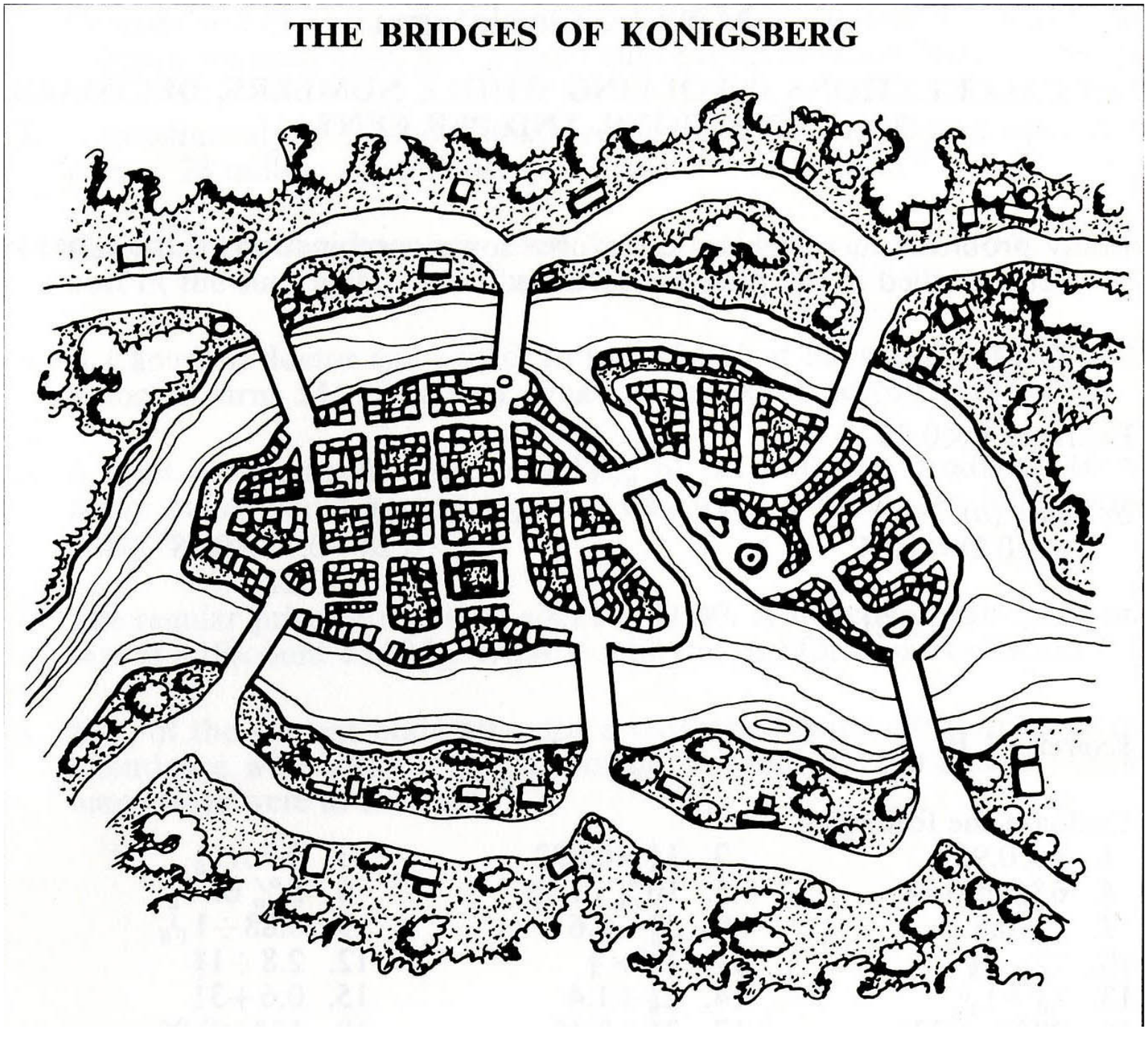
Graphs

or

Networks

The Seven Bridges of Königsberg

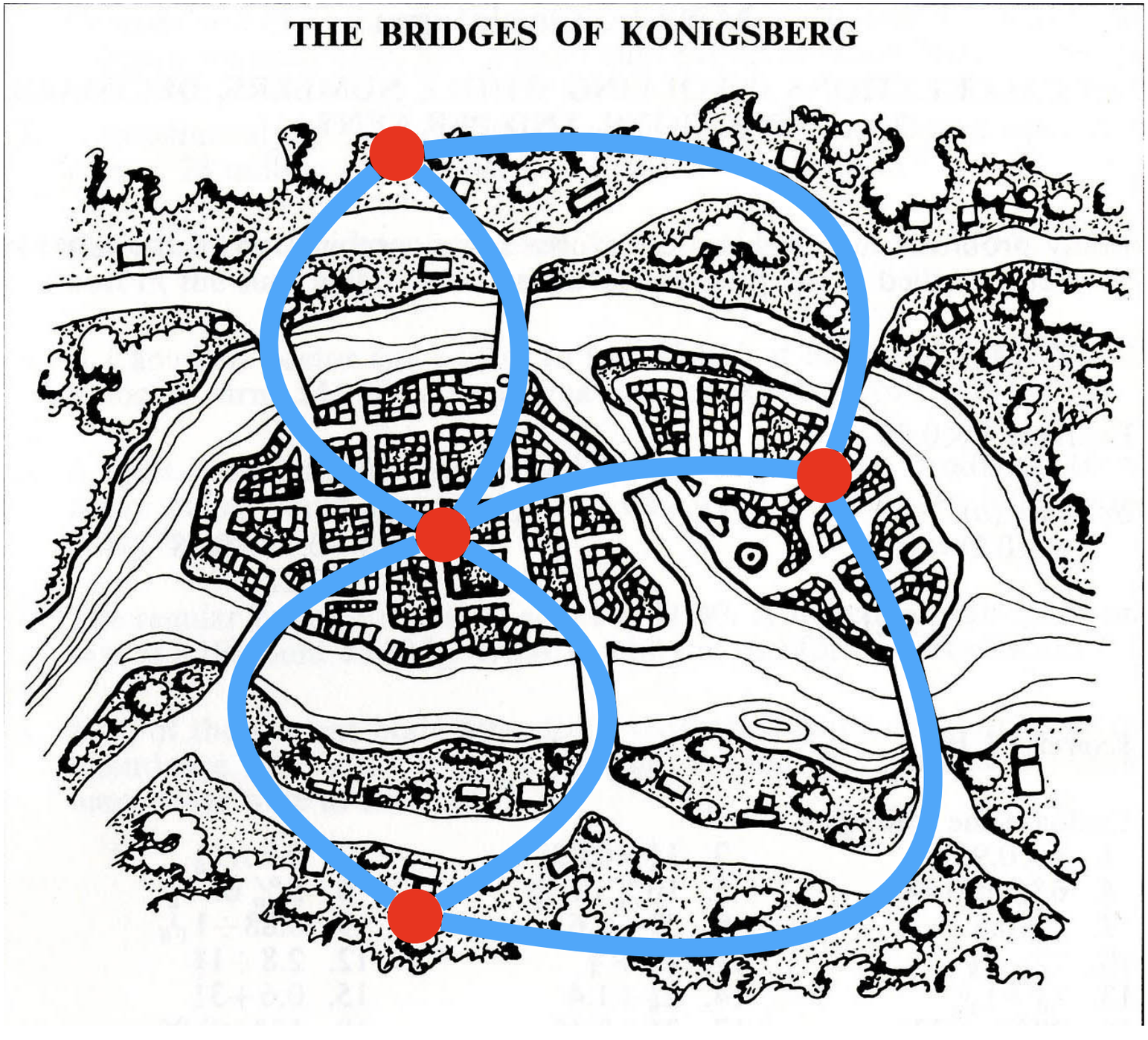
Leonhard Euler (1736)



Can one walk across the seven bridges and never cross the same bridge twice?

The Seven Bridges of Königsberg

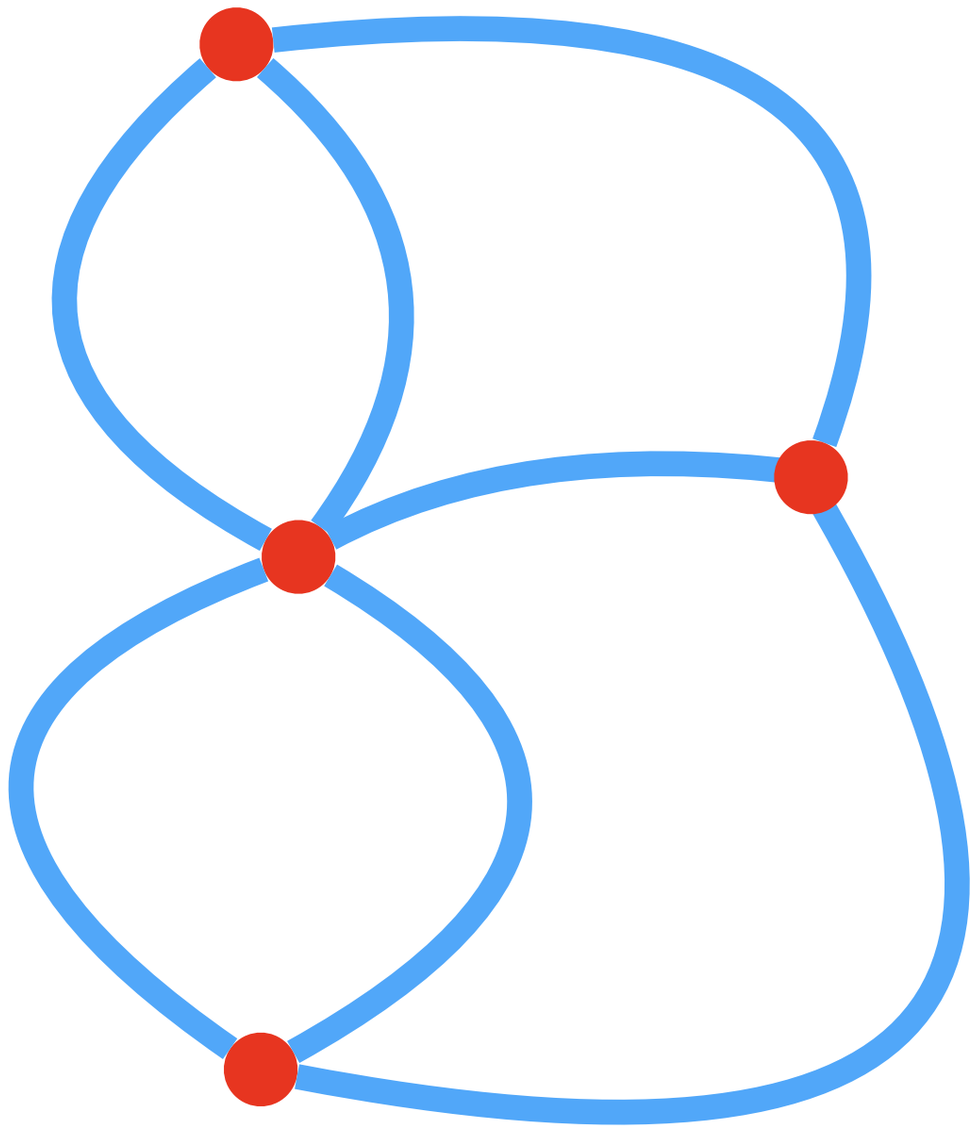
Leonhard Euler (1736)



Can one walk across the seven bridges and never cross the same bridge twice?

The Seven Bridges of Königsberg

Leonhard Euler (1736)



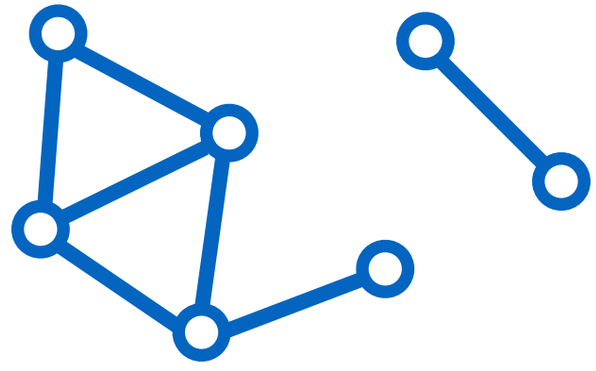
Can one walk across the seven bridges and never cross the same bridge twice?

Answer: **No**

Complex systems as networks

Networks are interpreted as graphs

$$G=(V, E)$$



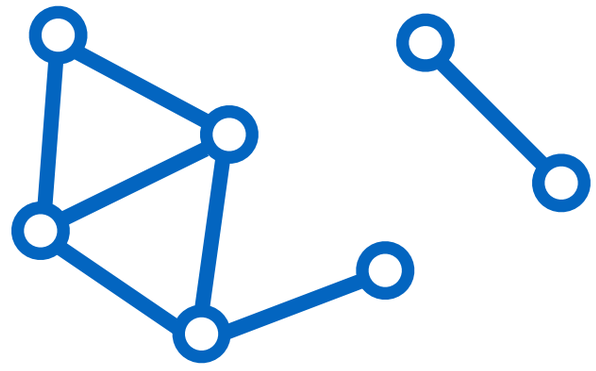
- Components ⇔ vertices $v \in V$
- Interactions between components ⇔ edges $(u,v) \in E$
- Identification of vertices and edges defines the type of the actual network

Vertex	Edge
person	friendship
neuron	synapse
WWW	hyperlink
company	ownership
gene	regulation

Graphs or Networks

Networks are sparse and often refer to real systems

- www,
- social network
- metabolic network.



Language: (Network, node, link)

Graph are dense and mathematically treatable

Language: (Graph, vertex, edge)

Vertex	Edge
person	friendship
neuron	synapse
WWW	hyperlink
company	ownership
gene	regulation

We will use in most cases the two terms interchangeably.

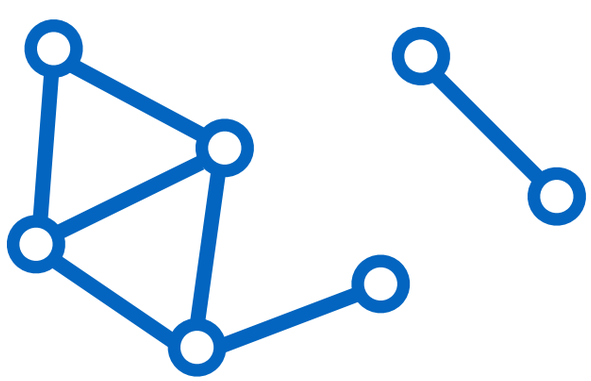
Types of Networks

Undirected networks

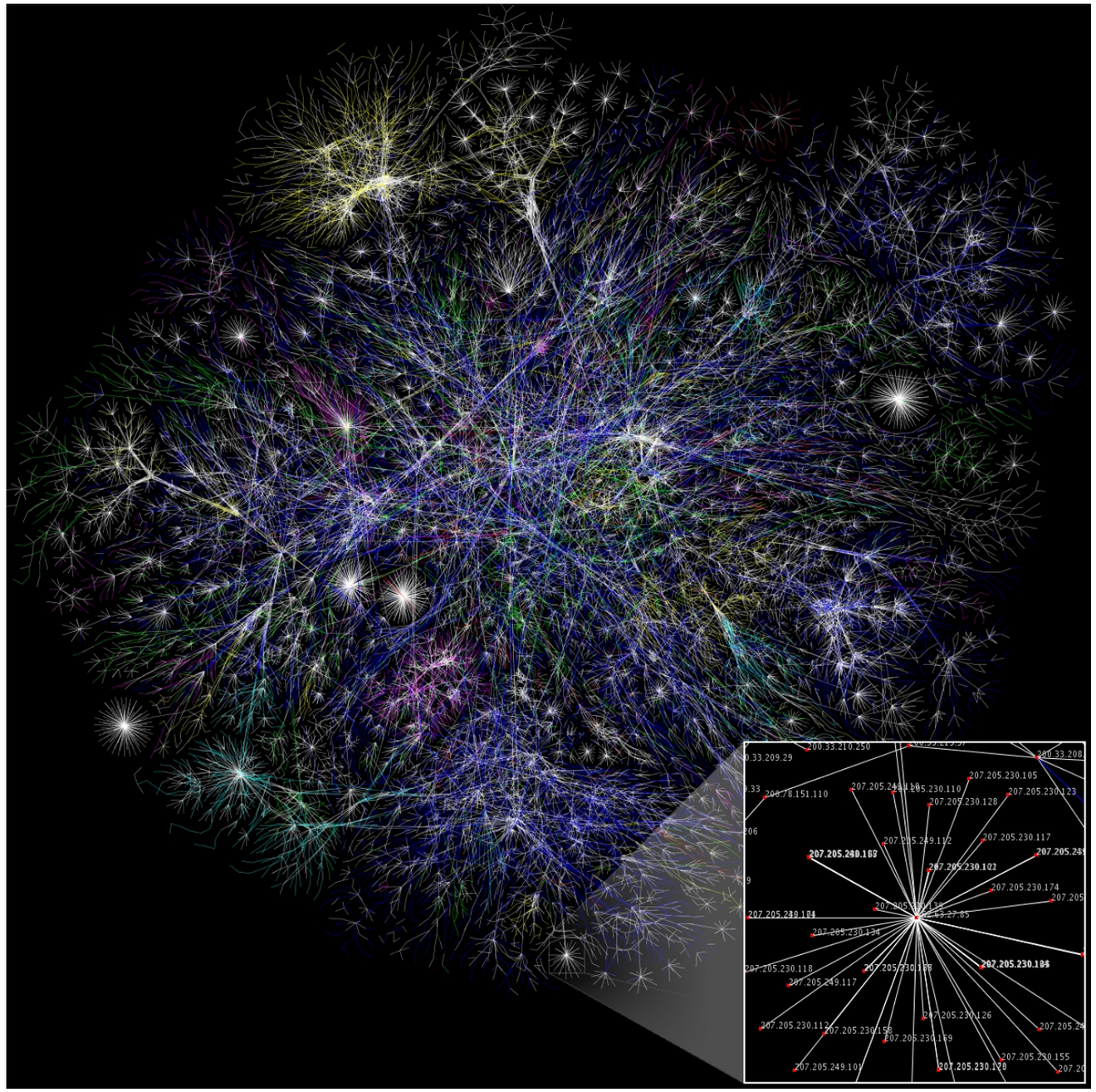
$$G=(V, E)$$

$$(u,v) \in E \equiv (v,u) \in E$$

- The directions of edges do not matter
- Interactions are possible between connected entities in both directions



