Maximum Entropy Network Ensembles

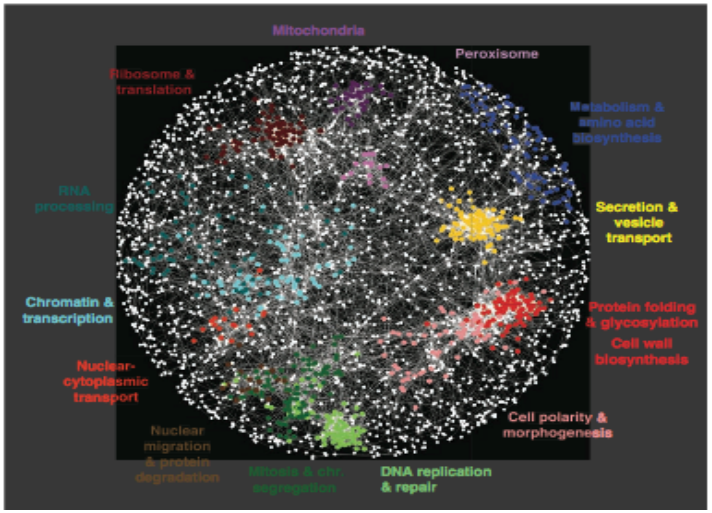
LTCC Course
Lesson 1

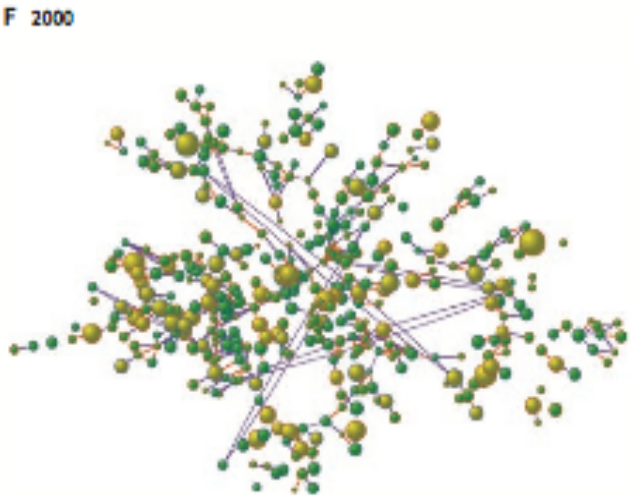
Ginestra Bianconi

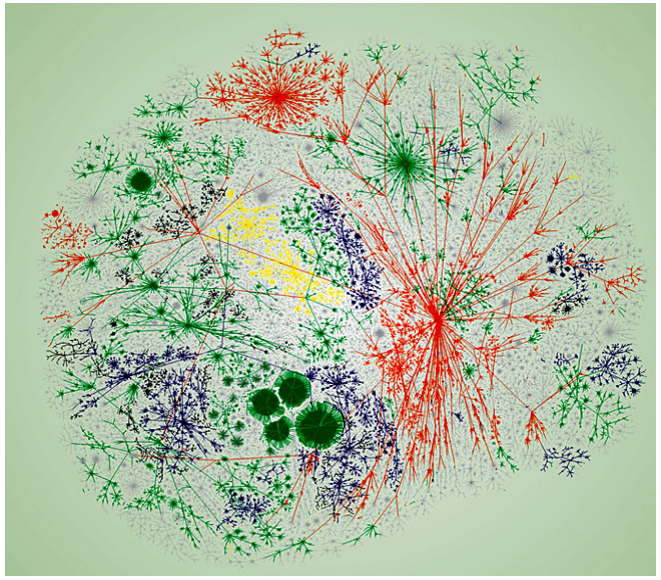
School of Mathematical Sciences Queen Mary University of London

Queen Mary
University of London

Complex networks







describe

the interactions between the elements of large complex

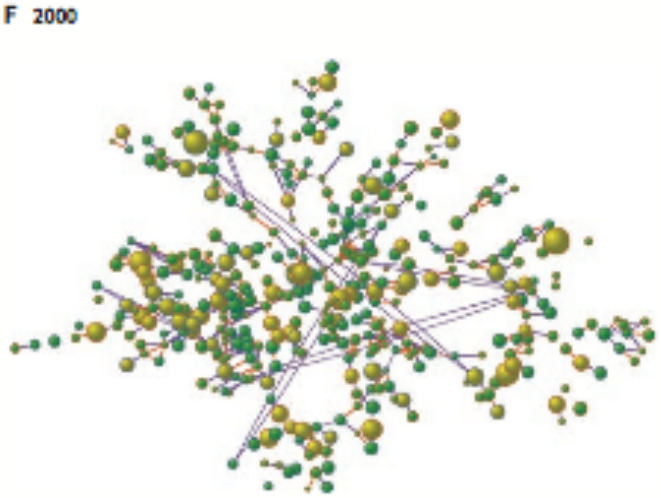
Biological, Social and Technological systems.