Opte project



The Internet: Nodes - routers, Links - physical wires

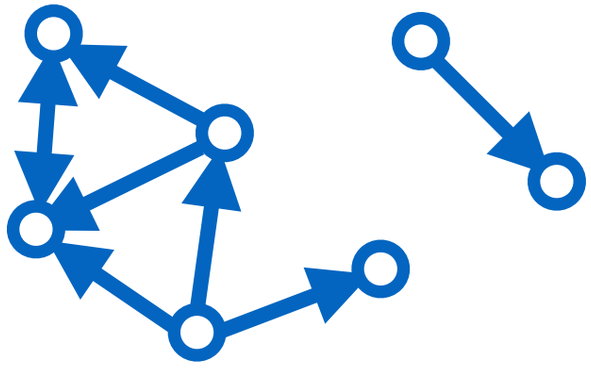
Directed networks

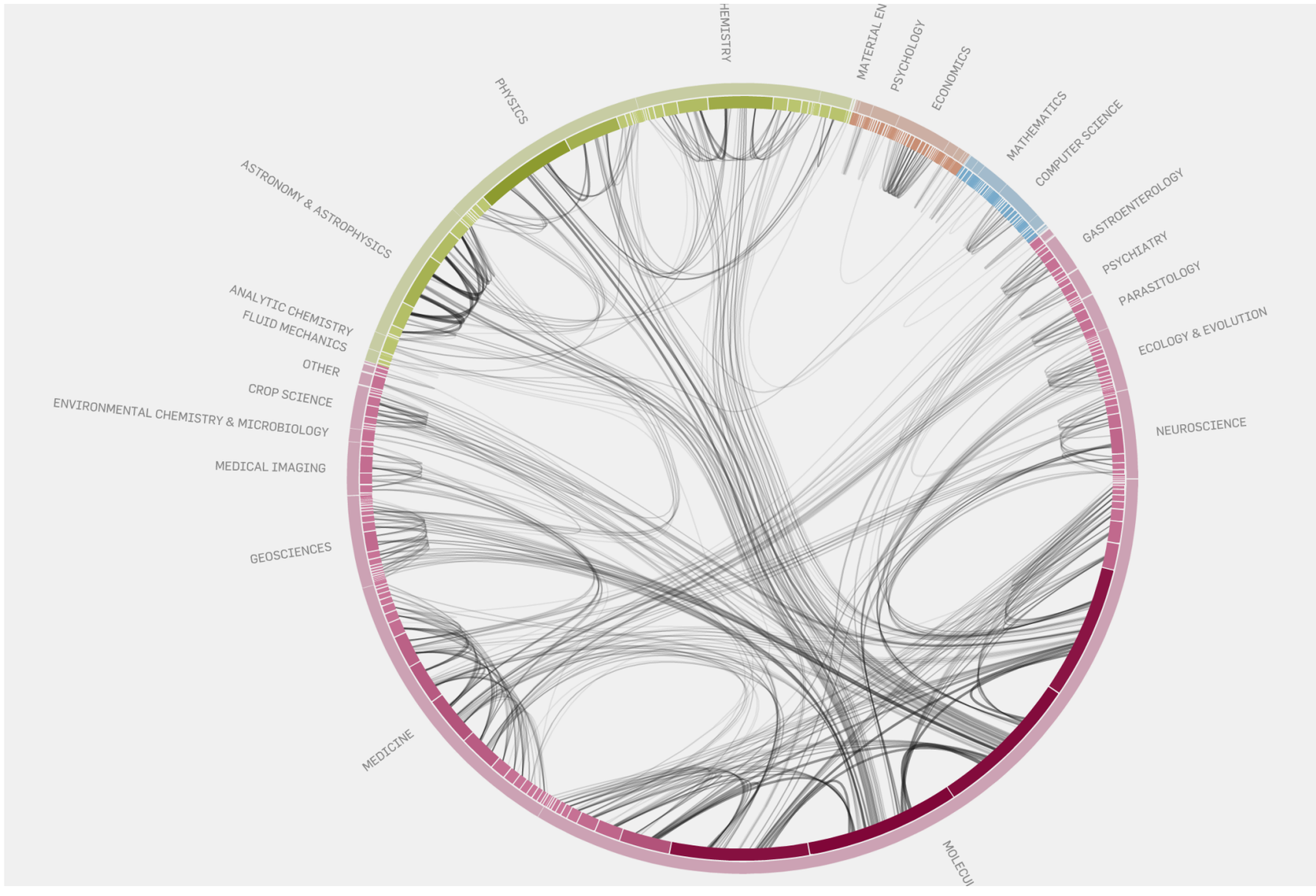
Moritz Stefaner, eigenfactor.com

$$G=(V, E)$$

$$(u,v) \in E \not\equiv (v,u) \in E$$

- The directions of edges matter
- Interactions are possible between connected entities only in specified directions





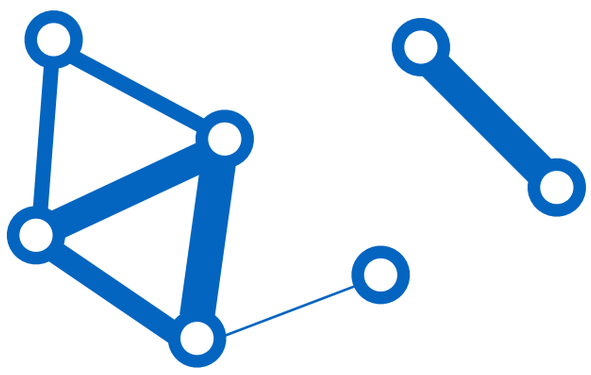
Citation network: Nodes - publications, Links - references

Weighted networks

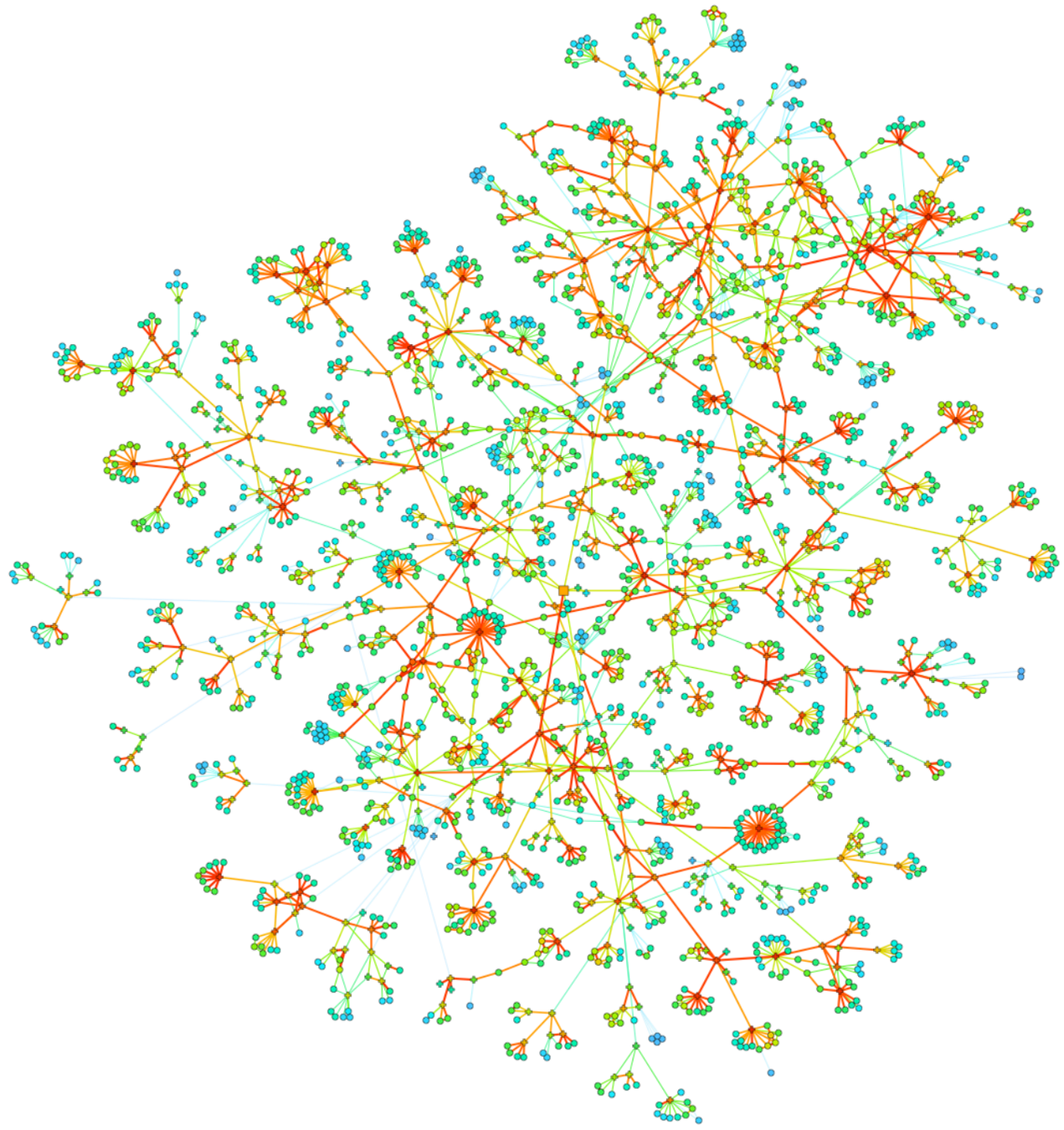
$G=(V, E, w)$

$w: (u,v) \in E \Rightarrow R$

- Strength of interactions are assigned by the weight of links

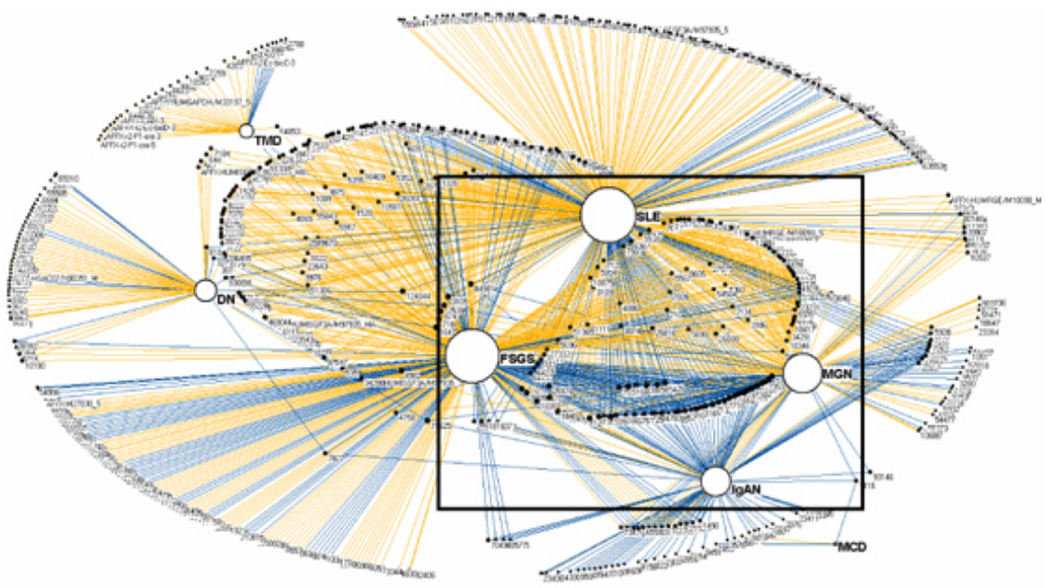


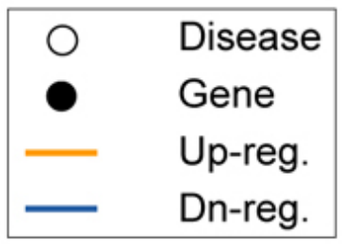
Onnela et.al. New Journal of Physics 9, 179 (2007).



Social interaction network: Nodes - individuals
Links - social interactions

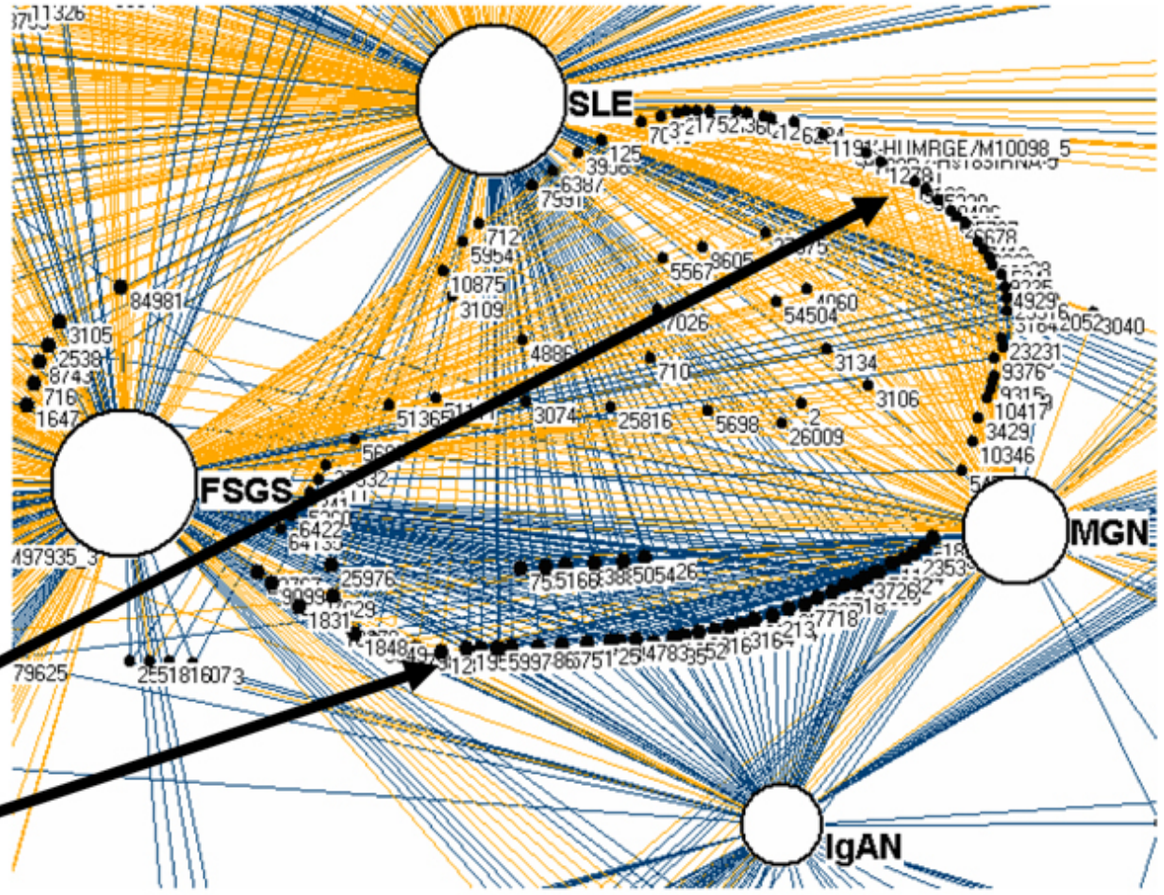
Bipartite network



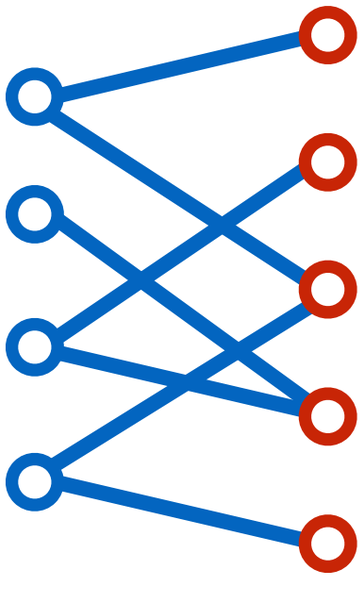


Genes (mostly up-regulated) by SLE, FSGS, and MGN

Genes (mostly down-regulated) by SLE, FSGS, MGN, and IgAN



Bhavnani et.al. BMC Bioinformatics 2009, **10**(Suppl 9):S3



Gene-desease network:

Nodes - Desease (7)&Genes (747)

Links - gene-desease relationship

$$G=(U, V, E)$$

$$U \cap V = \varnothing$$

$$\forall (u,v) \in E,\ u \in U \text{ and } v \in V$$

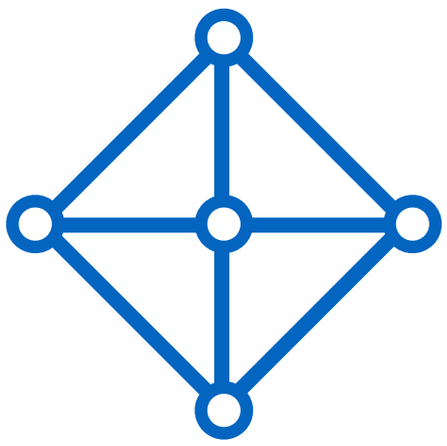
Planar networks

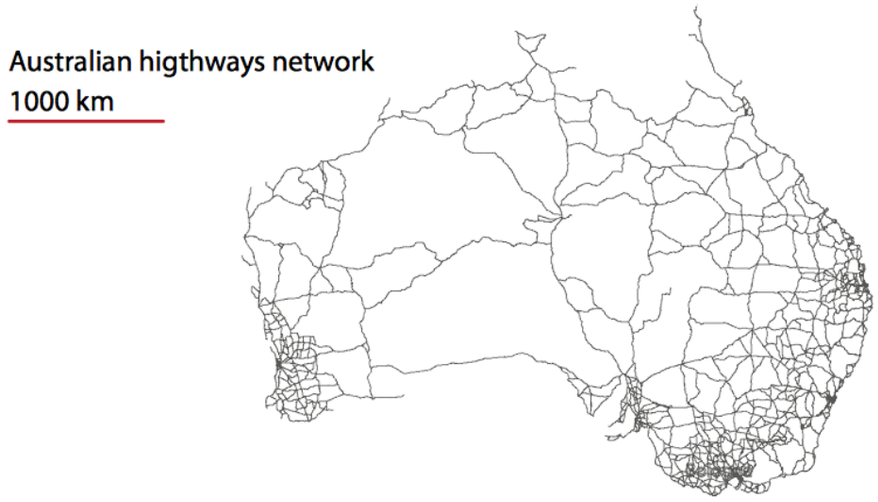
Viana et.al. Nature Scientific Reports 3:3495 (2013)

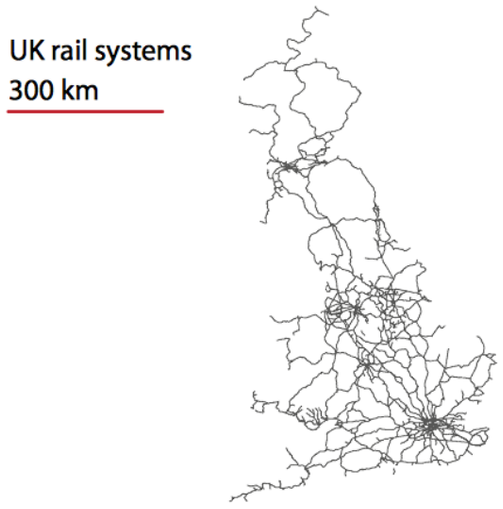
$G=(V, E, loc)$

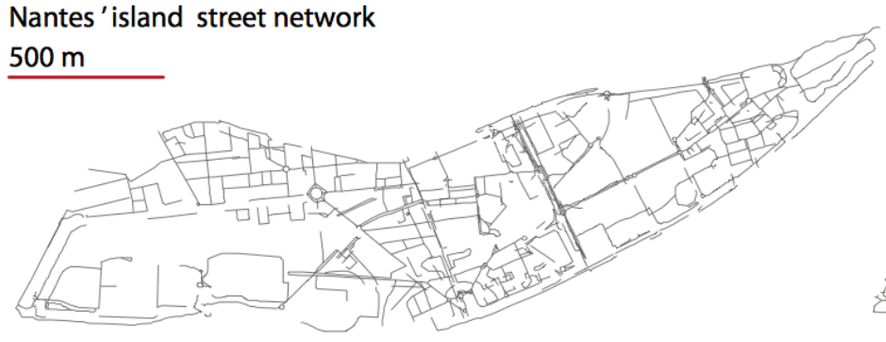
$loc: v \in V \Rightarrow (x,y)$

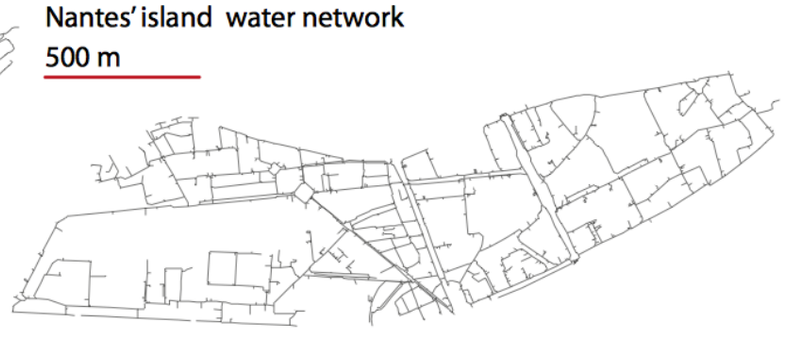
- Nodes can be embedded in a plane
- Geo-localised networks
- Spatial networks