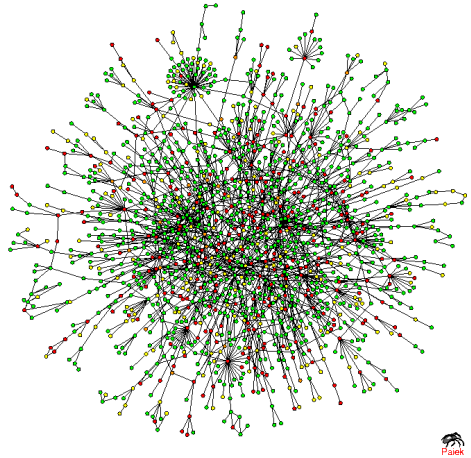
Networks are everywhere

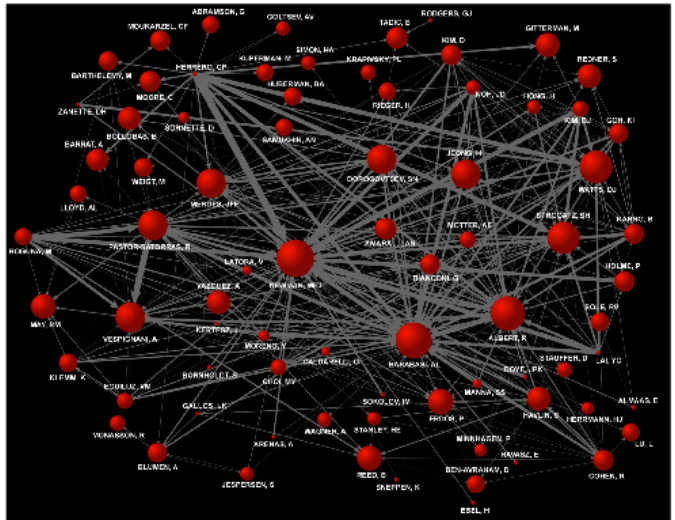
SOCIAL NETWORKS

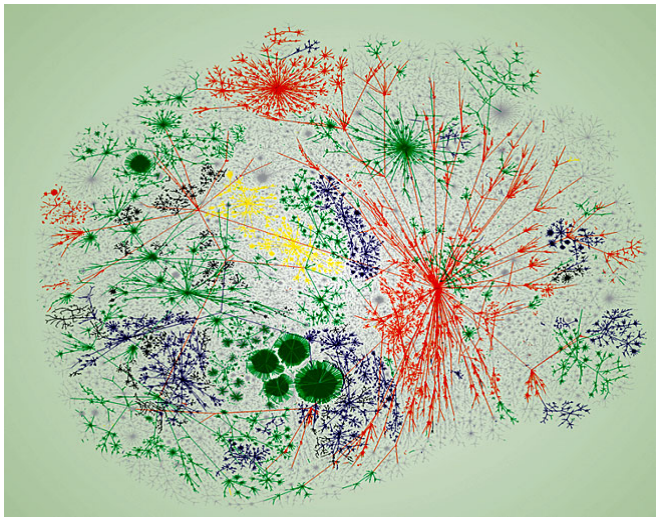
COMMUNICATION NETWORKS

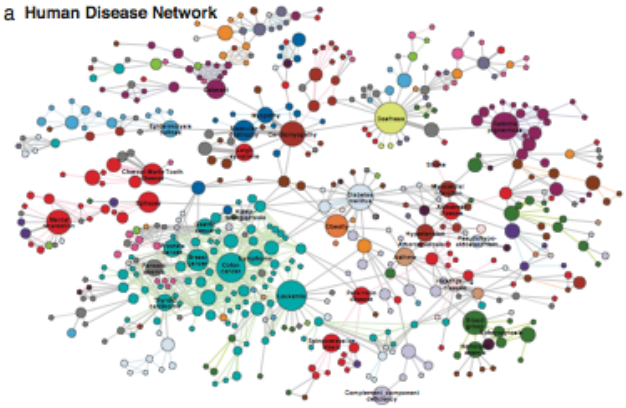
BIOLOGICAL NETWORKS







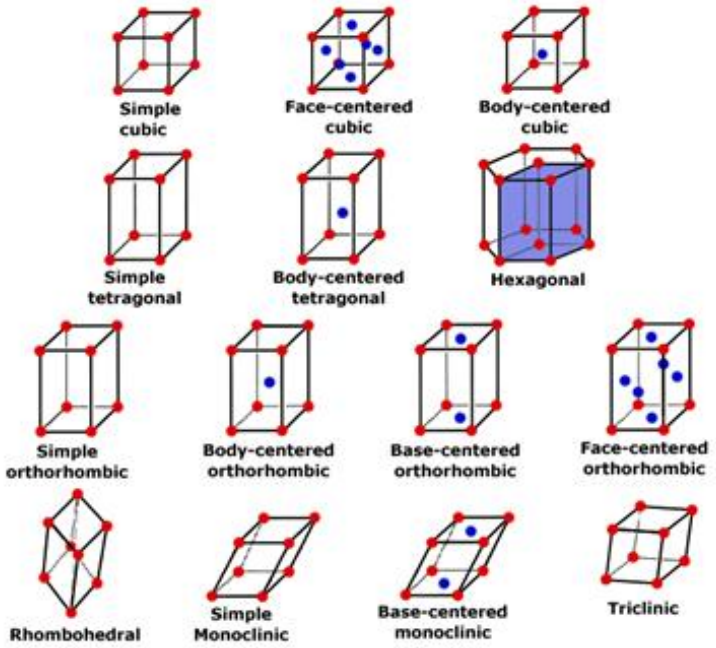




Complexity

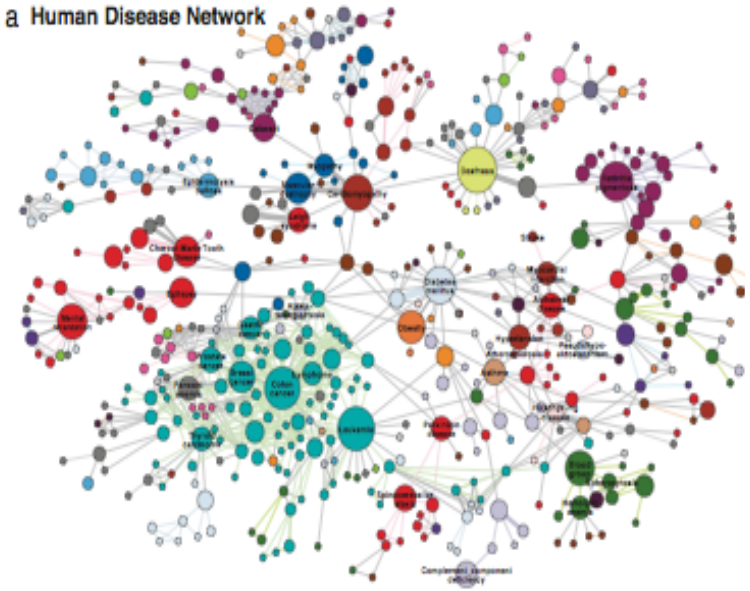
Between randomness and order

LATTICES



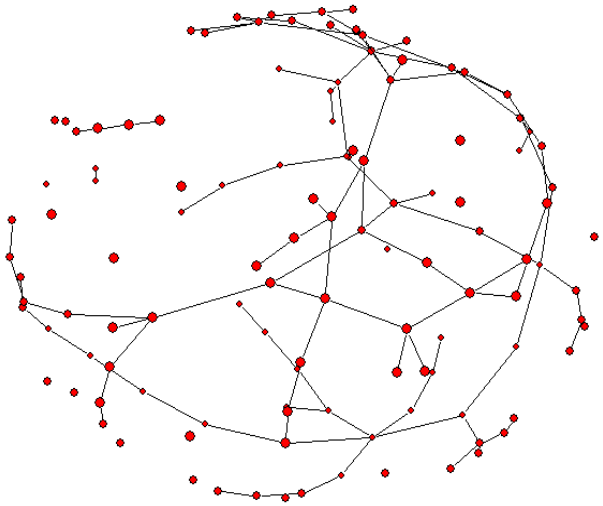
Regular networks
Symmetric

COMPLEX NETWORKS



Scale free networks
Small world
With communities
ENCODING INFORMATION
IN THEIR STRUCTURE

RANDOM GRAPHS



Totally random
Binomial degree
distribution

Why networks?

Because

NETWORKS

encode for the

THE INFORMATION CONTENT

of the entire complex system

Why modelling networks?

Because

NETWORKS MODELS

are essential to do

INFERENCE

starting from partial information about a network

Biological networks

The vast majority of biological networks includes

NOISY DATA
(false positives and false negatives)

or simply

PARTIAL INFORMATION
(example the human protein network:
only about 30% of interactions known)

Interbank networks

Typically financial institutions retain information about their financial contracts confidential.

Therefore interbank networks can be only inferred from

PARTIAL PUBLICLY AVAILABLE DATA

such as the banks' balance sheet

Social networks

In social network there is large interest in

inferring their

COMMUNITY STRUCTURE

and

PREDICTING MISSING LINKS

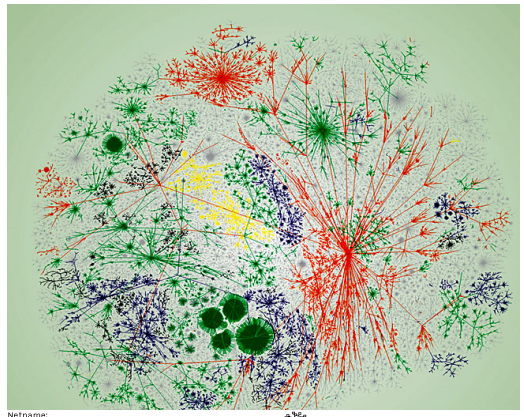
Networks
and
Generalized network structures

A classification

Types of networks

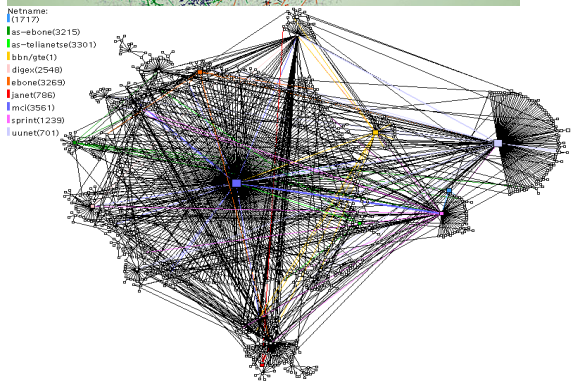
➢**Simple** Each link is either existent or non existent, the links do not have directionality

(protein interaction map, Internet,...)



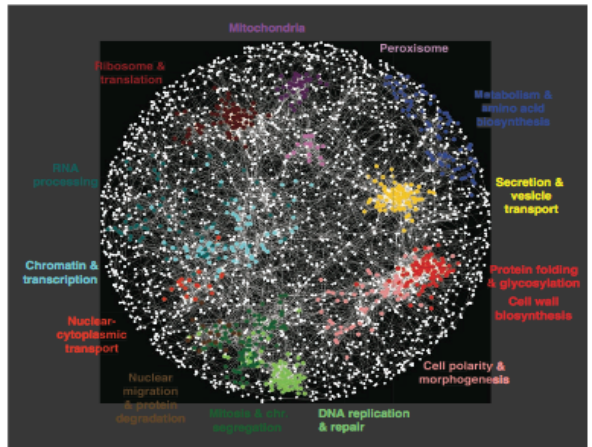
➢**Directed** The links have directionality, i.e., arrows

(World-Wide-Web, social networks...)



➢**Signed** The links have a sign

(transcription factor networks, epistatic networks...)



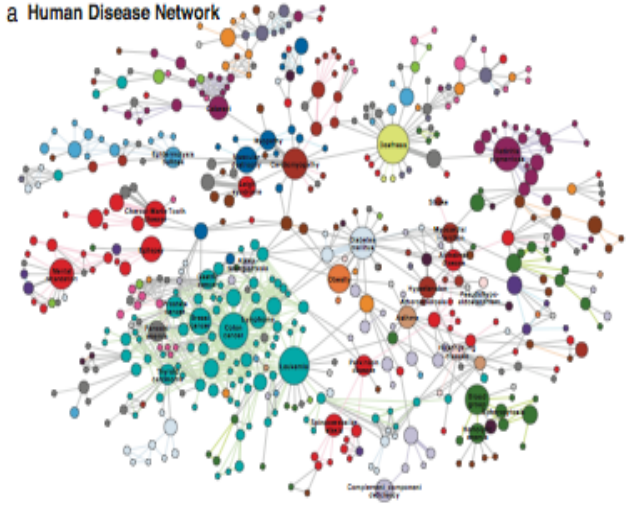
Types of networks

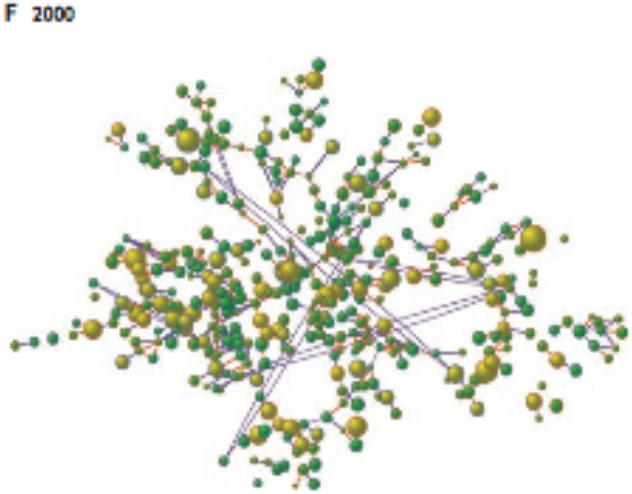
➢**Weighted** The links are associated to a real number indicating their weight

(airport networks, phone-call networks…)

➢**With features of the nodes** The nodes might have weight or associated feature

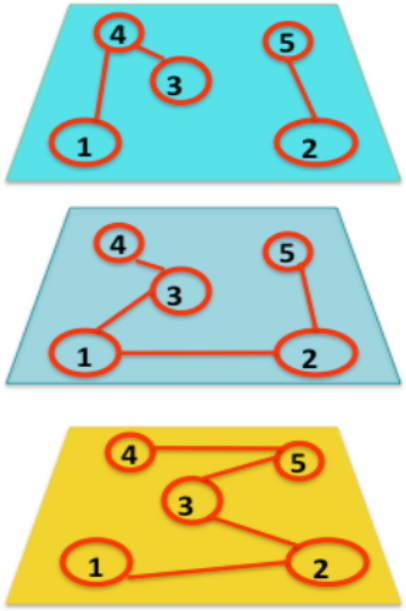
(social networks, diseasome, ect..)

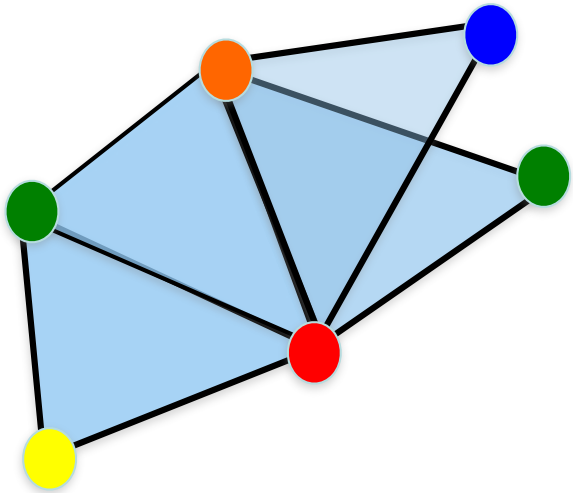


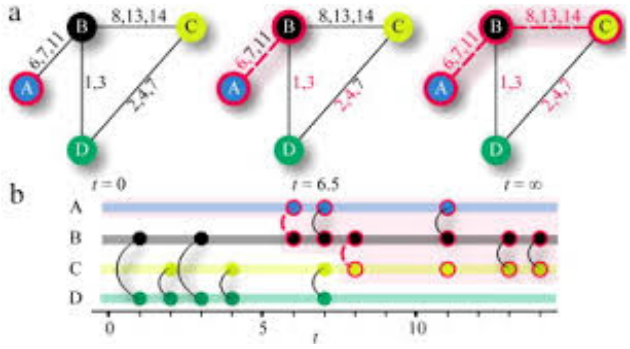


Generalized network structures

MULTILAYER NETWORKS **SIMPLICIAL COMPLEXES** **TEMPORAL NETWORKS**







Going beyond the framework of simple networks

is of fundamental importance

for a variety of applications ranging from

brain research to social and technological networks

Structure of the module

Lesson 1

Introduction of networks

Introduction to Maximum Entropy Principle

Lesson 2

Ensembles of networks

Random graphs

Exponential random graph models

Lesson 3

Microcanonical network ensemble with given degree sequence

Non-equivalence of the ensembles

Lesson 4

Hidden variable models and Block models

Growing network models and their entropy rate

Lesson 5

Maximum entropy models of

Multiplex networks

Temporal networks

Simplicial complexes

Contact

Professor Ginestra Bianconi

Queen Mary University of London

Email: *g.bianconi@qmul.ac.uk*

First lesson

- Introduction to Simple Networks

- Introduction to Maximum Entropy Principle

References Lesson 1

Network science books

A.L. Barabasi *Network Science* (Cambridge University Press)