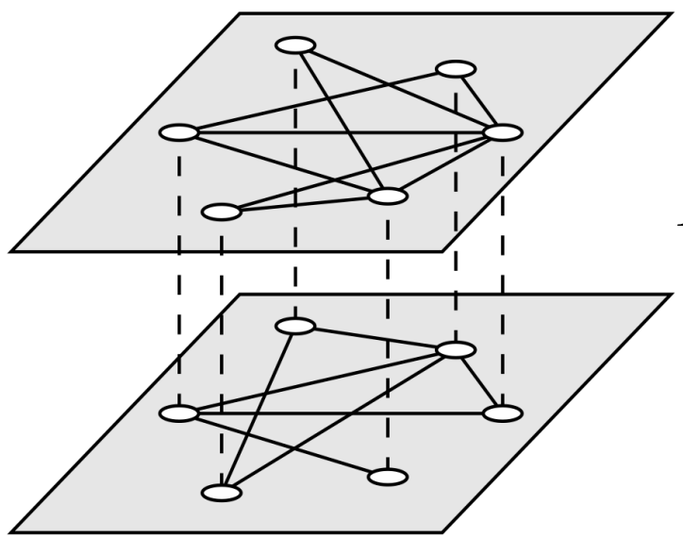
Street networks:

Nodes - junctions, Links - streets

Multiplex and multilayer networks

$G=(V, E_i), i=1 \ldots M$

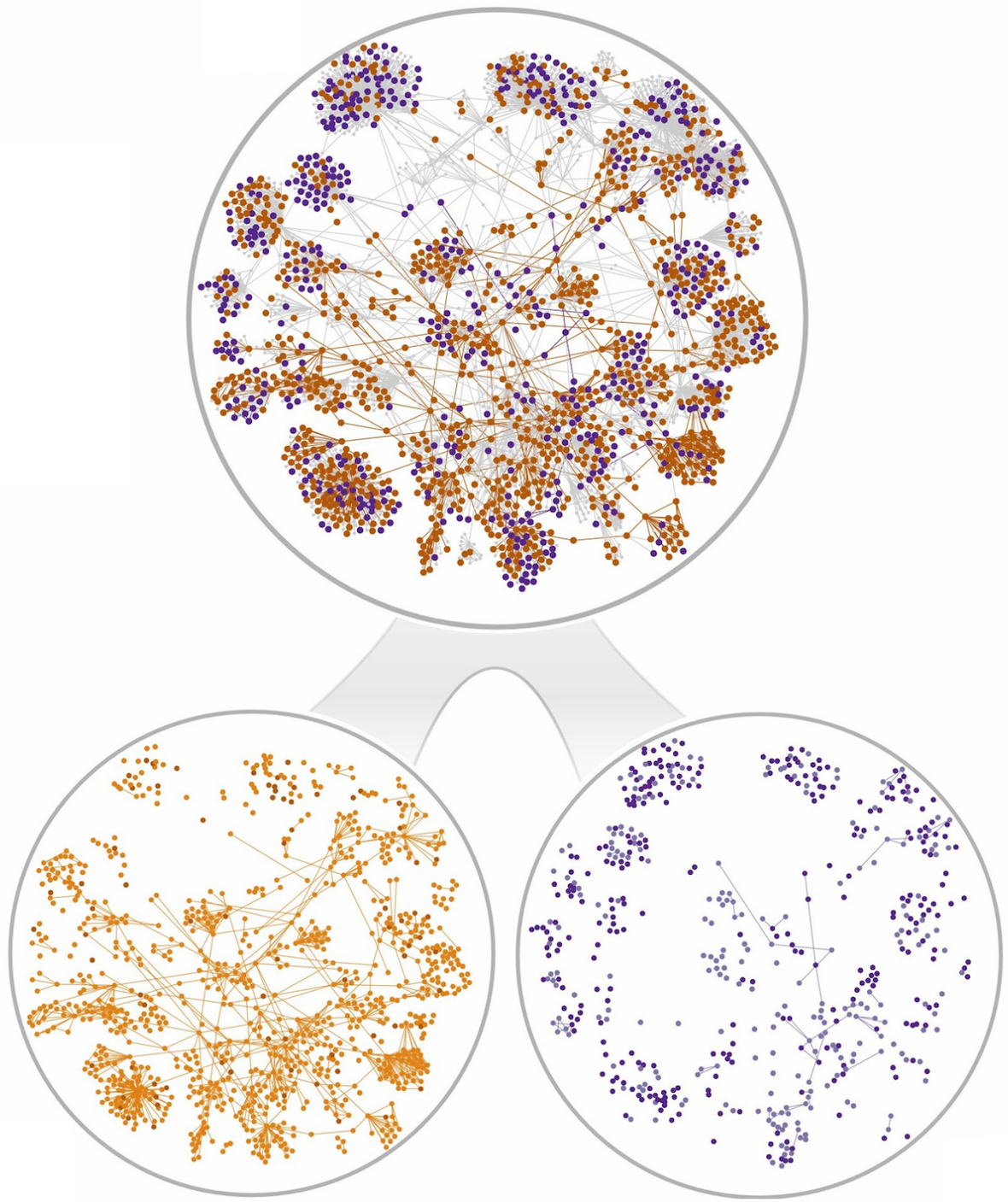
- Nodes can be present in multiple networks simultaneously
- These networks are connected (can influence each other) via the common nodes



$M=2$

Gomes et.al. Phys. Rev. Lett. 110, 028701 (2013)

Karsai et.al. (submitted)



Skype adoption network

Nodes - users, Links - social ties,

Colours - service adoption/termination

Temporal and evolving networks

$G=(V, E_t)$, $(u,v,t,d) \in E_t$

t - time of interaction (u,v)
d - duration of interaction (u,v,t)

- Temporal links encode time varying interactions

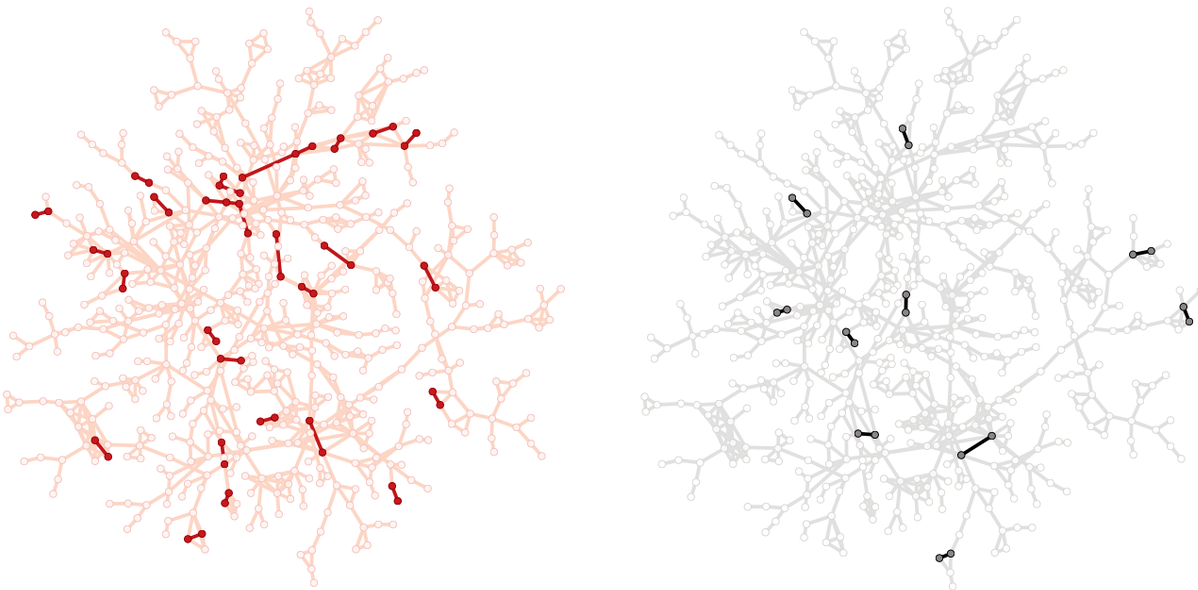
$G=(V_{t'}, E_{t'})$

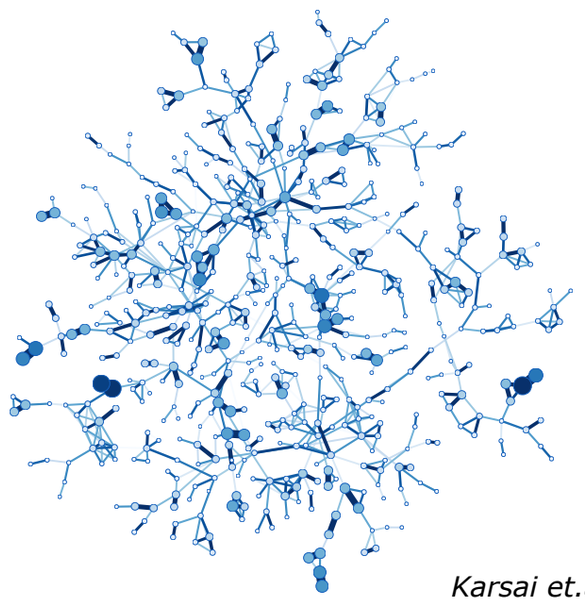
$v(t) \in V_{t'}$

$(u,v,t) \in E_{t'}$

- Dynamical nodes and links encode the evolution of the network
- Usually t<<t'

Mobile communication network
Nodes - individuals
Links - calls and SMS





Karsai et.al.

WHY

and

WHY

NOW?

Why?

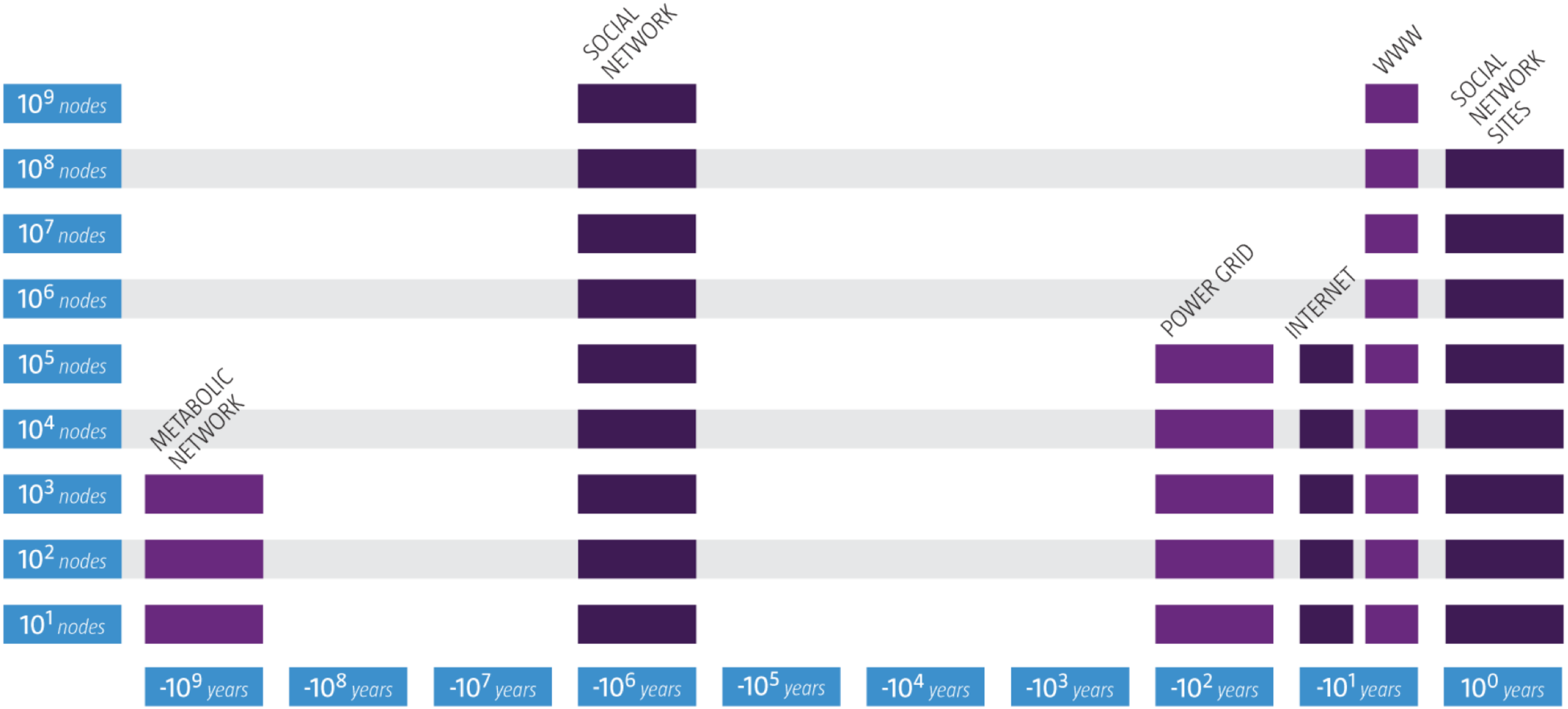
- A common framework applicable to many systems
- Different systems can be studied with same methods
- A “birds-eye” view on the system

MANY NETWORKS SHARE SIMILAR CHARACTERISTICS

- Similar processes shape the networks

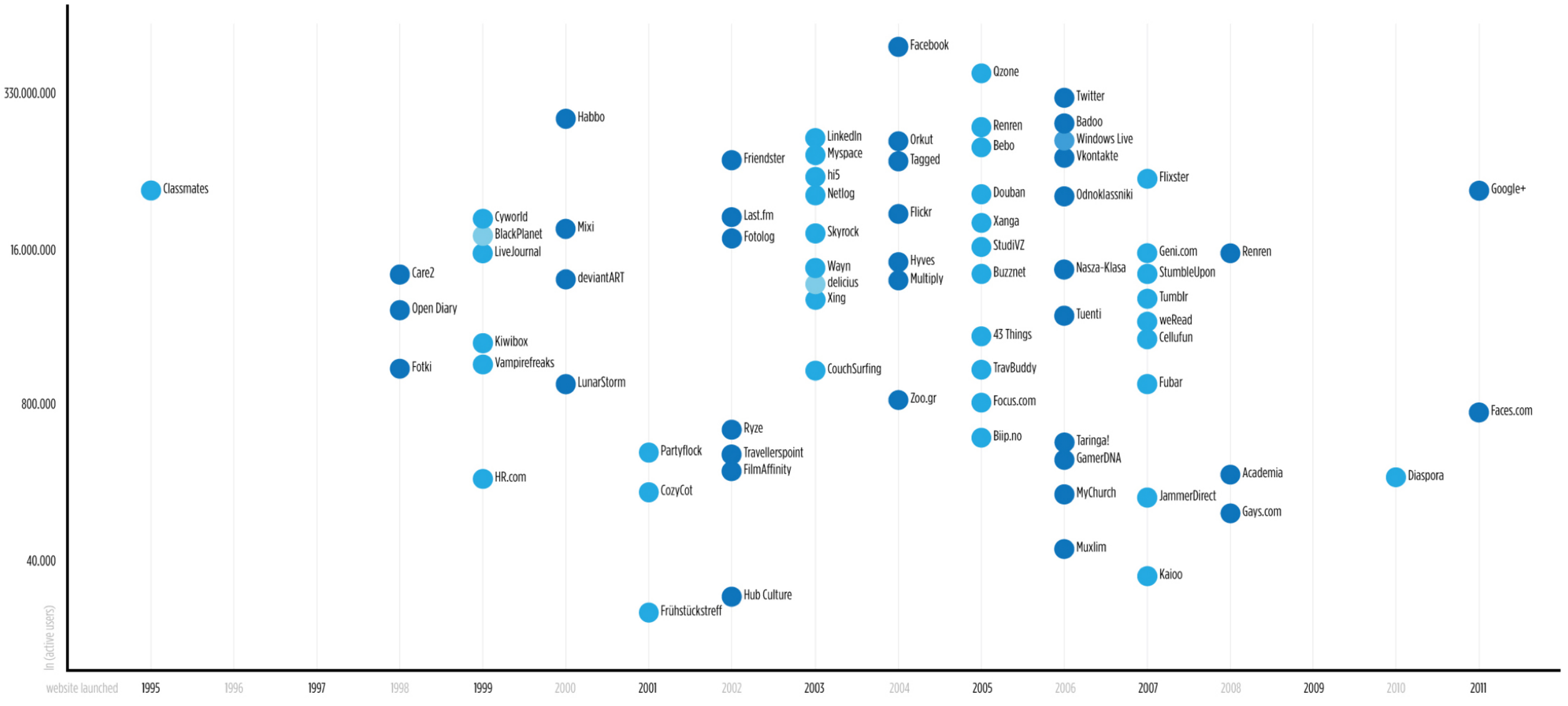
WE WILL NEVER UNDERSTAND COMPLEX SYSTEMS UNLESS WE MAP OUT AND UNDERSTAND THE NETWORKS BEHIND THEM *AL Barabási*

Why now?



Slide from CCNR Complex Networks Course
A. L. Barabási 2012

Why now?



Slide from CCNR Complex Networks Course
A. L. Barabási 2012

Why now?

- Data availability - the Big Data Revolution
- Universality - similar features of different systems
- Urgent need to understand complexity
  - Economic impact
  - Drug design, metabolic engineering
  - Human decease network
  - Fighting, terrorism and military
  - Epidemic forecast
  - Brain research

Network representations and network properties

(Large) Network representation

Adjacency matrix

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Advantage

Direct access

Disadvantage

Memory hungry

(Large) Network representation

Adjacency matrix

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Advantage

Direct access

Disadvantage

Memory hungry

Edge list

```
1 2
2 3
2 4
3 4
4 5
4 7
5 6
5 8
9 10
```

Memory friendly

Slow search

(Large) Network representation

Adjacency matrix

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Advantage: Direct access

Disadvantage: Memory hungry

Edge list

```
1 2
2 3
2 4
3 4
4 5
4 7
5 6
5 8
9 10
```

Advantage: Memory friendly

Disadvantage: Slow search

Neighbour (adjacency) list

```
1   2
2   1 3 4
3   2 4
4   2 3 5 7
5   4 6 8
6   5
7   4
8   5
9   10
10  9
```

Advantage: Memory optimised

Disadvantage: Fast search

(Large) Network representation

Directed networks

- Non-symmetric adjacency matrix
- Order of adjacency link list

```
   1 2 3 4 5 6 7 8 9 10

1  0 1 0 0 0 0 0 0 0 0
2  0 0 1 0 0 0 0 0 0 0
3  0 1 0 0 0 0 0 0 0 0
4  0 1 1 0 0 0 1 0 0 0
5  0 0 0 1 0 1 0 1 0 0
6  0 0 0 0 1 0 0 0 0 0
7  0 0 0 1 0 0 0 0 0 0
8  0 0 0 0 1 0 0 0 0 0
9  0 0 0 0 0 0 0 0 0 1
10 0 0 0 0 0 0 0 0 1 0
```

```
1   2
2   1 3 4
3   2
4   2 3 5 7
5   4 6 8
6
7   4
8
9   10
10  9
```

```
1 2
2 3
2 4
3 4
4 5
4 7
5 6
5 8
9 10
```

(Large) Network representation

Directed networks

- Non-symmetric adjacency matrix
- Order of adjacency link list

```
   1 2 3 4 5 6 7 8 9 10

1  0 1 0 0 0 0 0 0 0 0
2  0 0 1 0 0 0 0 0 0 0
3  0 1 0 0 0 0 0 0 0 0
4  0 1 1 0 0 0 1 0 0 0
5  0 0 0 1 0 1 0 1 0 0
6  0 0 0 0 1 0 0 0 0 0
7  0 0 0 1 0 0 0 0 0 0
8  0 0 0 0 1 0 0 0 0 0
9  0 0 0 0 0 0 0 0 0 1
10 0 0 0 0 0 0 0 0 1 0
```

```
1   2
2   1 3 4
3   2
4   2 3 5 7
5   4 6 8
6
7   4
8
9   10
10  9
```

```
1 2
2 3
2 4
3 4
4 5
4 7
5 6
5 8
9 10
```

Weighted networks

- Matrix element assign the weight of links
- Edges: non-zero elements

```
   1 2 3 4 5 6 7 8 9 10
1  0 6 0 0 0 0 0 0 0 0
2  6 0 1 3 0 0 0 0 0 0
3  0 1 0 3 0 0 0 0 0 0
4  0 3 3 0 6 0 4 0 0 0
5  0 0 0 6 0 1 0 3 0 0
6  0 0 0 0 1 0 0 0 0 0
7  0 0 0 4 0 0 0 0 0 0
8  0 0 0 0 3 0 0 0 0 0
9  0 0 0 0 0 0 0 0 0 4
10 0 0 0 0 0 0 0 0 4 0
```

- (e,w) tuples

```
1   (2,6)
2   (1,3) (3,1) (4,3)
3   (2,1) (4,3)
4   (2,3) (3,3) (5,6) (7,4)
5   (4,6) (6,1) (8,3)
6   (5,1)
7   (4,4)
8   (5,3)
9   (10,4)
10  (9,4)
```

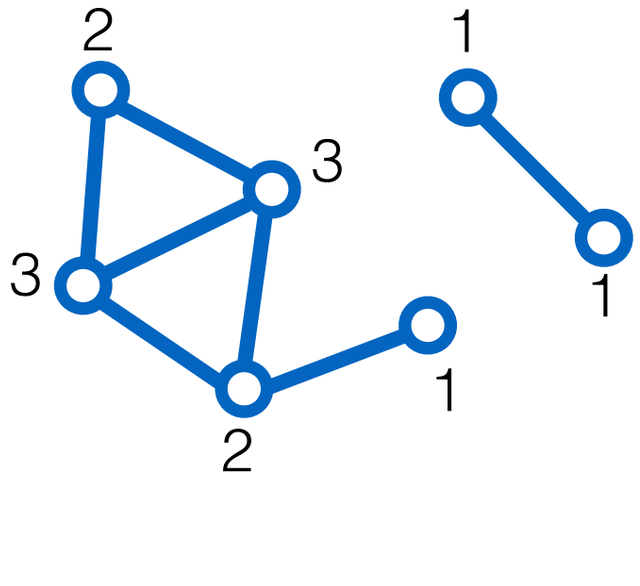
- (b,e,w) triplets

```
1 2 6
2 3 1
2 4 3
3 4 3
4 5 6
4 7 4
5 6 1
5 8 3
9 10 4
```

Node degree

Number of connections of a node

- Undirected network



$$k_i = A_{i1} + A_{i2} + ... + A_{iN} = \sum_j^N A_{ij}$$

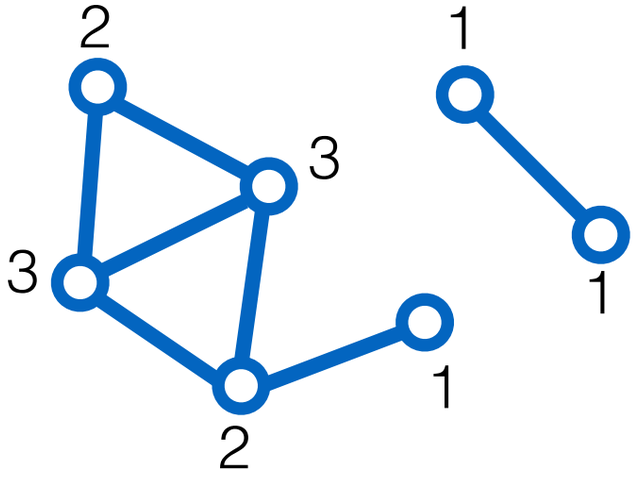
$$m = \frac{\sum_i k_i}{2} \qquad \text{where} \qquad m = |E|$$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Node degree

Number of connections of a node

- Undirected network

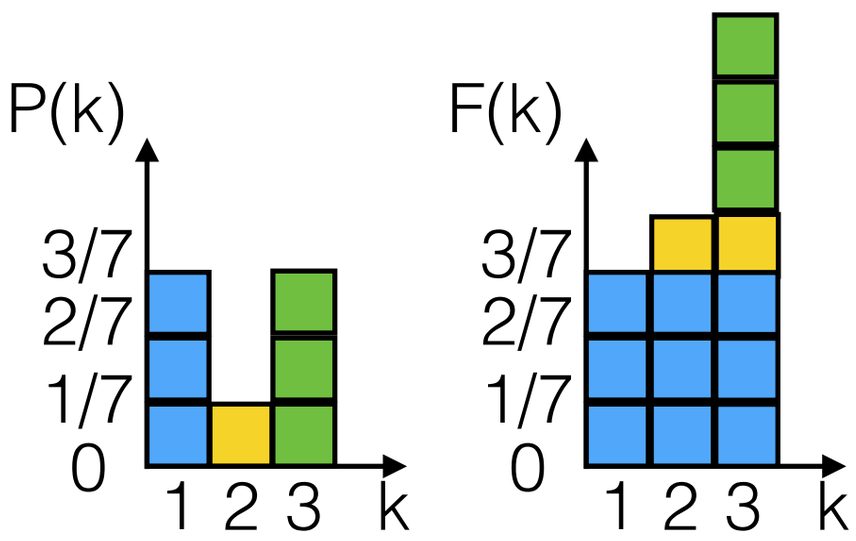


$$k_i = A_{i1} + A_{i2} + ... + A_{iN} = \sum_j^N A_{ij}$$

$$m = \frac{\sum_i k_i}{2} \qquad \text{where} \qquad m = |E|$$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

- Degree distribution:



PDF of k or CDF of k

$$P(k) = \frac{n_k}{\sum_k n_k} \qquad F(k) = \sum_{k_i=1}^{k} P(k_i)$$

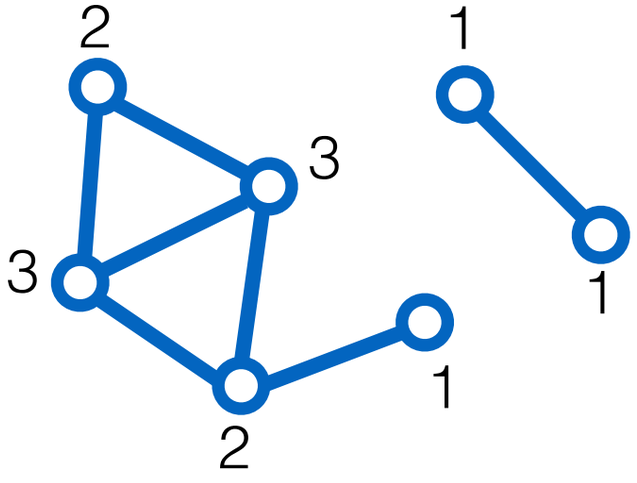
normalisation condition

$$\sum_{k=1}^{\infty} P(k) = 1 \qquad \lim_{k\to\infty} F(k) = 1$$

Node degree

Number of connections of a node

- Undirected network

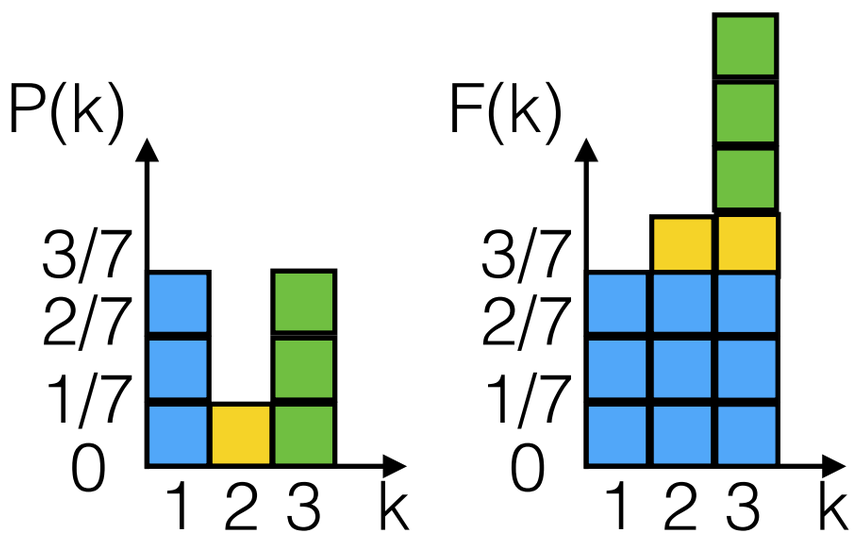


$$k_i = A_{i1} + A_{i2} + ... + A_{iN} = \sum_j^N A_{ij}$$

$$m = \frac{\sum_i k_i}{2} \qquad \text{where} \qquad m = |E|$$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

- Degree distribution:



PDF of k or CDF of k

$$P(k) = \frac{n_k}{\sum_k n_k} \qquad F(k) = \sum_{k_i=1}^{k} P(k_i)$$

normalisation condition

$$\sum_{k=1}^{\infty} P(k) = 1 \qquad \lim_{k\to\infty} F(k) = 1$$

moments

$$\langle k^n \rangle = \sum_{k=1}^{N-1} k^m P(k) = \frac{1}{N} \sum_{k=1}^{N-1} k^m n_k = \frac{1}{N} \sum_{i=1}^{N} k_i^m$$

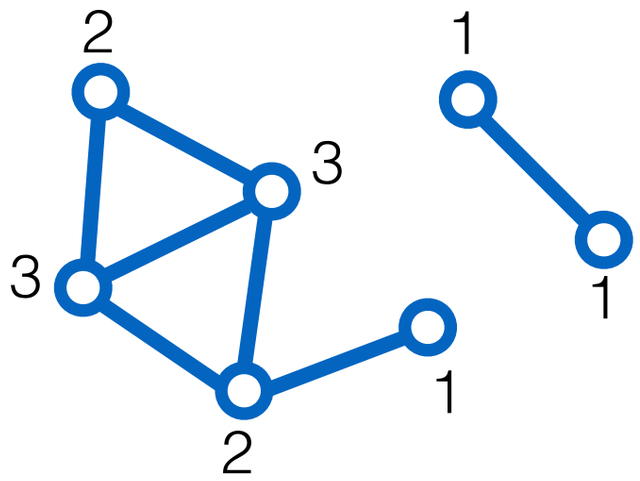
average degree

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i$$

Node degree

Number of connections of a node

- Undirected network



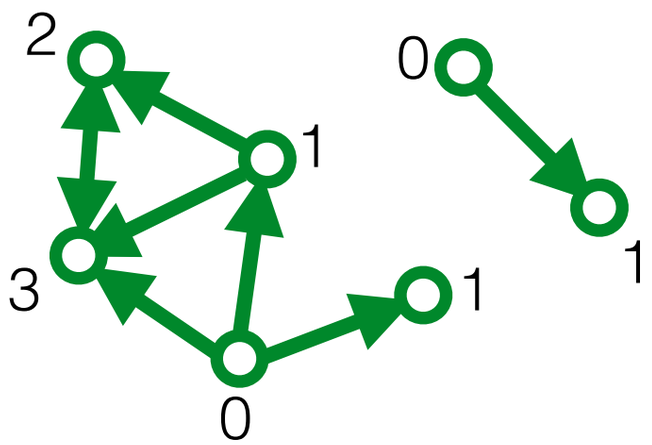
$$k_i = A_{i1} + A_{i2} + ... + A_{iN} = \sum_j^N A_{ij}$$

$$m = \frac{\sum_i k_i}{2} \qquad \text{where} \qquad m = |E|$$

mean degree $$\langle k \rangle = \frac{1}{N} \sum_i^N k_i$$

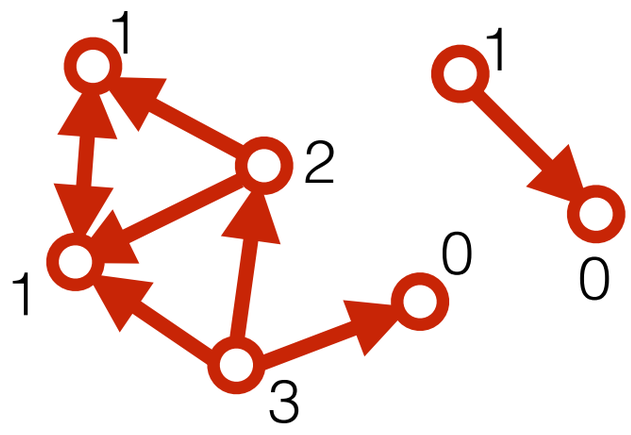
	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

- Directed network



In degree $$k_i^{in} = \sum_j^N A_{ij}$$

Out degree $$k_j^{out} = \sum_i^N A_{ij}$$



$$m = \sum_i^N k_i^{in} = \sum_j^N k_j^{out} = \sum_{ij} A_{ij}$$

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^N k_i^{in} = \frac{1}{N} \sum_{j=1}^N k_j^{out} = \langle k^{out} \rangle$$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

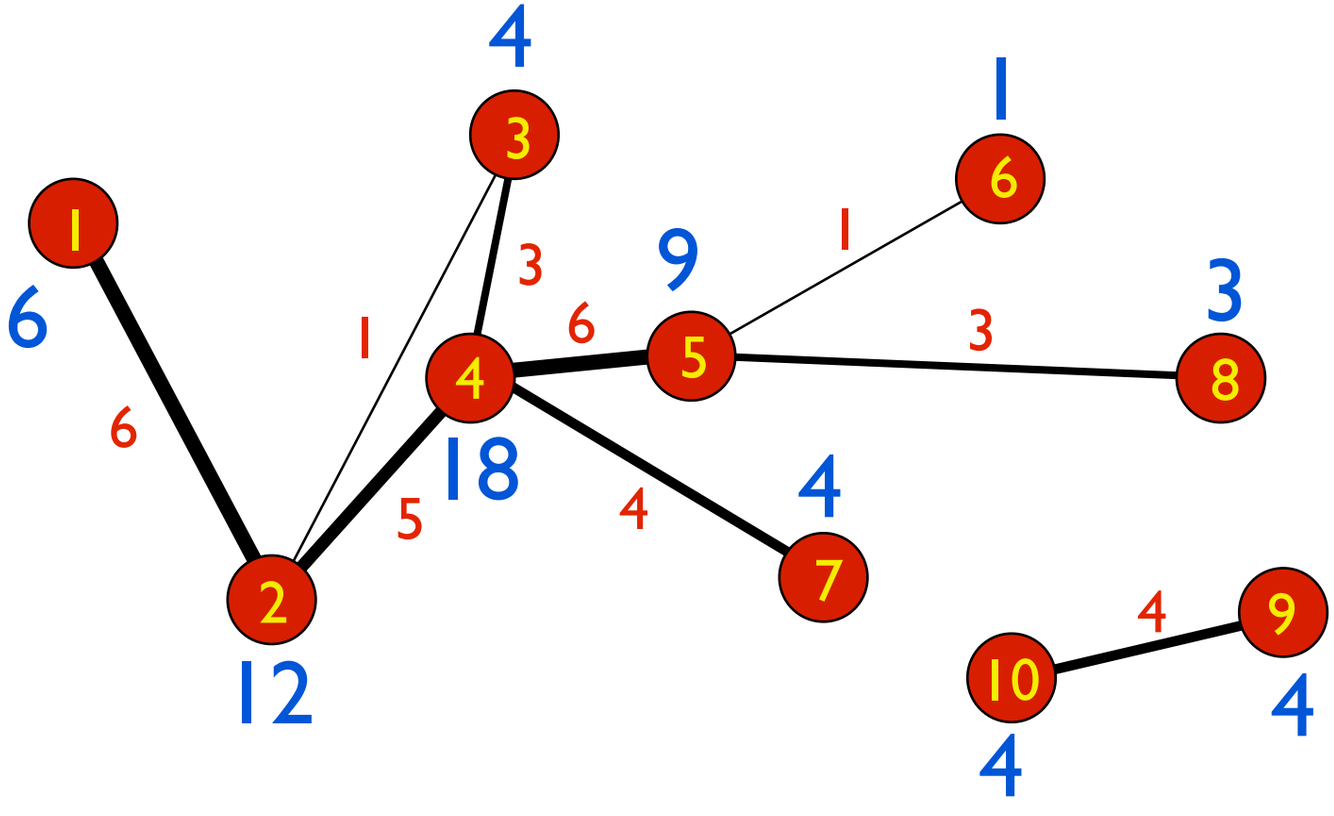
	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	1	1	0	1	0	1	0	0	0
5	0	0	0	1	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

Weighted degree: strength

- Weighted networks

The sum of the weights of links connected to node i

$$s_i = w_{i1} + w_{i2} + \ldots + w_{iN} = \sum_j w_{ij}$$



	1	2	3	4	5	6	7	8	9	10
1	0	6	0	0	0	0	0	0	0	0
2	6	0	1	3	0	0	0	0	0	0
3	0	1	0	3	0	0	0	0	0	0
4	0	3	3	0	6	0	4	0	0	0
5	0	0	0	6	0	1	0	3	0	0
6	0	0	0	0	1	0	0	0	0	0
7	0	0	0	4	0	0	0	0	0	0
8	0	0	0	0	3	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	4
10	0	0	0	0	0	0	0	0	4	0

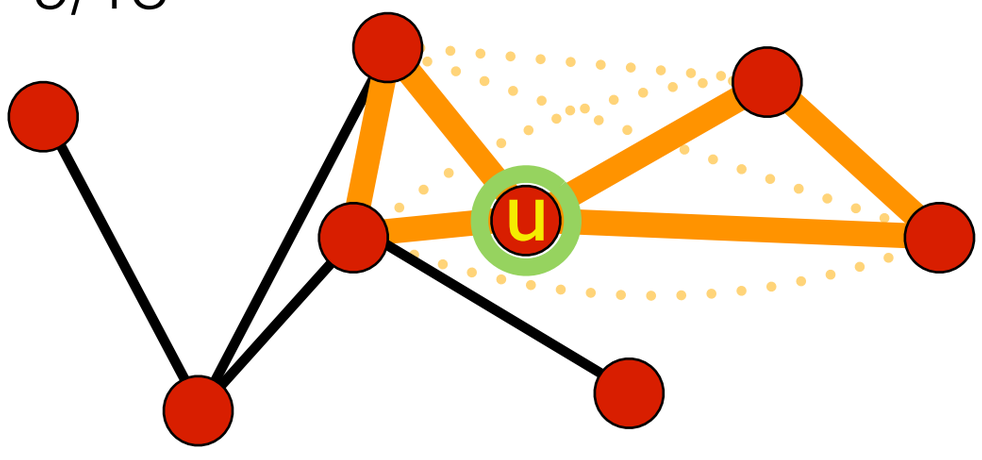
Node clustering coefficient

- Measure of interconnectivity
- What portion of neighbours of a node are connected to each other?

Global clustering coefficient

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}}.$$

C=9/18



Node clustering coefficient

- Measure of interconnectivity
- What portion of neighbours of a node are connected to each other?