Bianconi *Multilayer Networks: Structure and Function* (Oxford University Press)

Information Theory books

Cover & Thomas *Elements of information theory* (Wiley)

MacKay *Information Theory, Inference and Learning Algorithms*
(Cambridge University Press)

References Lesson 1

Network science books

A.L. Barabasi *Network Science* (Cambridge University Press)
Bianconi *Multilayer Networks: Structure and Function* (Oxford University Press)
S. N. Dorogovstev and J. F. F. Mendes *Evolution of networks* (Oxford University Press)

Papers

Barabási, A.L. and Albert, R., 1999. Emergence of scaling in random networks. *science*, *286*(5439), pp.509-512.
Bianconi, G. and Barabási, A.L., 2001. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, *54*(4), p.436.
Dorogovtsev, S.N., Mendes, J.F.F. and Samukhin, A.N., 2000. Structure of growing networks with preferential linking. *Physical review letters*, *85*(21), p.4633.
Bianconi, G. and Barabási, A.L., 2001. Bose-Einstein condensation in complex networks. *Physical review letters*, *86*(24), p.5632.
1.Bianconi, G., 2007. The entropy of randomized network ensembles. *EPL (Europhysics Letters)*, *81*(2), p.28005.
2. Bianconi, G., 2009. Entropy of network ensembles. *Physical Review E*, *79*(3), p.036114.
3.Clarke, L.E., 1958. On Cayley's formula for counting trees. *Journal of the London Mathematical Society*, *1*(4), pp.471-474.
4. Zhao, K., Halu, A., Severini, S. and Bianconi, G., 2011. Entropy rate of nonequilibrium growing networks. *Physical Review E*, *84*(6), p.066113.
5. Menichetti, G., Remondini, D., Panzarasa, P., Mondragón, R.J. and Bianconi, G., 2014. Weighted multiplex networks. *PloS one*, *9*(6), p.e97857.

Introduction

to

Simple Networks

Graphs and networks

Definition

A *graph* is an ordered pair $G = (V, E)$ comprising a set V of *vertices* connected by the set E of *edges*.

Definition

A *network* is the graph $G = (V, E)$ describing the set of interactions between the constituents of a complex system. The vertices of a network are called *nodes* and the edges *links*.

The *network size* N is the total number of nodes in the network $N=|V|$.
The total number of links L is given by $L=|E|$.

Dictionary

Graph theory and network theory use a different terminology.

In this course we will use the network theory terminology.

It might be useful to refer to this small dictionary when reading the literature.

	Graph Theory Terminology	Network Theory Terminology
N	Vertices	Nodes
L	Edges	Links
Links connecting a node with itself	Loops	Tadpoles

Examples of complex networks

Complex networks	Nodes	Links
Actors network	Actors	Co-acting on a movie
Collaboration networks	Scientists	Co-authors in one paper
Citation networks	Scientific papers	Citation
Facebook network	Individuals	Facebook friends
Metabolic network	Metabolites	Common chemical reaction
Protein-Interaction networks	Proteins	Physical interaction
Transcription networks	Genes	Regulation
Brain network	Neurons	Synaptic connections
Internet	Routers	Physical lines
World-Wide-Web	Webpages	URL's addresses
Airport network	Airports	Flight connections
Power-grids	Power plants	Electric grid

Labelled networks

A *labelled network*, of network size N, is formed by a set V of N distinguishable nodes indicated by a different and unique label

$$i \in \{1,2,\ldots,N\}$$

and by a set of links E characterising the interactions between pairs of nodes.

Simple networks

Definition

A *simple network* of N nodes is a network in which links are undirected and unweighted, and in which there are no tadpoles.

In other words the network is fully specified by a list of pairwise interactions of symmetric nature (if node i is linked to node j also node j is connected to node i).

Adjacency matrix

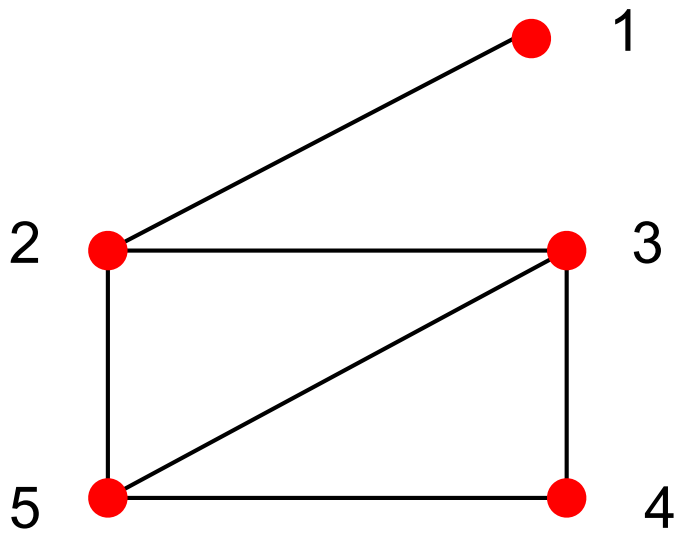
A simple network is fully determined by its adjacency matrix.

The adjacency matrix a of a simple network is a $N \times N$ matrix of elements given by

$$a_{ij} = \begin{cases} 1 \text{ if } i \text{ is linked to } j \\ 0 \text{ otherwise.} \end{cases}$$

The adjacency matrix of a simple network is symmetric.

Example of a simple network



This is a simple network of N=5 nodes and L=6 links with adjacency matrix

$$a = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Number of nodes and links

The most important characteristic of a simple network are:

- the total number of nodes N=|V|,
- the total number of links |L|=|E| with

$$L = \frac{1}{2} \sum_{i,j} a_{ij}$$

Middle size complex networks

When does a network starts to become “complex”?

Open question

	N
C.elegans Brain network	309
Venter minimal cell	256
Power-grids	$\sim 10^2 - 10^3$ (**but up to** 10^5)
Airport networks (single continent)	$\sim 10^2 - 10^3$
Zachary karate club network	34
Ecological networks datasets	$\sim 10^2 - 10^3$

Large complex networks

Many network datasets have large network size

	N
Brain network	Up to 10^{11}
Online Social networks	Up to 10^{9}
WWW	10^{9}
Internet	Up to 10^{5}

Network dataset repositories

For people that like to play with data

Three of the most popular network repositories are:

- SNAP: Stanford Network Analysis Project

http://snap.stanford.edu/

- Network Repository

http://networkrepository.com/

- Web of Life

http://www.web-of-life.es/

Sparse regime

Definition

A *sparse network* has a total number of links L of the same order of magnitude of the total number of nodes N.

$$L = \mathcal{O}(N)$$

Comments

These are networks in which in average every node has a finite number of connections.

A sparse network model will allow to model networks with tunable number of nodes N and with a average number of connections of every node independent of the total number of nodes N.

Degree

Definition

In a simple network the *degree* k_i of node i is given by the total number of links incident to node i

The degree k_i of node i can be expressed in terms of the adjacency matrix as

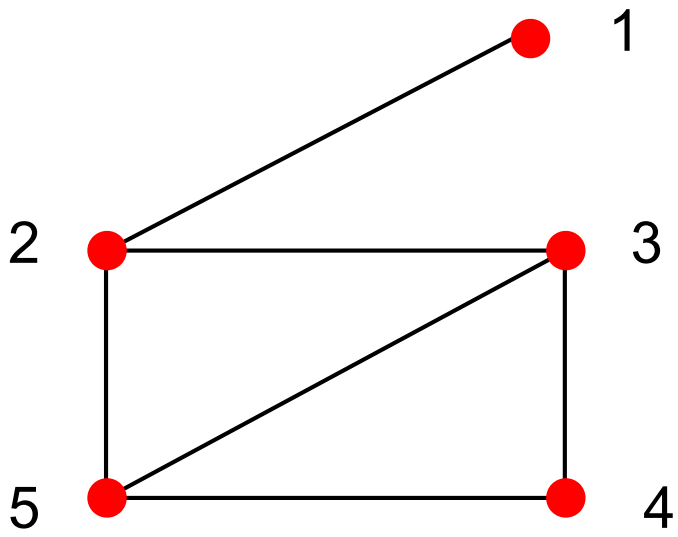
$$k_i = \sum_{j=1}^{N} a_{ij},$$

or equivalently

$$k_i = \sum_{j=1}^{N} a_{ji}\,.$$

The maximum degree K of a simple network of N nodes is **K=N-1**

Example of a simple network



This is a simple network of N=5 nodes and L=6 links with adjacency matrix

$$a = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Degrees

$$k_1 = 1, k_2 = 3, k_3 = 3, k_4 = 2, k_5 = 3$$

Degree sequence

Definition

The *degree sequence* of a simple network is the ordered sequence

$$\{k_i\}_{i=1,2\ldots,N} = \{k_1, k_2, \ldots, k_N\}$$

of the degrees k_i of all the nodes of the network .

$\langle k \rangle$

Given the degree sequence of a simple network, we can define its average degree.

Average degree

Definition

The *average degree* of the network with degree sequence $\{k_i\}_{i=1,2\ldots,N} = \{k_1, k_2, \ldots, k_N\}$ is defined as

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

Proposition $\langle k \rangle$

The average degree of a network is related to the number of nodes by

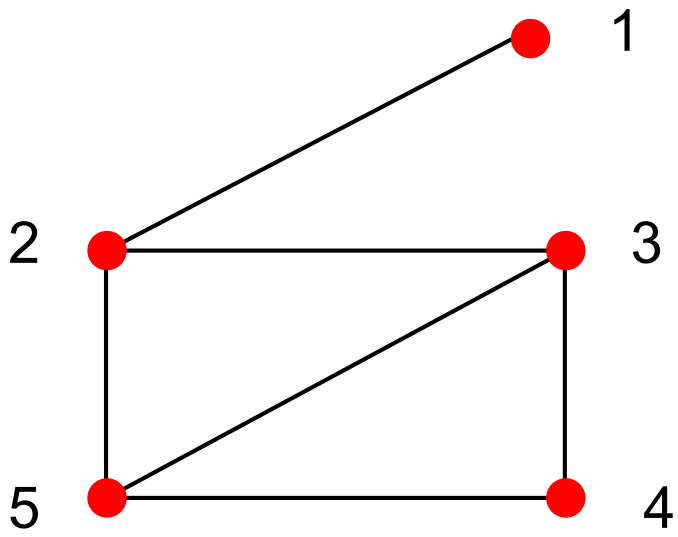
$$\langle k \rangle = \frac{2L}{N}$$

Proof

Indeed using the definition of the average degree and the expression of the degree of a node in terms of the adjacency matrix we have

$$\langle k \rangle N = \sum_{i=1}^{N} k_i = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} = 2L$$

Example of a simple network



This is a simple network of N=5 nodes and L=6 links with adjacency matrix

$$a = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Degrees

$$k_1 = 1, k_2 = 3, k_3 = 3, k_4 = 2, k_5 = 3$$

Degree Sequence

$$\{k_i\}_{i=1,\ldots,N} = \{1,3,3,2,3\}$$

Average degree, Maximum degree

$$\langle k \rangle = \frac{12}{5},$$

$$K = 3$$

Degree distribution

Definition

The *degree distribution* P(k) of a simple network is a function defined for $k \in \{0,1,2\ldots,N-1\}$
It indicates the fraction of nodes of degree k.
It also indicates the probability that a randomly chosen node of the network has degree k.

Let us indicate with N(k) is the total number of nodes of the network with degree k, i.e.

$$N(k) = \sum_{i=1}^{N} \delta(k, k_i)$$

where $\delta(k, k_i)$ indicates the Kronecker delta, i.e. $\delta(k, k_i) = 1$ **if** $k_i = k$ and $\delta(k, k_i) = 0$ otherwise.

The degree distribution P(k) of simple network is given by

$$P(k) = \frac{N(k)}{N} = \frac{1}{N} \sum_{i=1}^{N} \delta(k_i, k)$$

Properties of the degree distribution

The degree distribution is non-negative $P(k) \geq 0 \ \forall k$, and normalized

$$\sum_{k=0}^{N-1} P(k) = 1.$$

Definition

The *n-moment* of a degree distribution is defined as the expectation of k^n

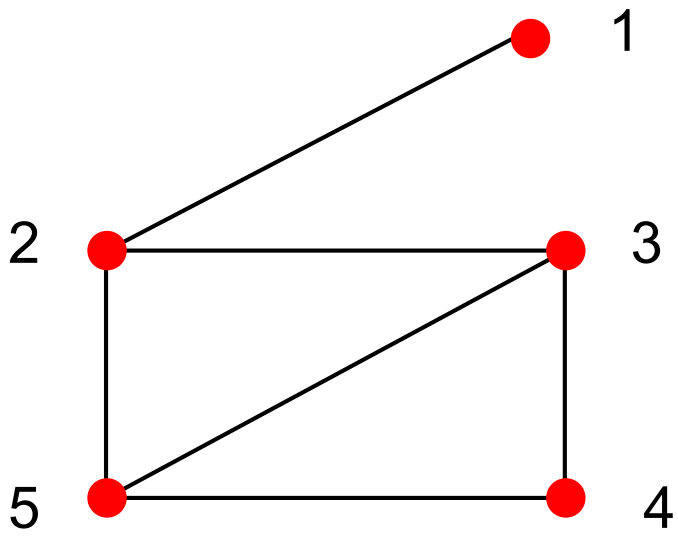
and will be indicated as $\langle k^n \rangle$. Therefore

$$\langle k^n \rangle = \sum_{k=0}^{N-1} k^n P(k) = \mathbb{E}(k^n)$$

The average degree is the first moment of the degree distribution

$$\langle k \rangle = \sum_{k=0}^{N-1} kP(k) = \frac{1}{N} \sum_{i=1}^{N} k_i$$

Example of a simple network



This is a simple network of N=5 nodes and L=6 links with adjacency matrix

$$a = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Degrees

$$k_1 = 1, k_2 = 3, k_3 = 3, k_4 = 2, k_5 = 3$$

Degree Sequence

$$\{k_i\}_{i=1,\ldots,N} = \{1,3,3,2,3\}$$

Average degree, Maximum degree

$$\langle k \rangle = \frac{12}{5},$$

$$K = 3$$

Degree distribution

$$P(0) = 0, \qquad P(1) = 1/5,$$

$$P(2) = 1/5, \qquad P(3) = 3/5, \quad P(4) = 0.$$

Clustering coefficient

Definition

In a simple network the *clustering coefficient* C_i o a node *i* of degree k_i is defined as

$$C_i = \frac{\sum_{j<k} a_{ij} a_{jk} a_{ki}}{k_i(k_i - 1)/2} \qquad \textbf{if } k_i > 1$$