Global clustering coefficient

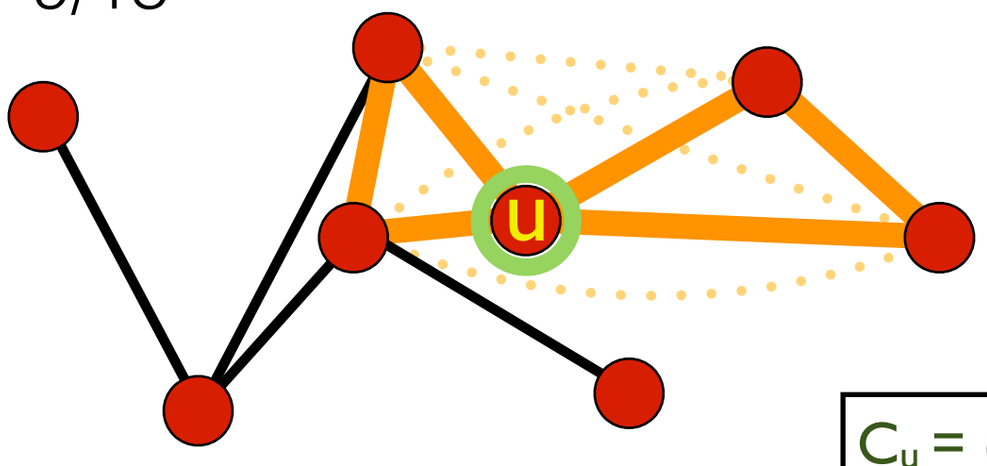
$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}}.$$

Local clustering coefficient

$$C_u = \frac{2e_u}{k_u(k_u - 1)}$$

- e_u - number of links between the neighbours of node u
- $(k_u(k_u-1))/2$ - maximum number of triangles

C=9/18



$C_u = (2x2)/(4x3) = 1/3$

Average local clustering coefficient

$$C = \frac{1}{N} \sum_u C_u$$

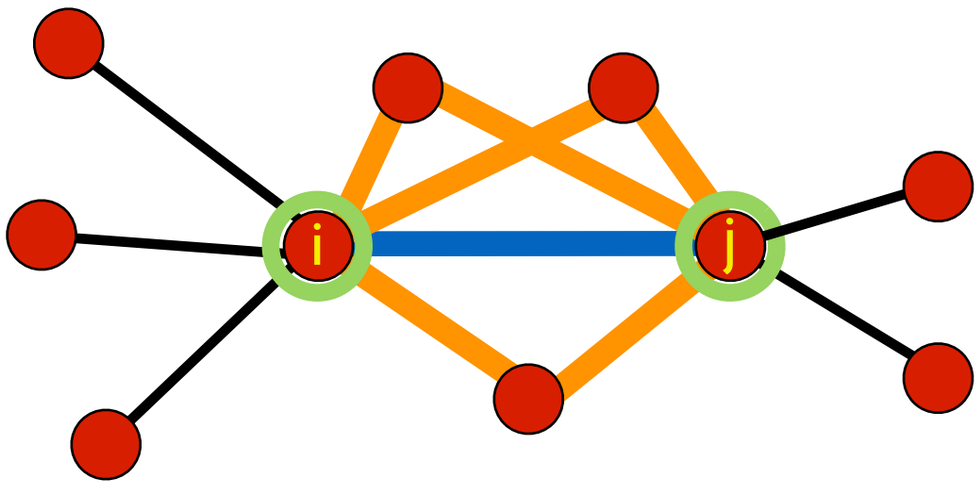
Definition: Watts and Strogatz 2002

Link clustering coefficient: Overlap

- Link property
- Fraction of common neighbours of a connected pair
- Jaccard index of common neighbours

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_i - 1) - n_{ij}}$$

- n_i - number of common neighbours of nodes i and j
- $(k_i\text{-}1)+(k_j\text{-}1)\text{-}n_{ij}$ maximum number possible triangles between nodes i and j



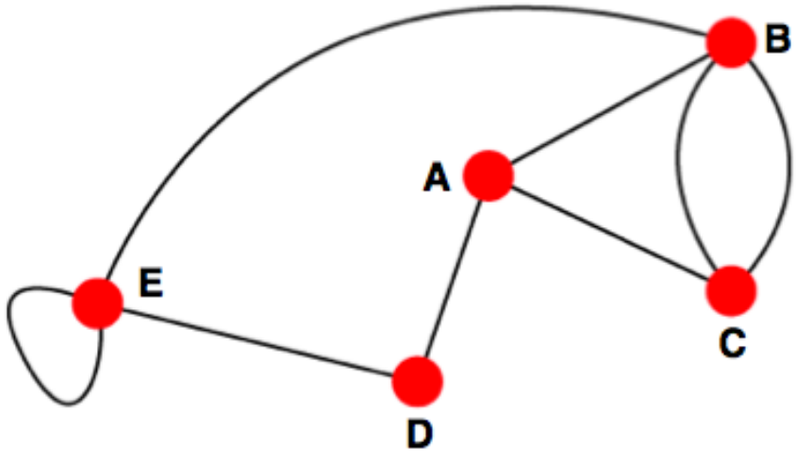
O_{ij} = 3/(6+5-3) = 3/8

Path length

A ***path is*** a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length ***n*** between nodes i_0 and i_n is an ordered collection of ***n+1*** nodes and ***n*** links

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

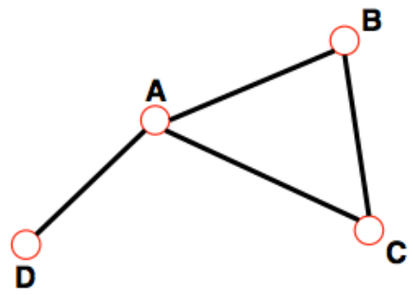


•A path can intersect itself and pass through the same link repeatedly. Each time a link is crossed, it is counted separately

•A legitimate path on the graph on the right:
ABCBCADEEBA

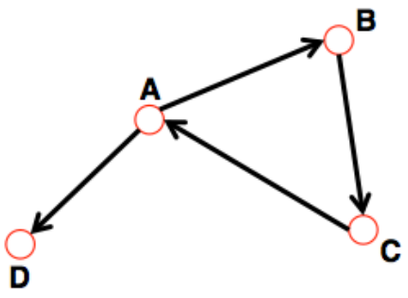
• In a directed network, the path can follow only the direction of an arrow.

Path length



The ***distance (shortest path, geodesic path)*** between two nodes is defined as the number of edges along the shortest path connecting them.

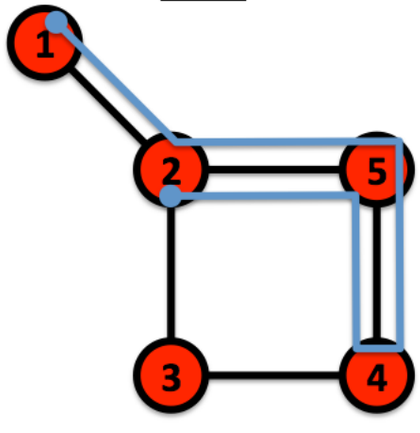
*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

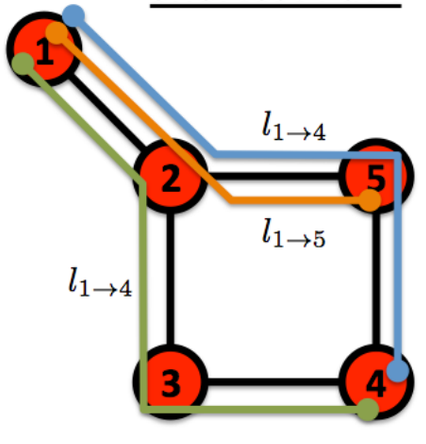
Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

Path



A sequence of nodes such that each node is connected to the next node along the path by a link.

Shortest Path



$l_{1\to4} = 3$

$l_{1\to5} = 2$

The path with the shortest length between two nodes (distance).

Path length

N_{ij},number of paths between any two nodes *i* and *j*:

Length n=1: If there is a link between ***i*** and ***j***, then $A_{ij}=1$ and $A_{ij}=0$ otherwise.

Length n=2: If there is a path of length two between ***i*** and ***j***, then $A_{ik}A_{kj}=1$, and $A_{ik}A_{kj}=0$ otherwise.
The number of paths of length 2:

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik}A_{kj} = [A^2]_{ij}$$

Length n: In general, if there is a path of length ***n*** between ***i*** and ***j***, then $A_{ik}\ldots A_{lj}=1$ and $A_{ik}\ldots A_{lj}=0$ otherwise.
The number of paths of length ***n*** between ***i*** and ***j*** is*

$$N_{ij}^{(n)} = [A^n]_{ij}$$

****holds for both directed and undirected networks.***

Path length

- d_{max} diameter- the maximum distance between any pairs of nodes
- $\langle d \rangle$ average path length - for directed graphs

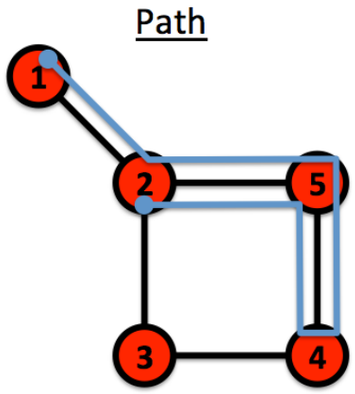
$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}$$

- where d_{ij} is the shortest distance between nodes i and j
- multiplicative is *(2 x max number of links)*
- distance between unconnected nodes is 0

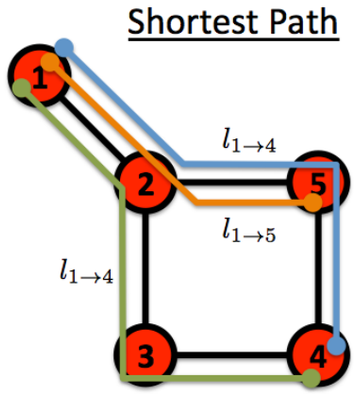
- $\langle d \rangle$ average path length - for un-directed graphs

$$\langle d \rangle = \frac{2}{N(N-1)} \sum_{i < j} d_{ij}$$

- since $d_{ij} = d_{ji}$
- multiplicative is *(max number of links)*



A sequence of nodes such that each node is connected to the next node along the path by a link.



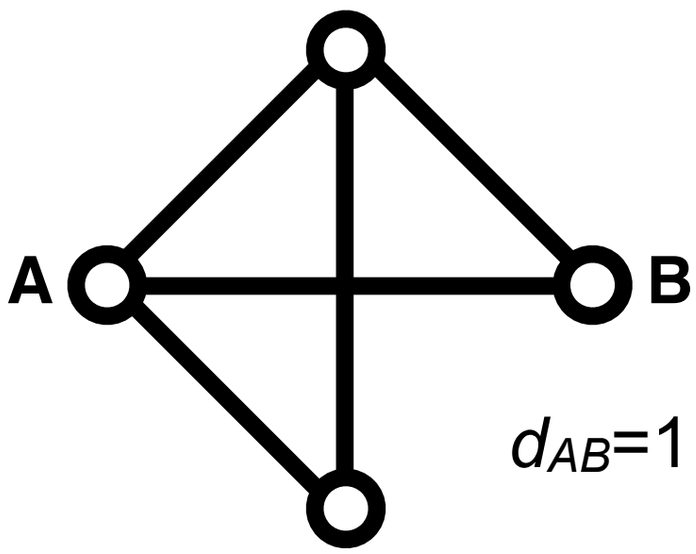
The path with the shortest length between two nodes (distance).

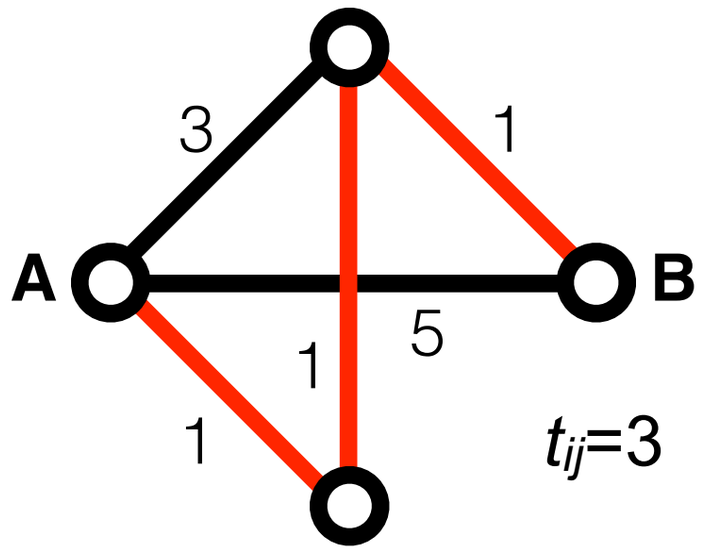
Weighted path length

length of a shortest path $P(i \to j)$ $\neq$ length of a weighted shortest path $P(i \to j)$

$$d_{ij} = \sum_{e_{mn} \in P(i \to j)} A_{mn} \qquad t_{ij} = \sum_{e_{mn} \in P(i \to j)} w_{mn}$$

Shortest path ≠ Weighted shortest path





Central quantities in network analysis

- Degree distribution: $P(k)$
- Clustering coefficient: C
- Average path length: $<d>$

Advanced global network characteristics

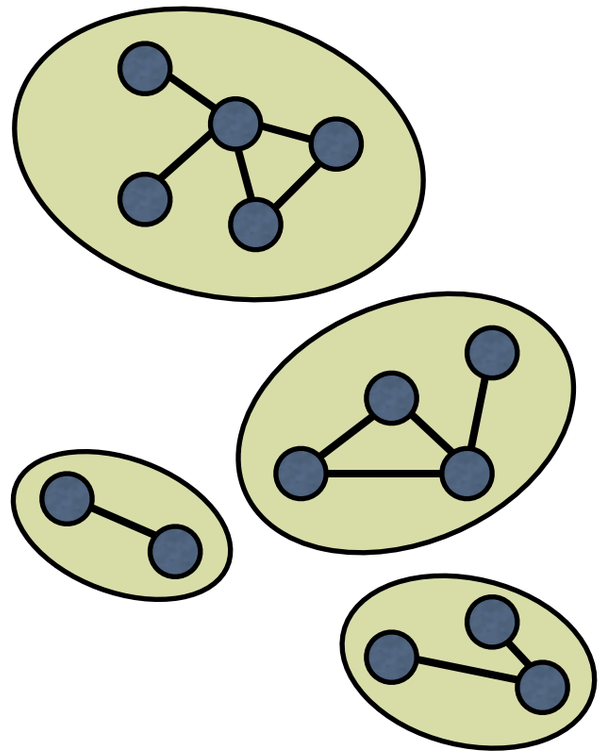
Network density

$$\rho = \frac{m}{\binom{N}{2}} = \frac{2m}{N(N-1)}$$

- Ratio of the numbers of existing edges m and possible number of edges $(N(N-1))/2$ in a network of size N
- If the network is fully connected: $\rho = 1$
- Sparse networks: if $\rho \to 0$ (or tends to a constant) as network size $N \to \infty$
- Real networks are usually sparse

Connectivity and components

- A connected component is a subset of vertices with at least one path connecting each of them
- A network may consist of a single connected component (a connected network) or several of those
- Distances between nodes in disjoint components are not defined (infinite)
- Bridge: if we remove it, the graph becomes disconnected.
- The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero



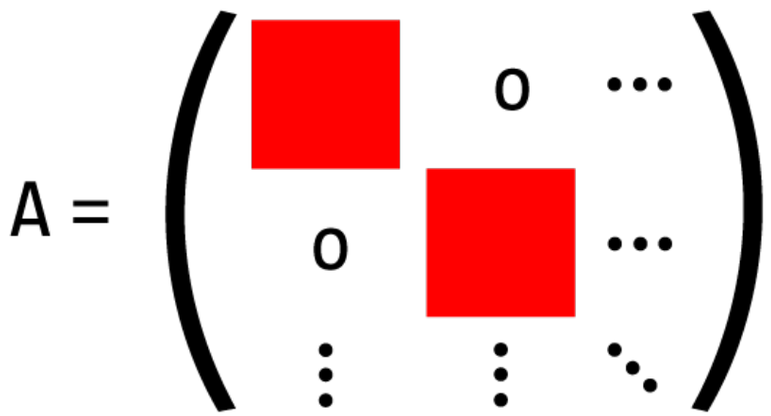


Figure after Newman, 2010

Connectivity and components - directed networks

- Strongly connected component (SCC): has a path from each node to every other node in the component
- Weakly connected component (WCC): it is connected if we disregard the directions
- In-component: nodes that can reach the SCC
- Out-component: nodes that can be reached from SCC

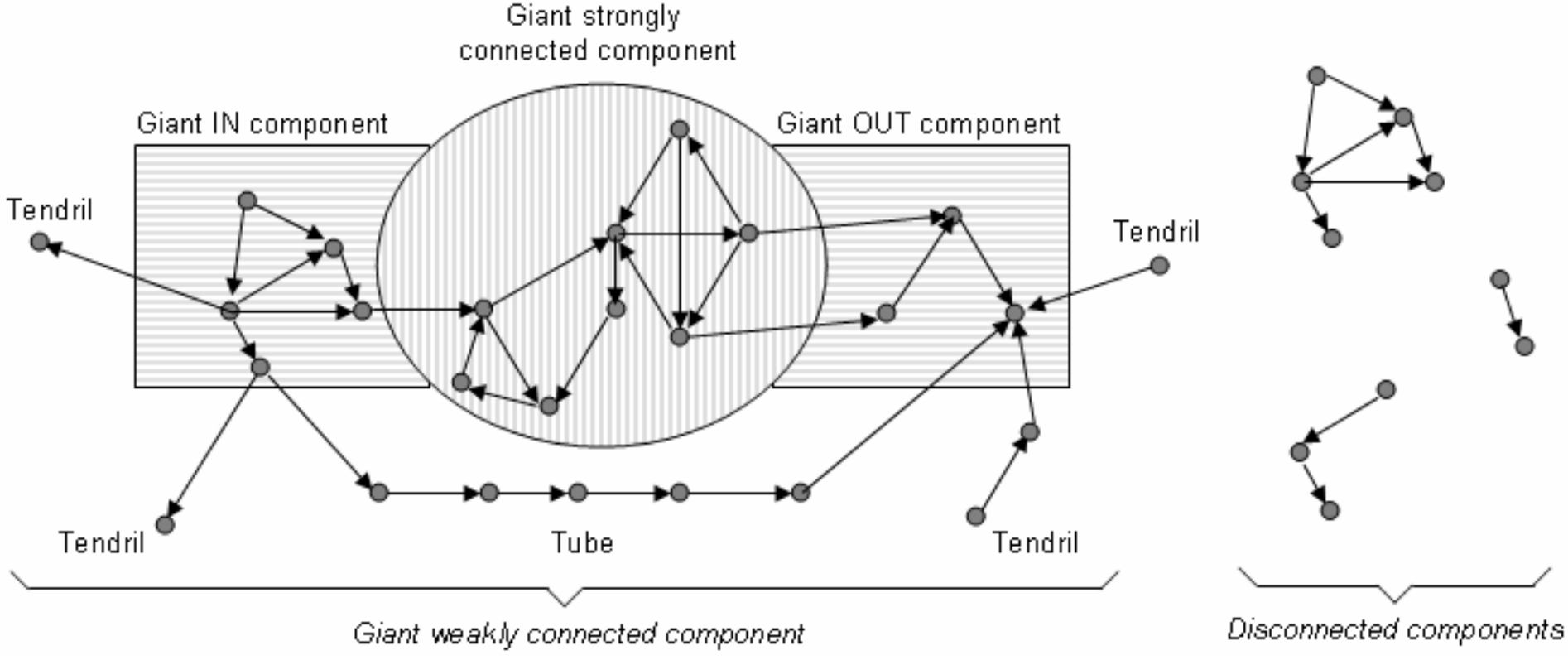


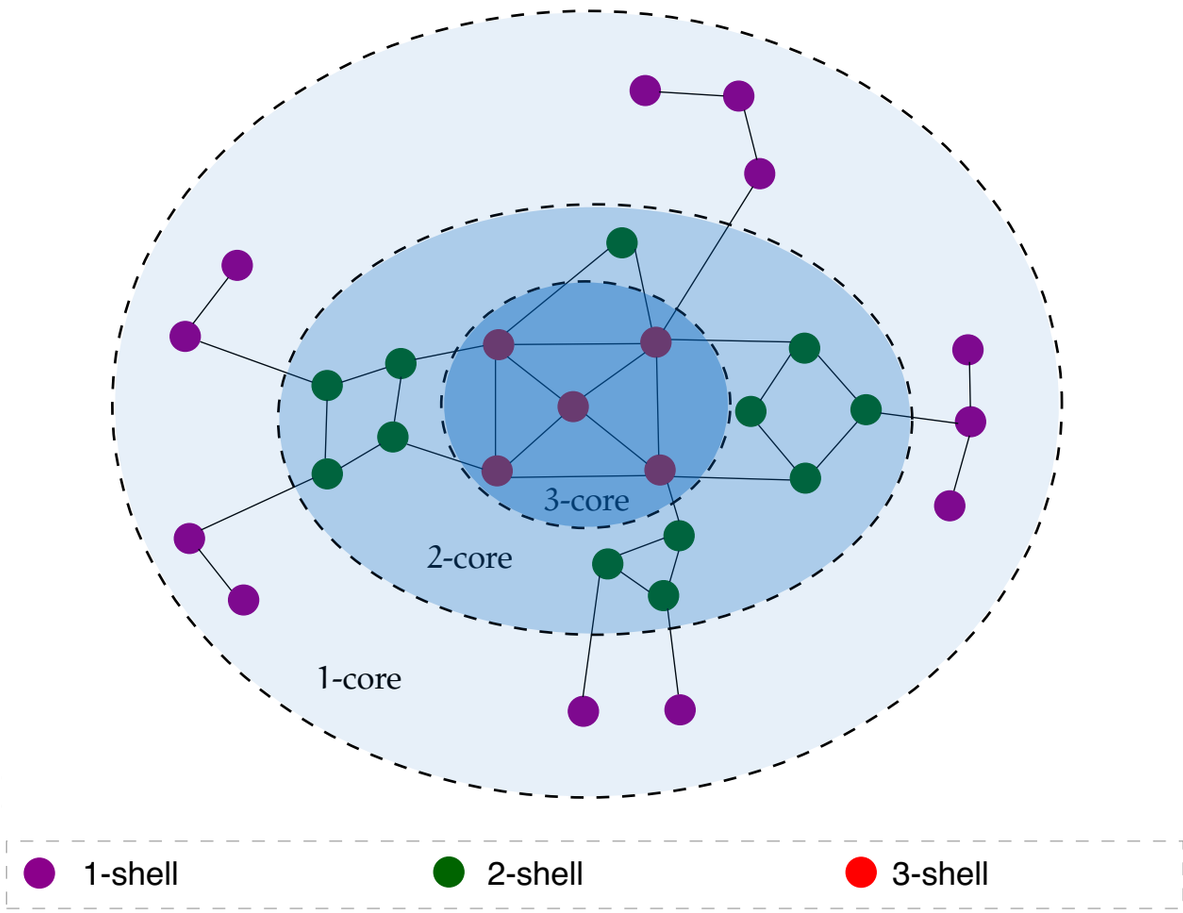
Figure from Broder et. al. (2000)

k-core decomposition

Goal: To identify dense cores of high degree nodes in a network

Given graph $G = (V, E)$

Definition: A subgraph $H = (C,E|C)$ induced by the set $C \subseteq V$ is a **k-core or a core of order k** iff $\forall v \in C : degree(H(v)) \geq k$, and H is the maximum subgraph with this property.

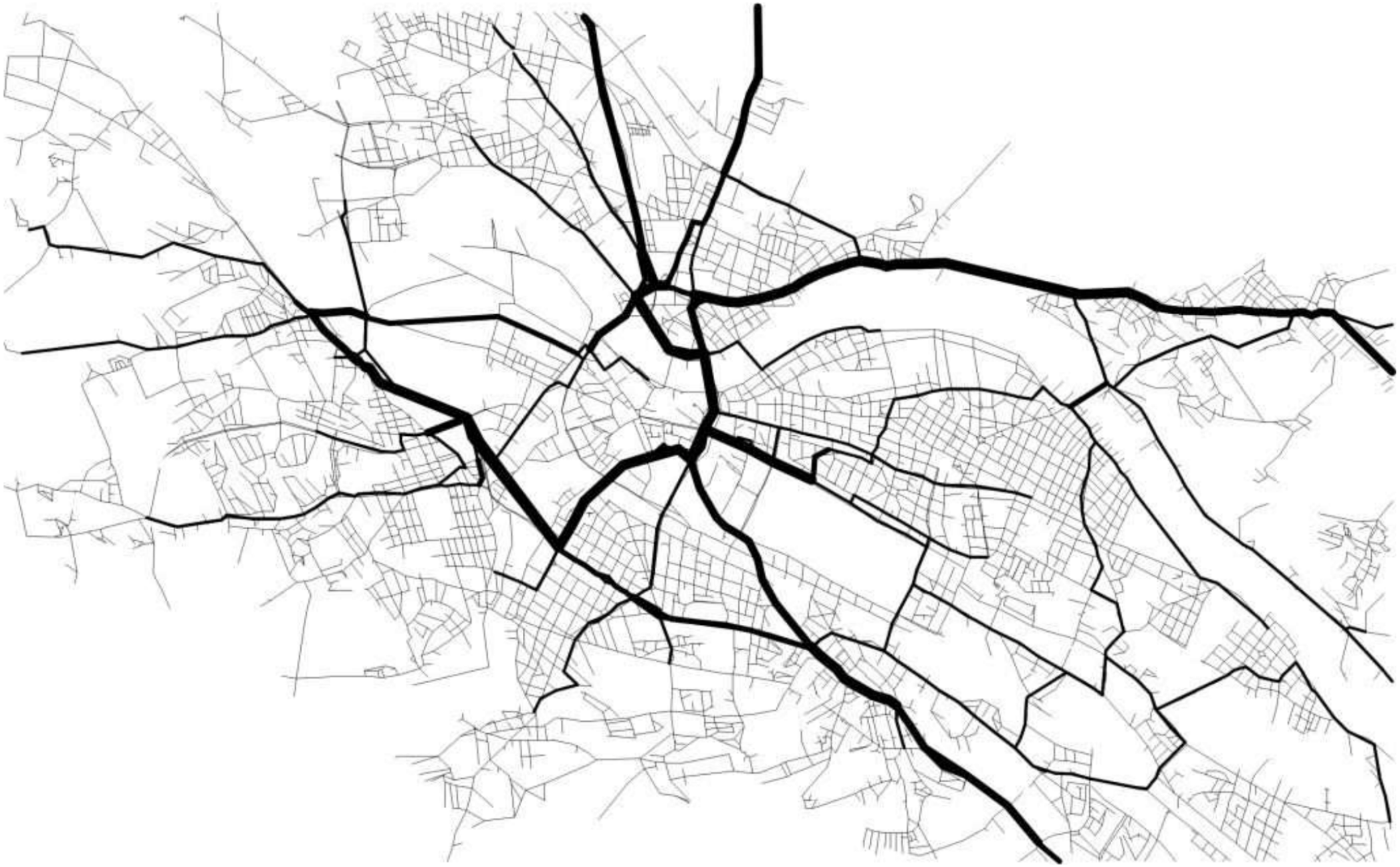


- A k-core of G can be obtained by recursively removing all the vertices of degree less than k, until all vertices in the remaining graph have at least degree k.

Definition: A vertex i has **coreness c** if it belongs to the c-$core$ but not to $(c + 1)$-$core$.

Definition: A **c-shell** is composed by all the vertices whose coreness is c. The k-core is thus the union of all shells with $c \geq k$.

Betweenness Centrality



from S. Lammer, B. Gehlsen, and D. Helbing. Scaling laws
in the spatial structure of urban road networks. Physica A, **363**:89, 2006.

Betweenness Centrality

Assumption: important vertices are bridges over which information flows

Practically: if information spreads via shortest paths, nodes are important which are along many shortest paths

Notation: $\sigma_{jk}(i) =$ number of geodesic path from j to k via i: $j \to \ldots \to i \to \ldots \to k$

$\sigma_{jk} =$ number of geodesic path from j to k: $j \to \ldots \to k$

Definition:

$$C_b(i) = \sum_{j \neq k} \frac{\#\{\text{geodesic path: } j \to \ldots \to i \to \ldots \to k\}}{\#\{\text{geodesic path: } j \to \ldots \to k\}} = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

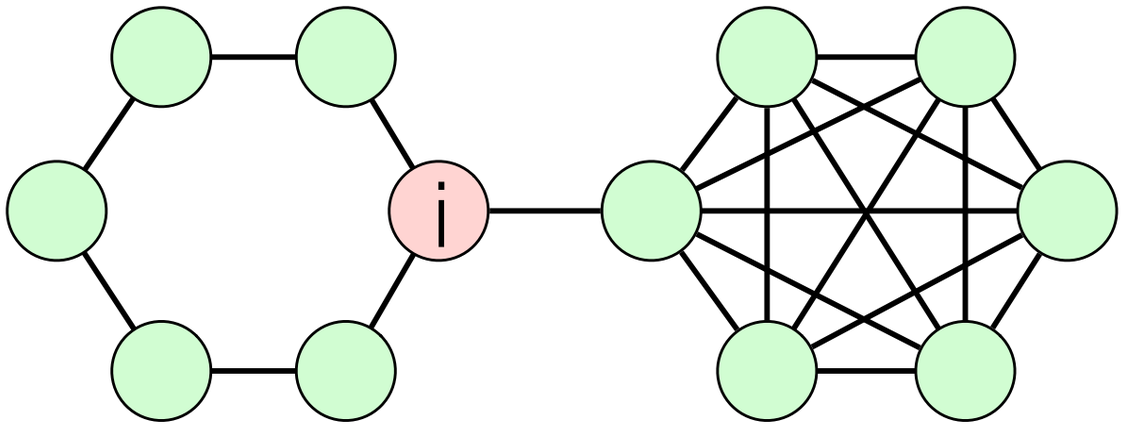
Normalised definition:

$$C_b(i) = \frac{1}{N^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \qquad \text{where} \qquad C_b \in [0,1]$$

Total number of ordered vertex pairs

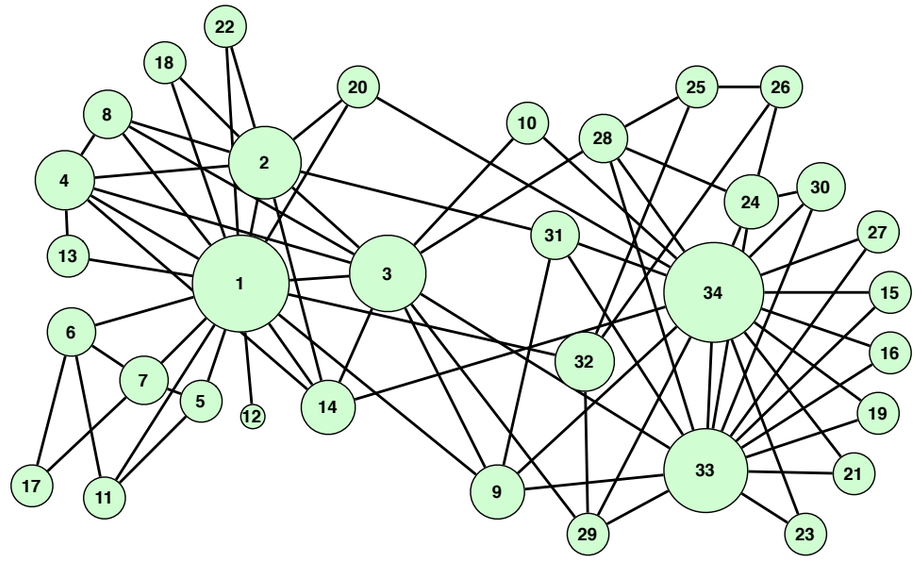
Betweenness Centrality

$$C_b(i) = \frac{1}{N^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad \text{where} \quad C_b \in [0,1]$$

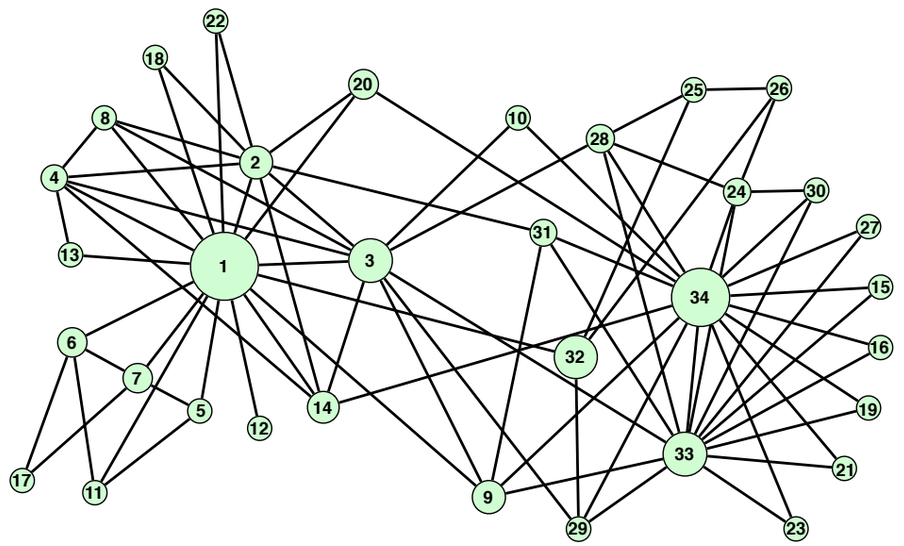


$$C_b(i) = \frac{78}{144}$$

Zachary's karate club network



degree



betweenness

Exact computation:

Floyd-Warshall: *O(n³) time complexity*
O(n²) space complexity

Approximate computation

Dijskstra: *O(n(m+n log n)) time complexity*

Community detection methods

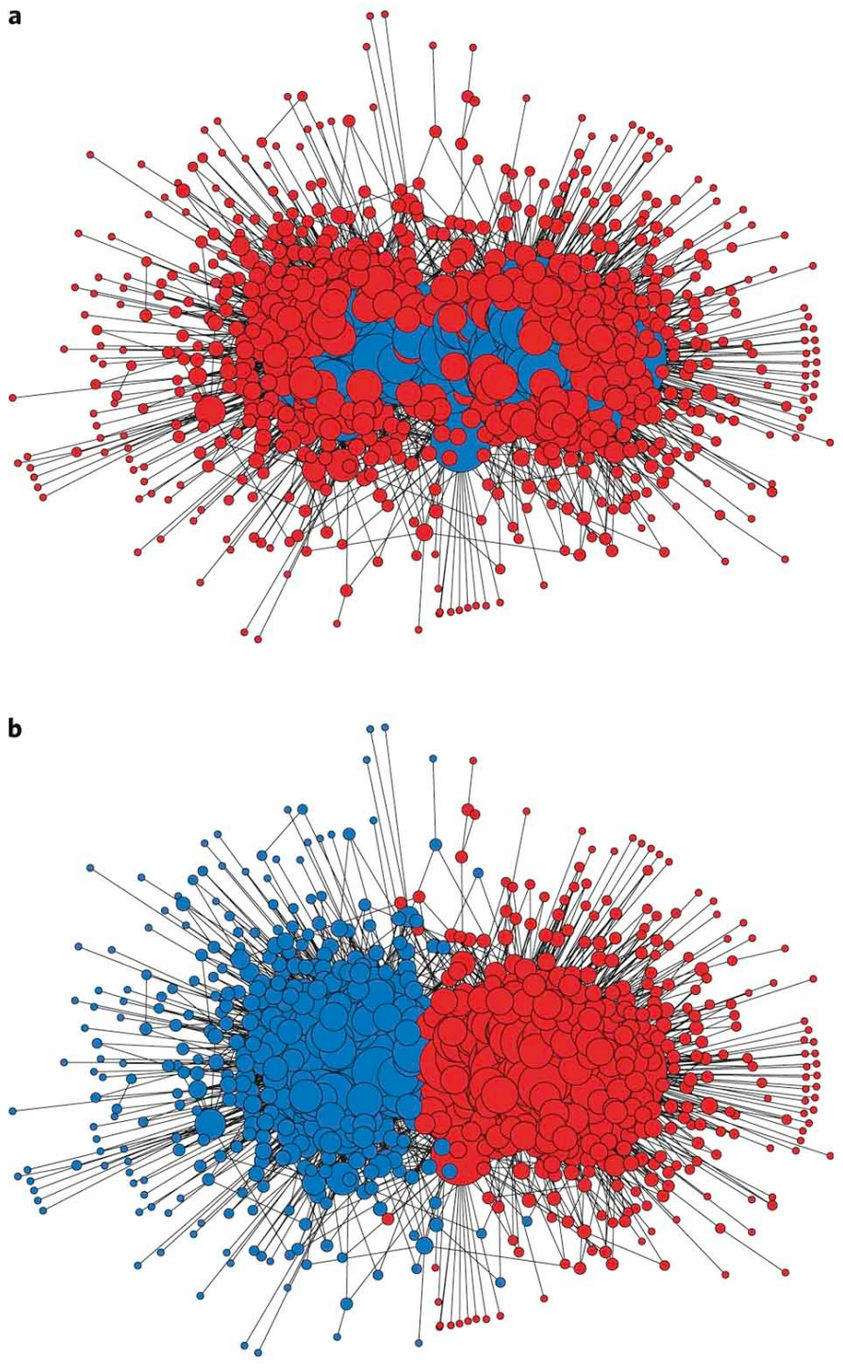
Complex networks - Mesoscopic view

Mesoscopic structures

- Motifs
- Partitions
- Modules
- Communities

Questions:

- Methods to find
- Measures to quantify
- Structure and frequency
- Applications
  - Visualization
  - Recommendation systems
  - Unknown functionality
  - …

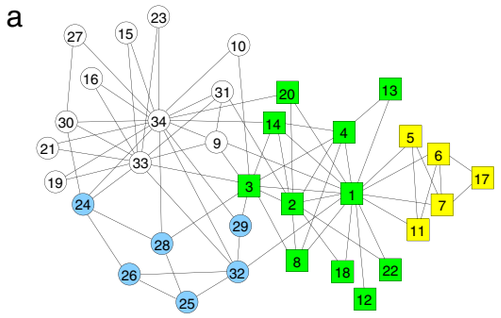


Newman, Nature Physics (2012)