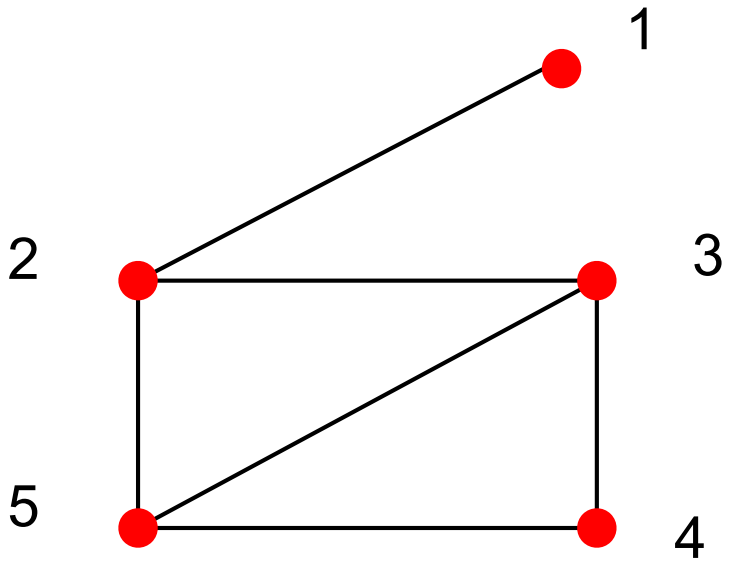
$$C_i = 0 \qquad \textbf{if } k_i \leq 1$$

Comments

The clustering coefficient measures the density of triangles around a node
(i.e. in the ego-centered network of node i) and is one of the most popular network measures.

Note however that low clustering coefficient does not imply that the network has only large loops
(take for instance a square grid, which is highly clustered but does not contain any triangle)

Clustering coefficient



Definition

$$C_i = \frac{\sum_{j<k} a_{ij} a_{jk} a_{ki}}{k_i(k_i - 1)/2} \qquad \textbf{if } k_i > 1$$

$$C_i = 0 \qquad \textbf{if } k_i \leq 1$$

Clustering coefficient C_i

$$C_1 = 0$$
$$C_2 = 1/3$$
$$C_3 = 2/3$$
$$C_4 = 1$$
$$C_5 = 2/3$$

Paths and shortest distance

Definitions

A *path* of a network, is a sequence of nodes, such that every consecutive pair of nodes is connected by a link.

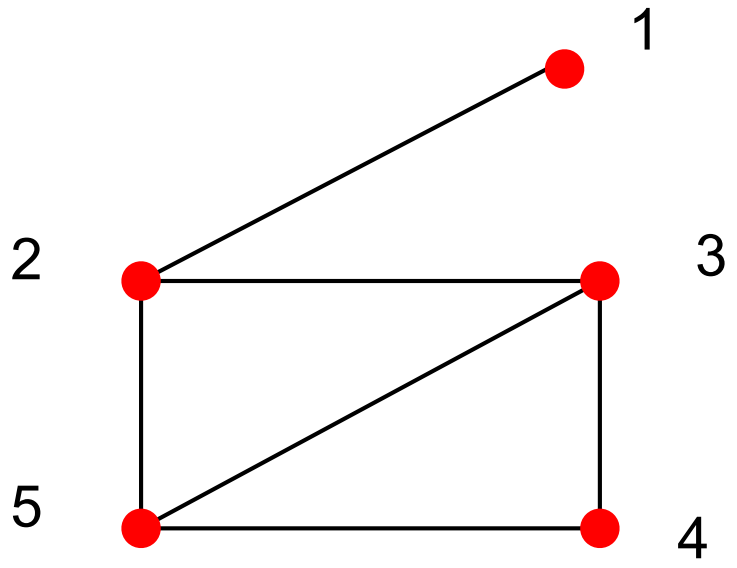
The *path length* is equal to the number of links traversed along the path, including eventual repetitions in the case of paths that intersect themselves.

A *shortest path* between node i and node j is a path of minimum length.

The *shortest distance* d_{ij} between node i and node j is the length of any shortest path between node i and node j.

Shortest distance

The *shortest distance or simply distance* between two nodes *is the minimal number of links than a path must traverse to go from one node to the other*



Shortest distance (or distance)

$$d_{1,2} = 1, d_{1,3} = 2, d_{1,4} = 3, d_{1,5} = 2, d_{3,5} = 1$$

Average shortest distance and diameter of a network

Definitions

The average shortest distance ℓ of a connected network is the average of the shortest distances between any two distinct nodes of the network.
Therefore, in a connected network we have

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1,N|j\neq i} d_{ij} .$$

The diameter D of a connected network is the maximum of the shortest distances between any two nodes of the network given by

$$D = \max_{i,j\neq i} d_{ij} .$$

Comment

It follows that we always have $D \geq \ell$

Subgraph

Definitions

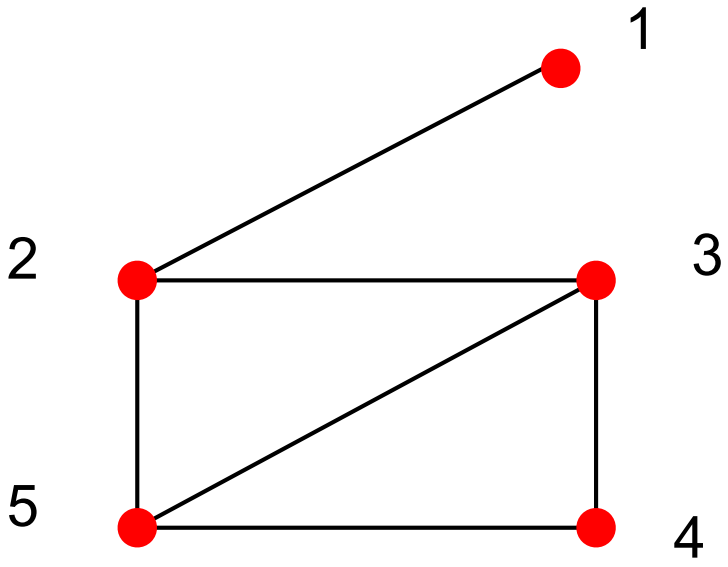
A *subgraph* H=(V',E') of a network G=(V,E) is formed by a set of nodes $V' \subset V$ and by a set of links E' such that $E' \subset E$ with all the links in E' are incident only to nodes included in V'.

A *loop of size* L is a connected subgraph formed by L links such that every node has degree 2.

A *clique* of size c is a fully connected subgraph of the network of c nodes and c(c-1)/2 links

Loops of size L

A loop of size L is a connected subgraph formed by L links such that every node has degree 2



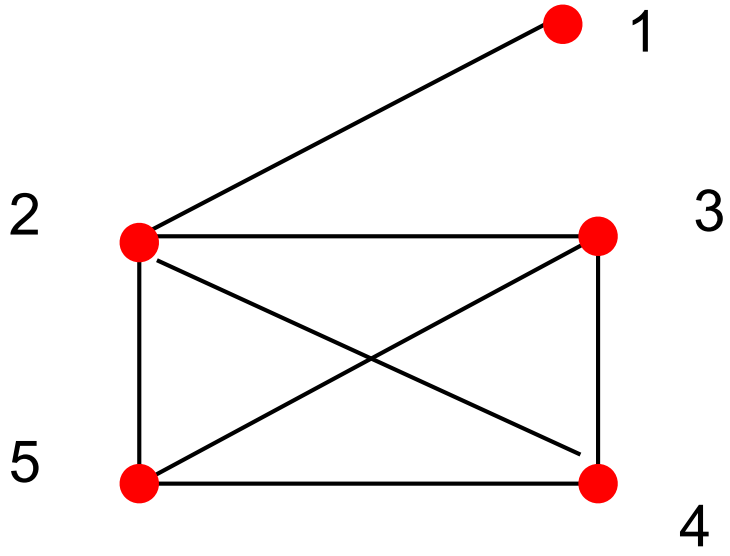
This network has

➢2 loops of size 3

➢and 1 loop of size 4

Cliques

Clique of size c is a fully connected subgraph of the network of c nodes and c(c-1)/2 links



This network contains
4 cliques of size 3 (triangles)
1 clique of size 4

Network Universalities

Small world networks

Complex networks have at the same time
a small diameter
like Cayley trees, Bethe lattices, and random graphs
i.e. L scales like the logarithm of the number of nodes N or slower

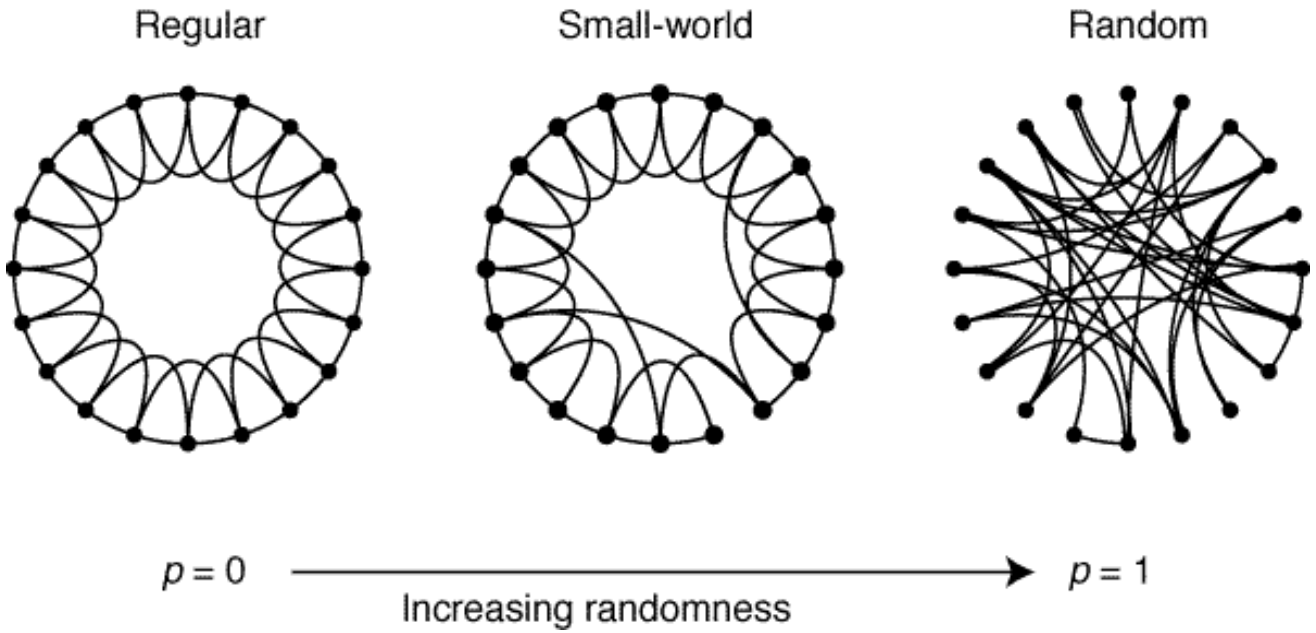
$$D = \mathcal{O}(\ln N) \textbf{ or } D = o(\ln N)$$

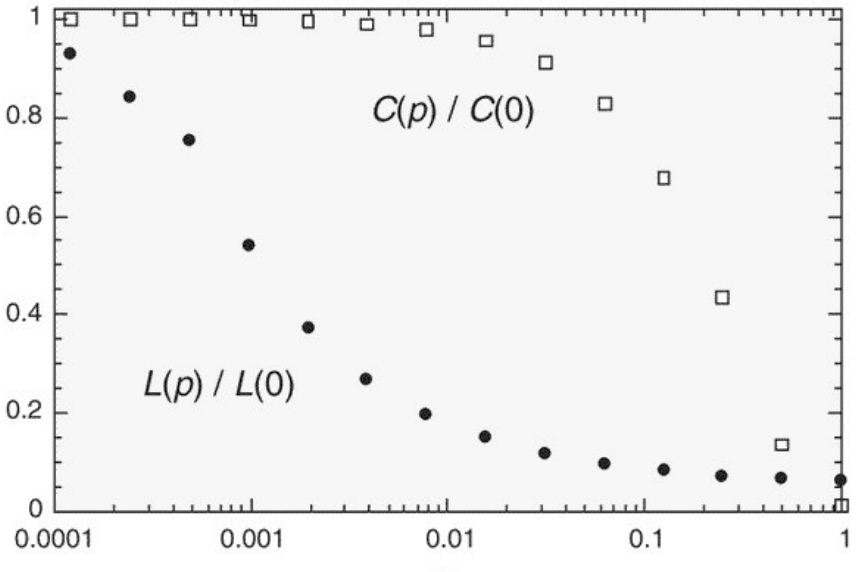
and significant density of small loops
like lattices
which is measured typically by a large average clustering coefficient of the nodes

$$\langle C \rangle = \mathcal{O}(1)$$

Watts and Strogatz (1998)

Watts and Strogatz small world model



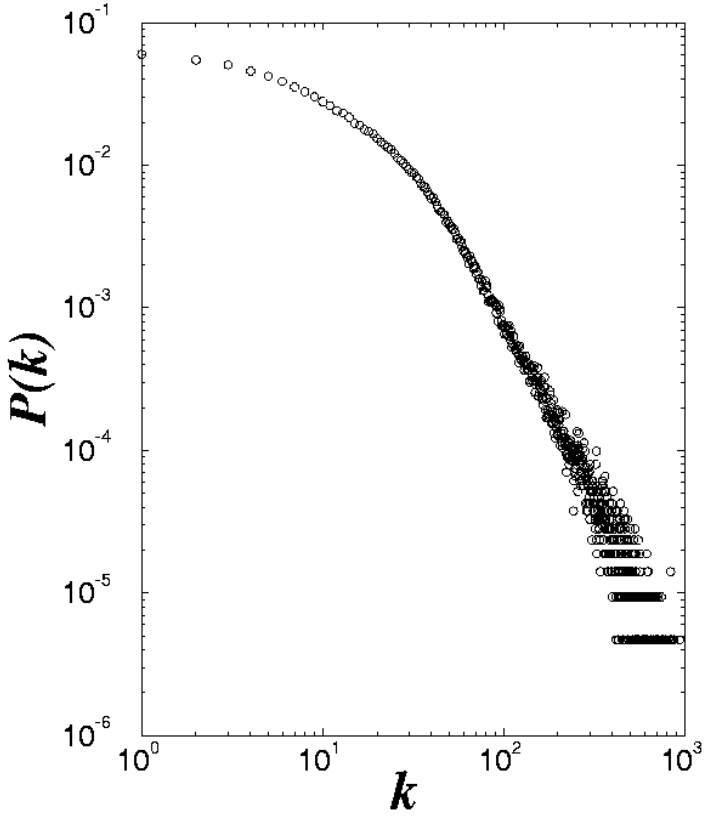


There is a wide range of values of p in which high clustering coefficient coexist with small average distance