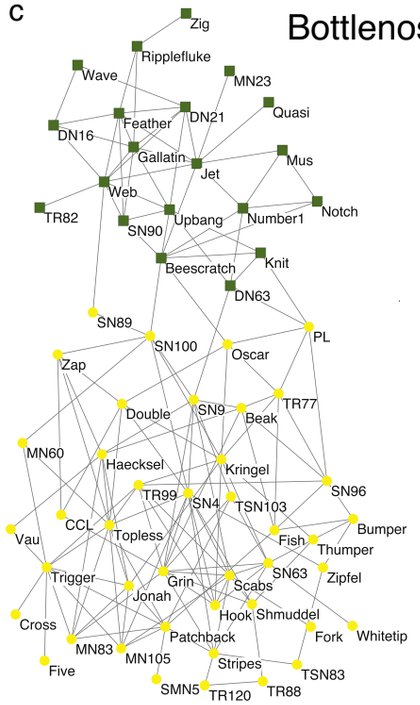
Communities in real world

Communities in social networks

Zachary's karate club

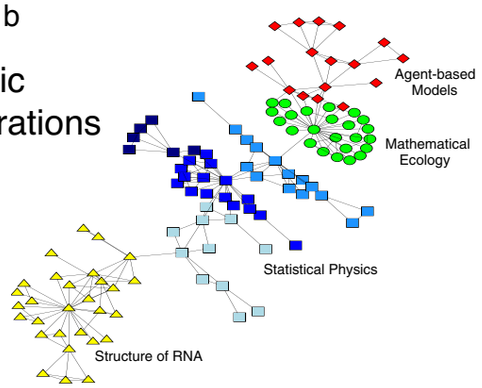


Bottlenose dolphins



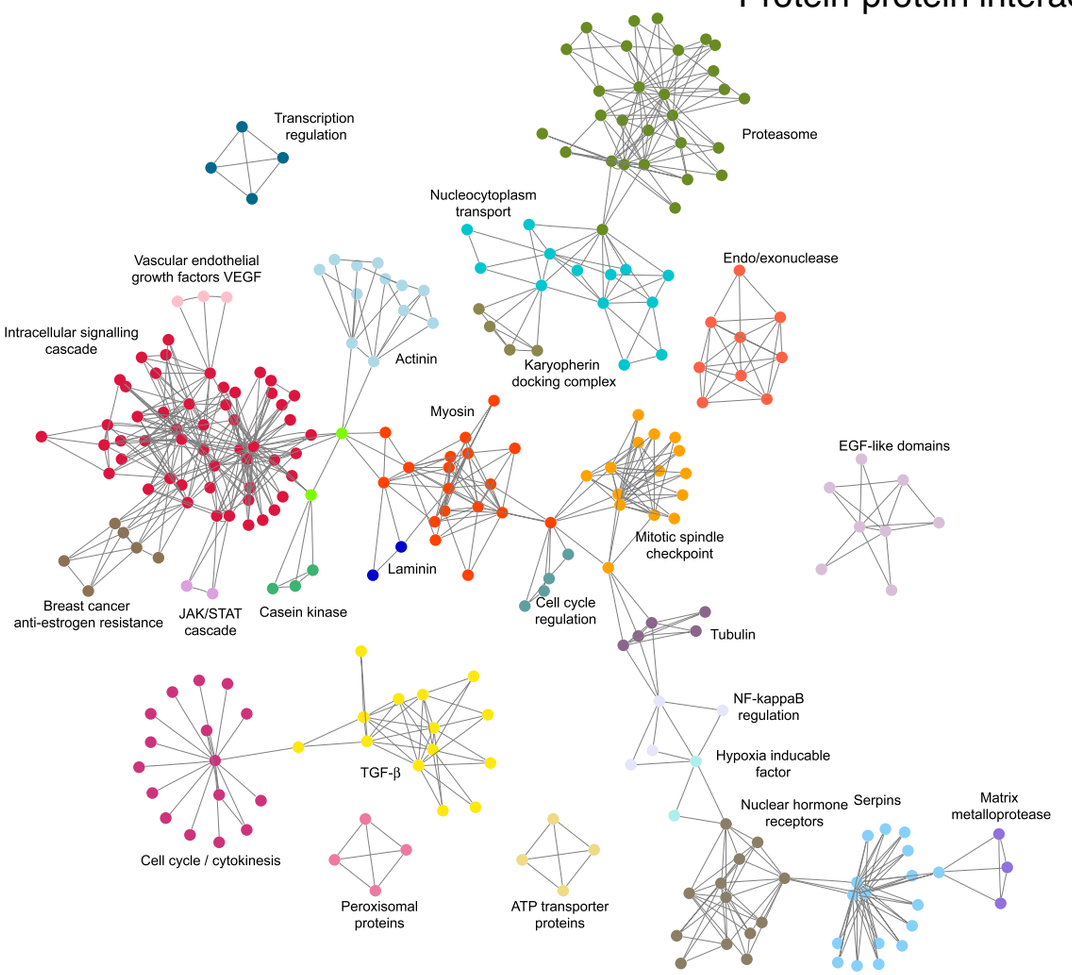
b

Scientific collaborations



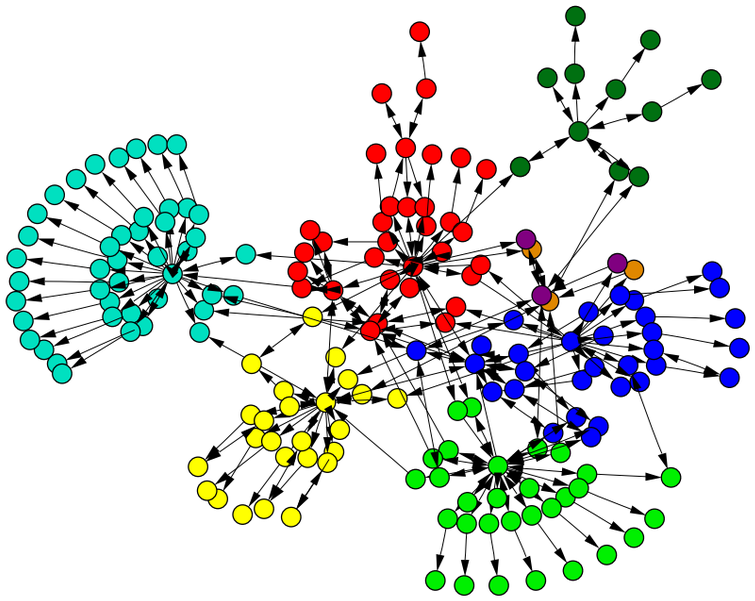
Communities in biological networks

Protein-protein interactions



Communities in information networks

www



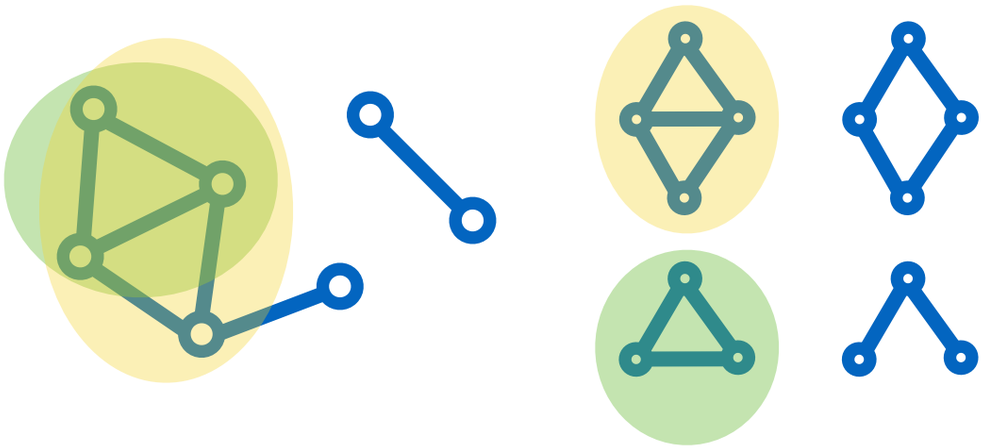
etc.

Fortunato (2010)

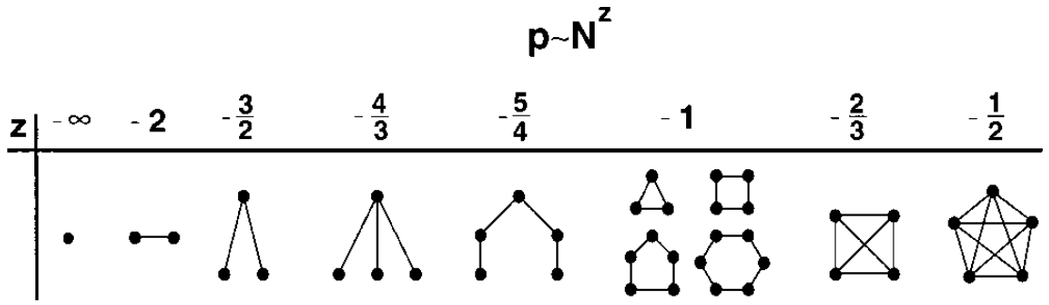
Subgraphs and motifs

Subgraphs

- $G'=(V',E')$ is a subgraphs of $G=(V,E)$ if $G' \subseteq G$ and $E' \subseteq E$



- In random graphs the appearance of different subgraphs is a function of the system size and the p probability

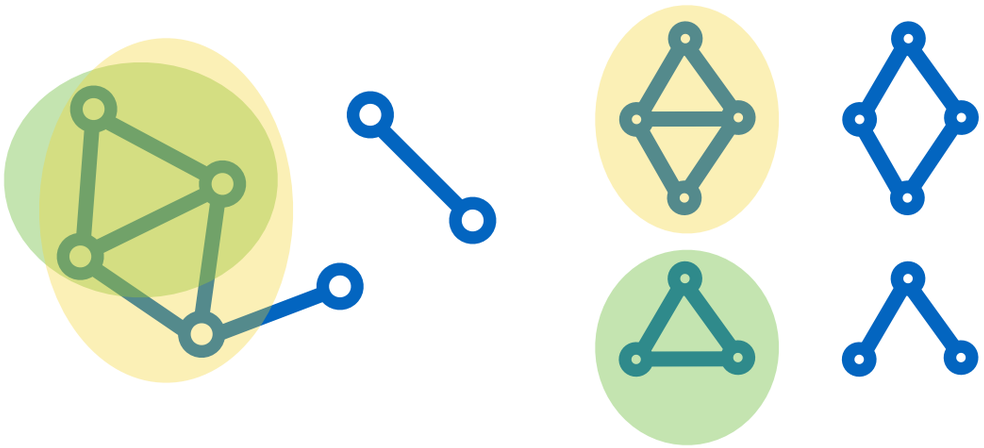


Albert et al., (2002)

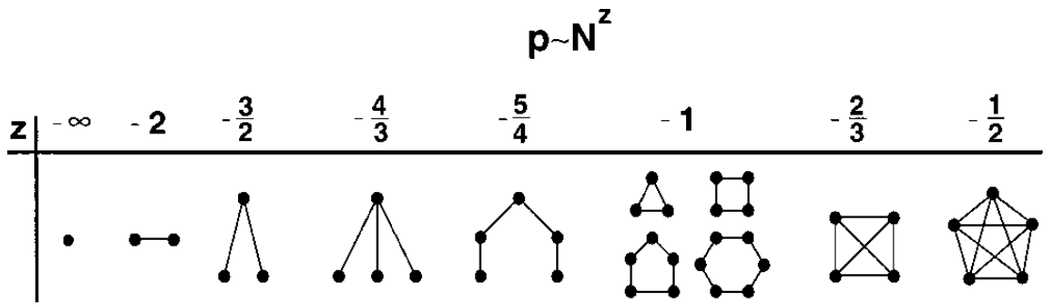
Subgraphs and motifs

Subgraphs

- $G'=(V',E')$ is a subgraphs of $G=(V,E)$ if $G' \subseteq G$ and $E' \subseteq E$



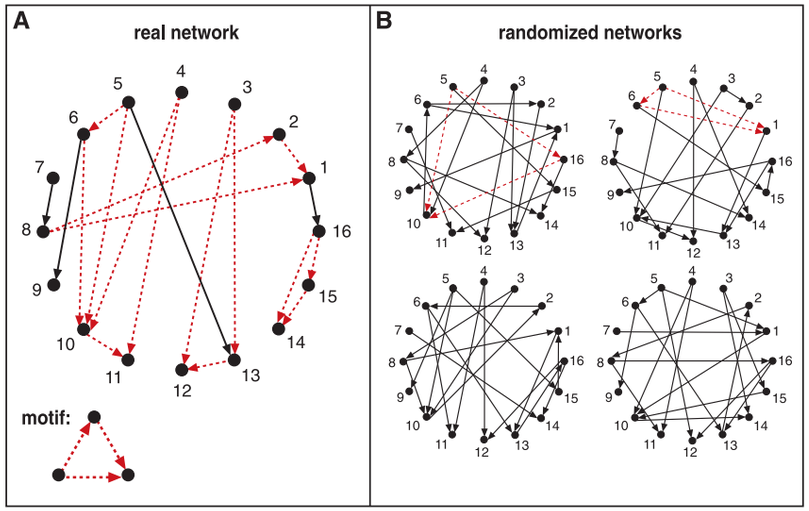
- In random graphs the appearance of different subgraphs is a function of the system size and the p probability



Albert et al., (2002)

Motifs

- Sub-graphs that appear with a significantly higher frequency in the real network than in the randomised version of the studied network



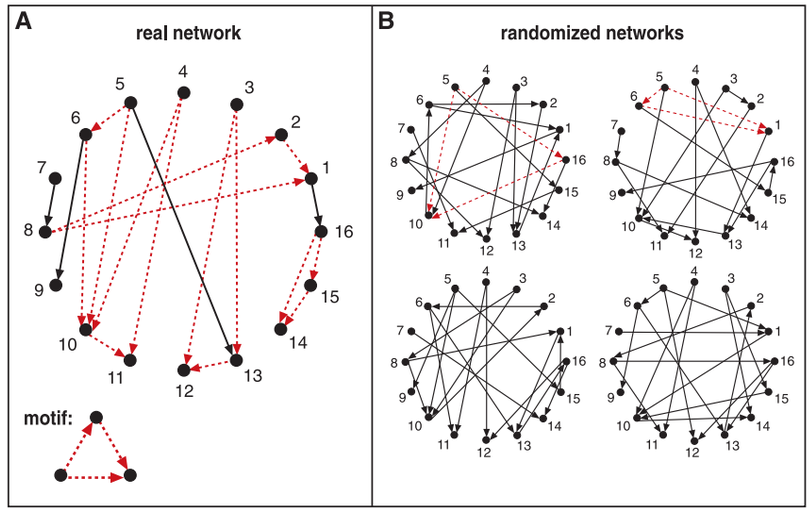
R. Milo et al., Science 298, 824 (2002)

- Randomised networks: Ensemble of maximally random networks preserving the degree distribution of the original network
- Algorithms: mfinder, FPF, ESU, …

Subgraphs and motifs

Z-score

1. Count subgraphs in the original network
2. Repeat:
   (i) rewire original network with configuration model
   (ii) count subgraphs in the rewired network
3. Compare original count to the reference ensemble as:



number of times the subgraph M occurs in the empirical network

average number of times the subgraph M occurs in the randomised reference network

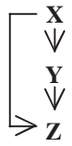
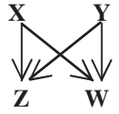
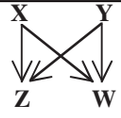
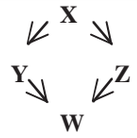
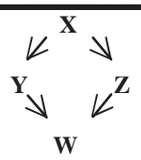
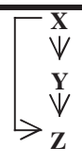
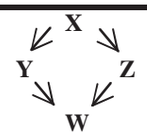
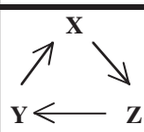
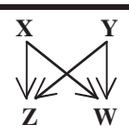
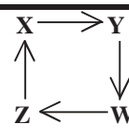
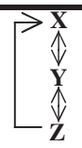
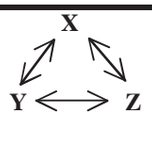
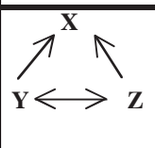
$$Z_M = \frac{n_M - \langle n_M^{rand} \rangle}{\sigma_{n_M}^{rand}}$$

standard deviation of n_M in the reference system

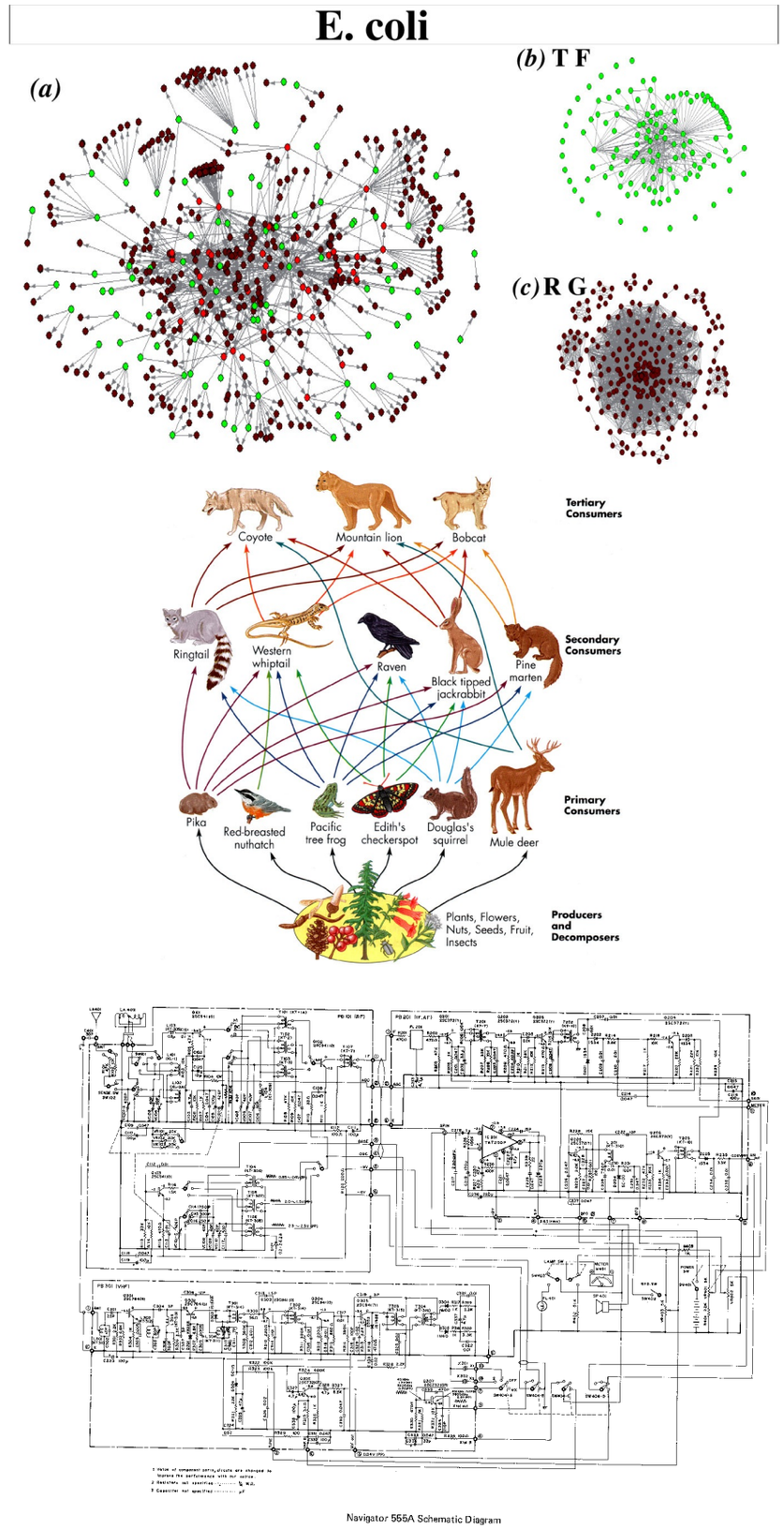
- Subgraph frequency in the configuration model depends on the system size
- To compare different networks Z-scores need to be normalised as:

$$\hat{Z}_i = \frac{z_i}{\sqrt{\sum_i z_i^2}}$$

Motifs in real networks

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

R. Milo et al., Science 298, 824 (2002)



Communities - basics

General conditions

- Community structure inferred only from structural informations, relations with actual groups is unclear
- The number of m edges of the graph is of the order of the n number of vertices

$$m \sim n$$

otherwise the problem becomes similar to data clustering

Important features

- Computational complexity: how a method performs on very large networks
- Possible network structure: bipartite, large, dense, temporal, …
- Possible community features: disjoint, overlapping, temporal, …
- Possible validation: other methods, benchmarks, meta informations

Limitations

- Communities are usually implicitly defined by the specific algorithm adopted, without an explicit definition!
- The practical definition may depend on the specific system/application

Modularity based algorithms

Modularity

Newman & Girvan, 2004

Principle: Random graphs have no community structure

Method: comparing the edge density in each cluster with the edge density of the cluster in a randomised version of the graph

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

Null model:

- Principally it is arbitrary
- ER (Bernoulli) random graph
- Random graph preserving the original node degree sequence (generated by a Configuration Model process)

Modularity

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

Take a network $G=(V,E)$ with n nodes, m links, and A_{ij} adjacency matrix

Assume C communities:

- assume $i \in V$ is in community C_i and $j \in V$ is in community C_j
- membership: $\delta(C_i, C_j) = 1$ if $C_i=C_j$, and 0 otherwise

Modularity: fraction of edges fall within communities, minus the expected fraction of such edges in a reference model

Expected fraction of edges in a reference model (using a configuration network model):

- It keeps the degrees of nodes unchanged
- It cuts each link in two stubs and rewire links randomly

Total number of stubs: $l_b = \sum_i k_i = 2m$

Expected number of connected edges between i and j: $\frac{k_i k_j}{l_b} = \frac{k_i k_j}{2m}$

Modularity

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

Adjacency matrix of the original graph

Expected number of edges between nodes with degree k_i and k_j in the configuration model network

Total number of stubs, number of possible rewiring of a link ($n \to \infty$)

δ function: 1 if both nodes are in the same module $C_i = C_j$, 0 otherwise

Modularity

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

Features:

- Q < 1
- Q = 0 for a partition where the whole graph is a single cluster
- -1/2 ≤ Q: negative for multipartite structures
- Modularity depends on the network size: partitions of different graphs cannot be compared to each other
- High modularity value does not necessarily assign good partitioning - random graphs can have high partition as well…

Partitioning

Divide the graph in n parts, such that the number of links between them (cut size) is minimal

Problems:

- Number of partitions must be specified in advance
- Size of clusters must be specified in advance

Traditional methods:

- Graph bi-sectioning
- Kerninghan-Lin algrorithm
- Spectral partitioning
- Partitional clustering
- K-means clustering
- …

One would like methods that can predict the number and the size of the partition and indicate a subset of “good” partitions

Partitioning

Graph partition: division of a graph as a union of non-overlapping and non-empty subgraphs

- Number of possible partitions of a graph with n vertices into k clusters is given by the S(n,k) Stirling number of the second kind

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$$

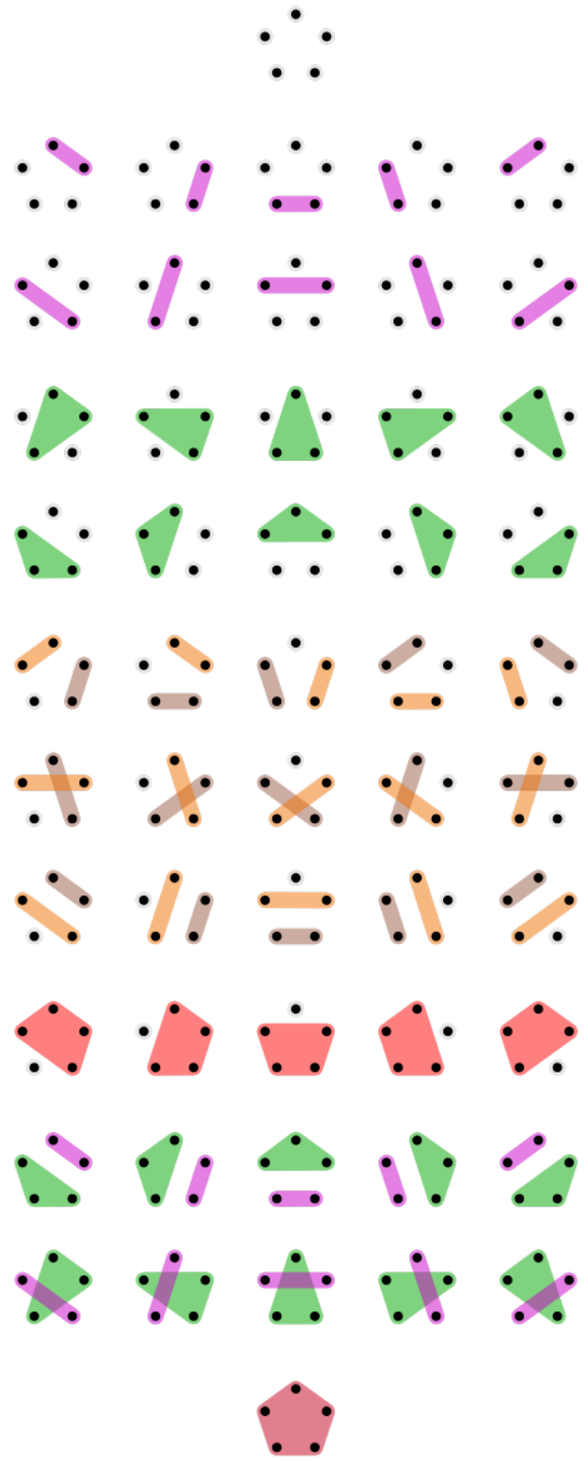
- Total number of possible partitions: Bell-number

$$B_n = \sum_{k=0}^{n} S(n,k)$$

- In the large n limit:

$$B_n \sim \frac{1}{\sqrt{n}}[e^{W(n)}]^{n+1/2}e^{e^{W(n)}-n-1}$$, $W(n)$: Lambert function

- It is a double exponential ➡ very large number



The 52 partitions of a set with 5 elements

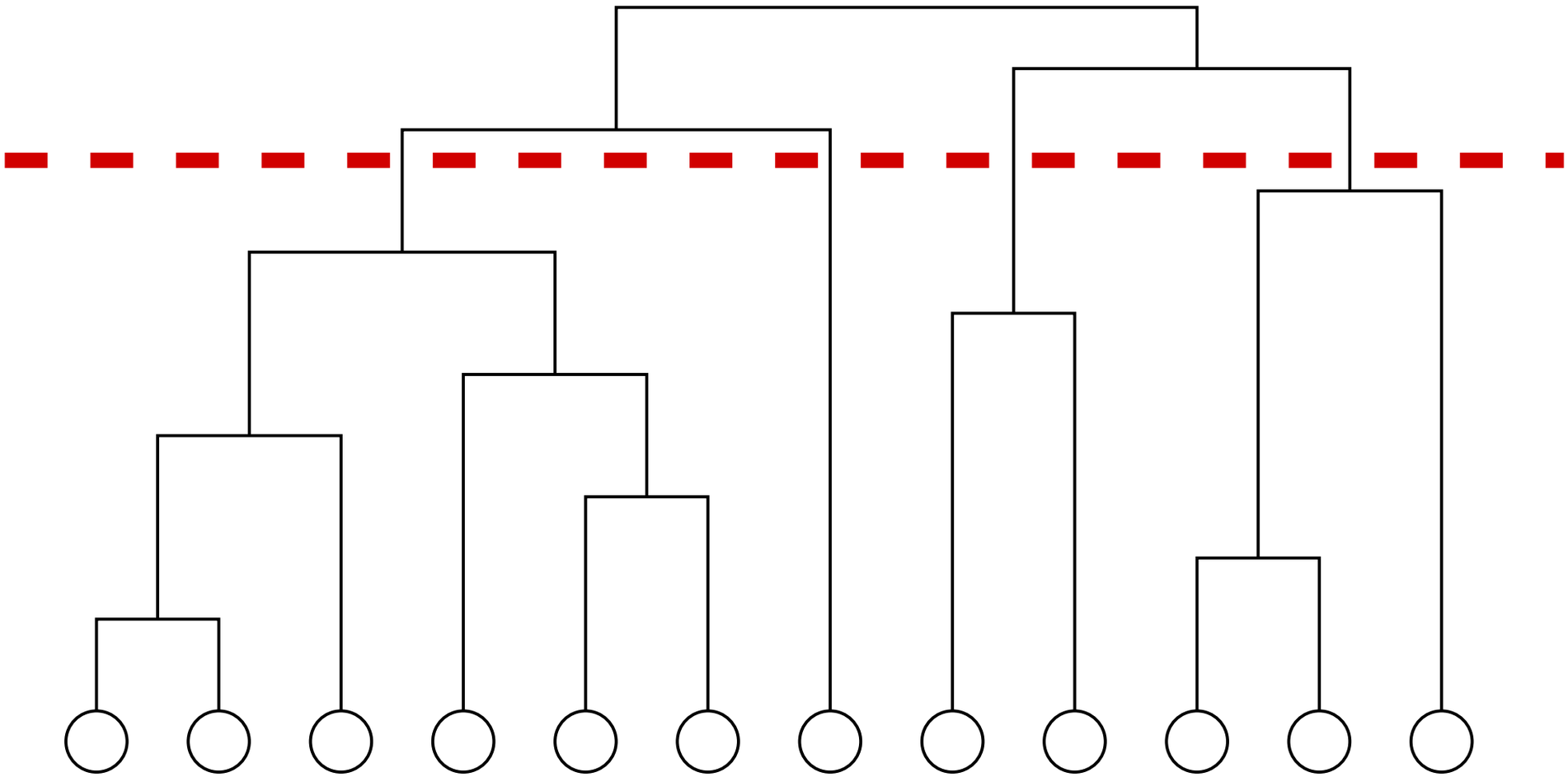
Hierarchical clustering

- Very common in social network analysis
- Two methods: agglomerative (bottom-up approach), divisive (top-down approach)

General algorithm:

1. A criterion is introduced to compare nodes based on their similarity
2. A similarity matrix X_{ij} is constructed: the similarity of nodes i and j is X_{ij}
3. (Agglomerative) Starting from the individual nodes, larger groups are built by joining groups of nodes based on their similarity
4. (Divisive) Starting from the graph as a single cluster, separate the most dissimilar parts, etc.

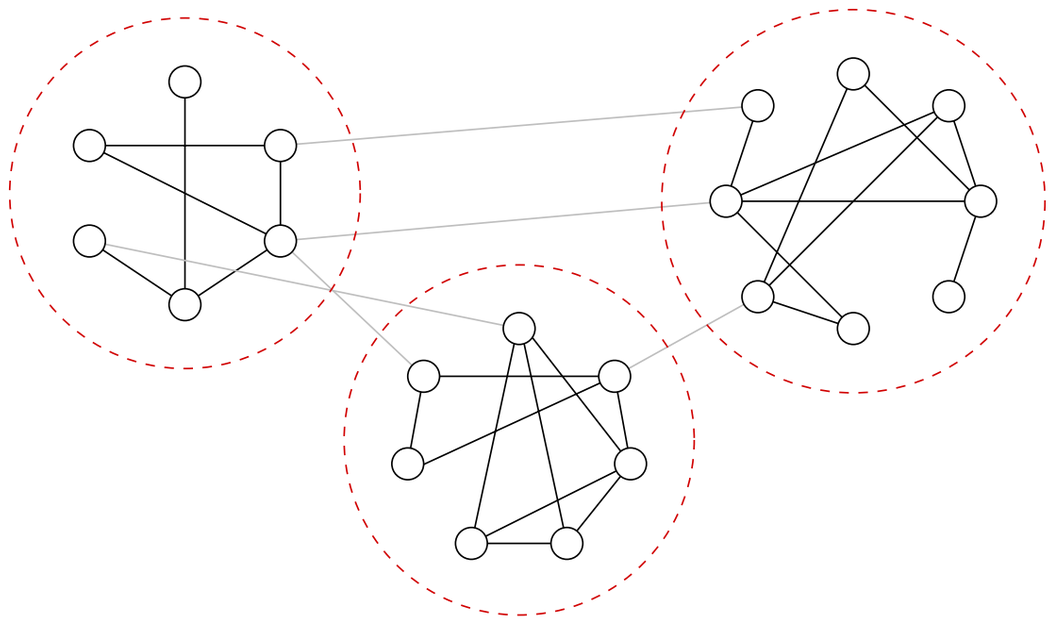
Hierarchical clustering - dendogram



Newman (2003)

Girvan-Newman method

- Divisive method: one removes the links that connect the clusters, until the clusters are isolated
- To identify inter-community links it uses edge betweenness measure
- If there are more geodesic paths between the same pair of vertices crossing the edge one divides the contribution of each path by their multiplicity
- Computable with algorithms based on breadth-first-search algorithm, with complexity $O(mn)$ (Brandes, 2001)



Newman (2003)

Girvan-Newman method

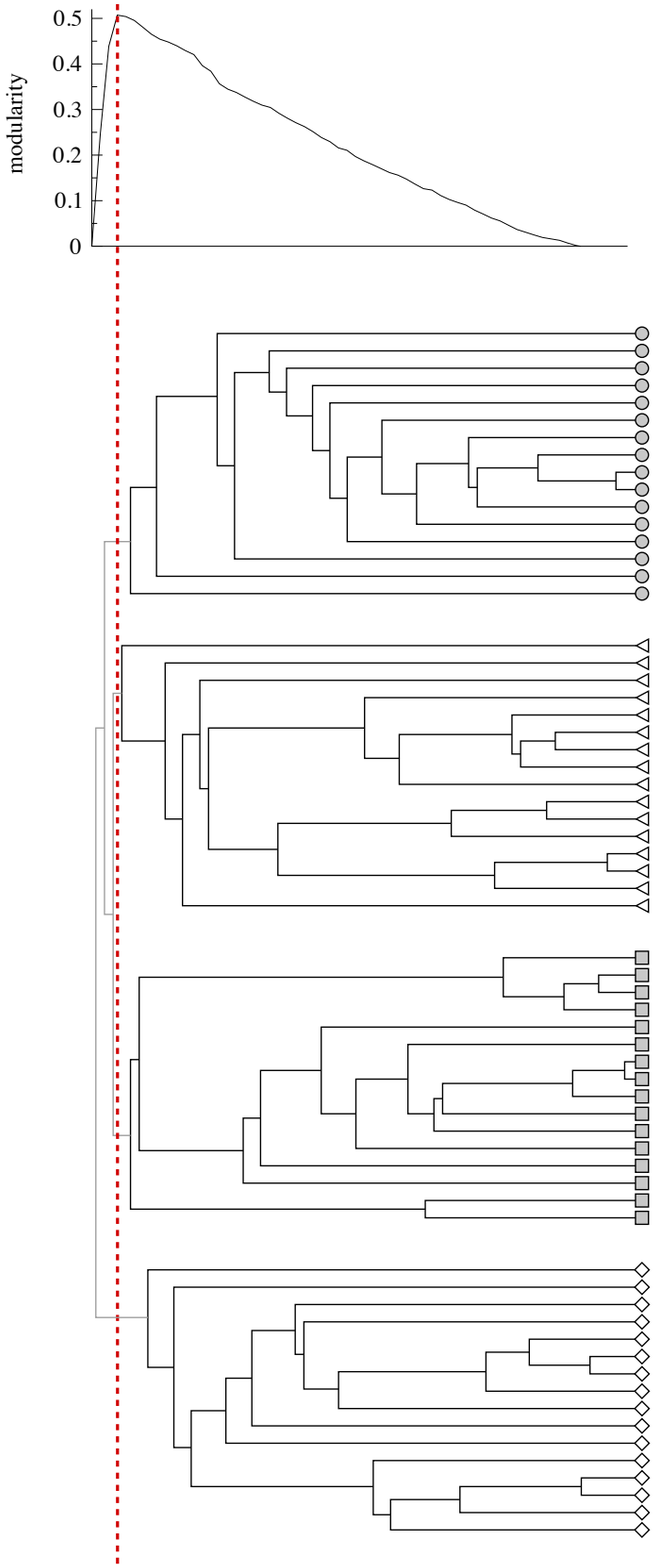
Algorithm

1. Calculate the betweenness of all edge
2. Remove the one with the highest betweenness
3. Recalculate the betweenness of the remaining edges
4. Repeat from step 2

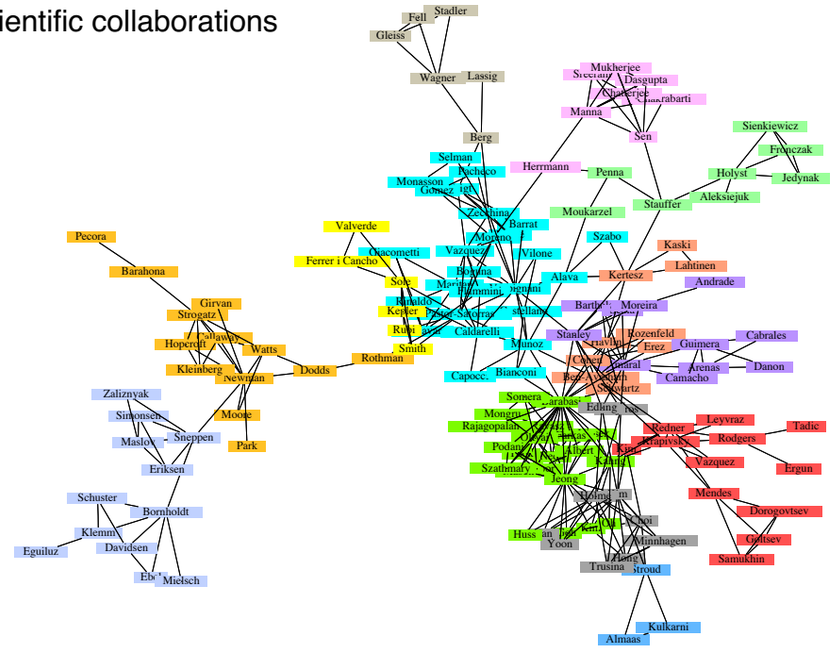
Features

- Complexity: $O(m^2n)$ (or $O(n^3)$ on sparse graphs and may be lowered by calculating step 3 only for a sample of node pairs)
- It delivers a hierarchy of partitions! Which one is the best?
- Girvan&Newman (2004): the best partition is the one corresponding to the highest modularity

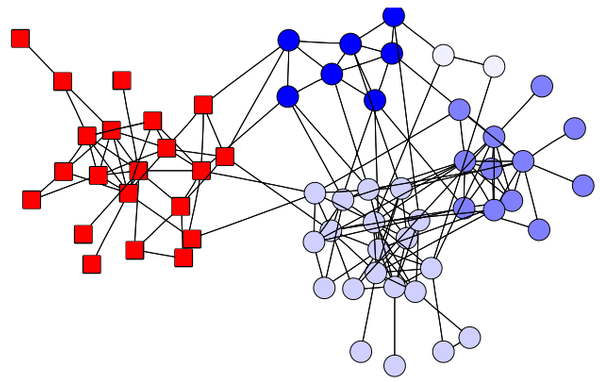
Girvan-Newman method



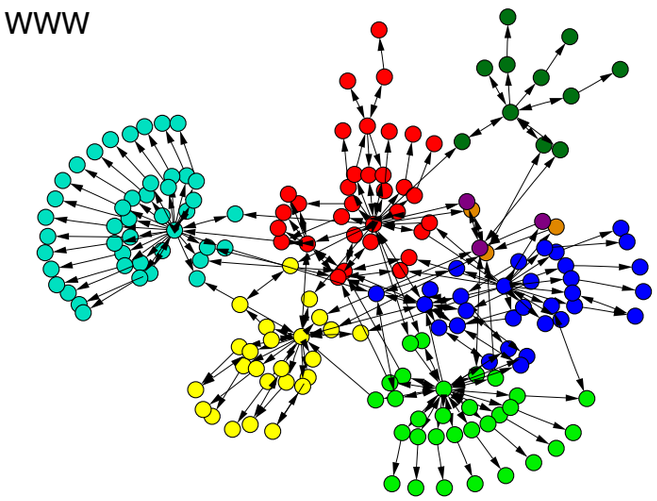
Scientific collaborations



Dolphins



WWW



Newman 2003

Other methods

Agglomerative approach: Louvain method

Information theoretical approach: Infomap method

Statistical approach: Stochastic block models

Structural approach: Clique percolation

Representation learning approach: node2vec

… and about 1000 more algorithms