Power law networks

Power-law networks are networks with degree distribution

$$P(k) = Ck^{-\gamma}$$



For $\gamma > 3$

$\langle k \rangle \rightarrow$ **const** **for** $N \rightarrow \infty$

$\langle k^2 \rangle \rightarrow$ **const** **for** $N \rightarrow \infty$

For $\gamma \in (2,3]$

$\langle k \rangle \rightarrow$ **const** **for** $N \rightarrow \infty$

$\langle k^2 \rangle \rightarrow \infty$ **for** $N \rightarrow \infty$

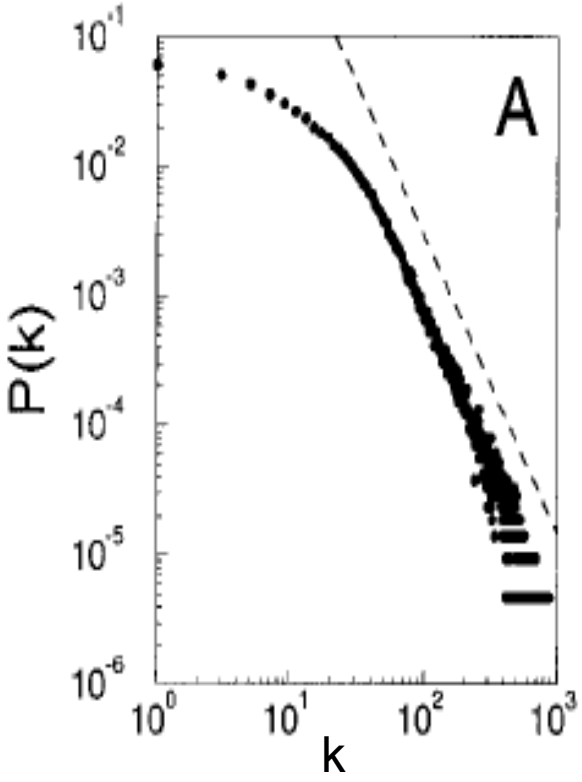
For $\gamma \in (1,2]$

$\langle k \rangle \rightarrow \infty$ **for** $N \rightarrow \infty$

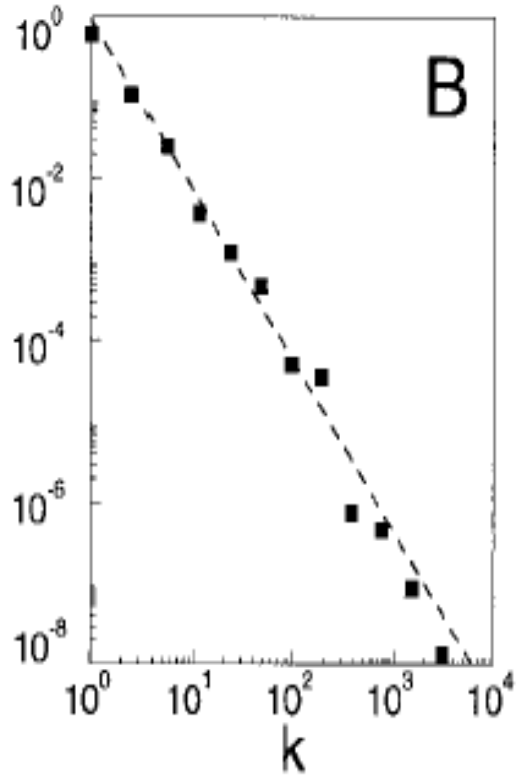
$\langle k^2 \rangle \rightarrow \infty$ **for** $N \rightarrow \infty$

Scale-free networks

Actor networks

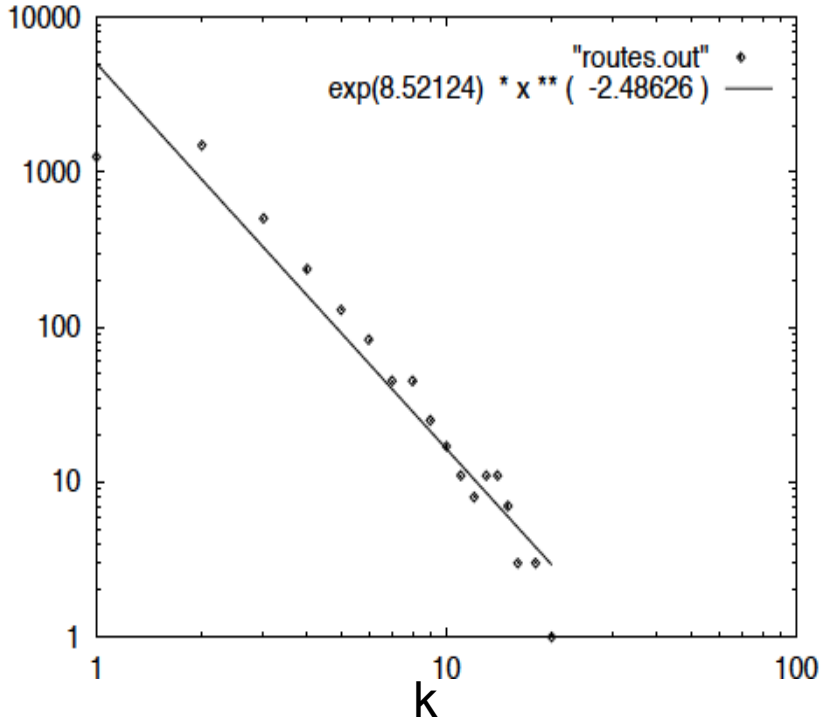


WWW



Barabasi-Albert 1999

Internet



Faloutsos, Faloutsos and Faloutsos 1999

$P(k) \simeq Ck^{-\gamma}$ **for** $k \gg 1$ **with** $\gamma \in (2,3]$

$\langle k \rangle \to$ **const** **for** $N \to \infty$

$\langle k^2 \rangle \to \infty$ **for** $N \to \infty$

Scale-free networks

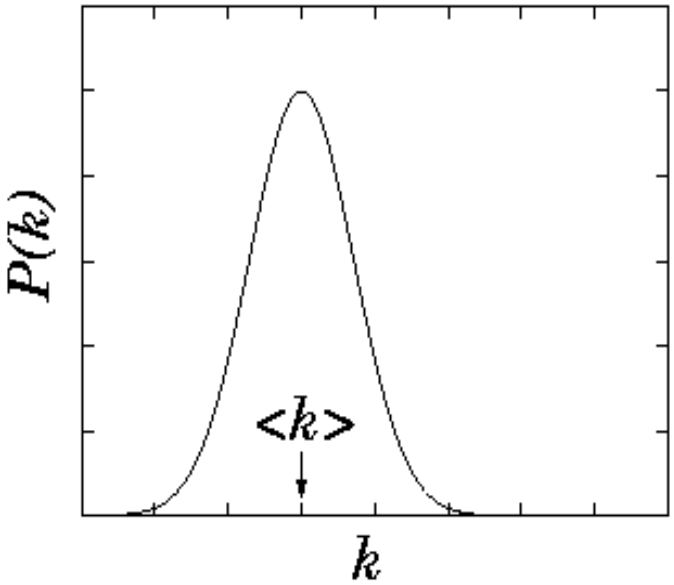
- A large variety of networks are scale-free.
- *Scale-free networks* are networks whose degree distribution $P(k)$ can be approximated ,for large values of the degree *k*, by a power-law, i.e.

$$P(k) \simeq Ck^{-\gamma} \text{ for } k \gg 1$$

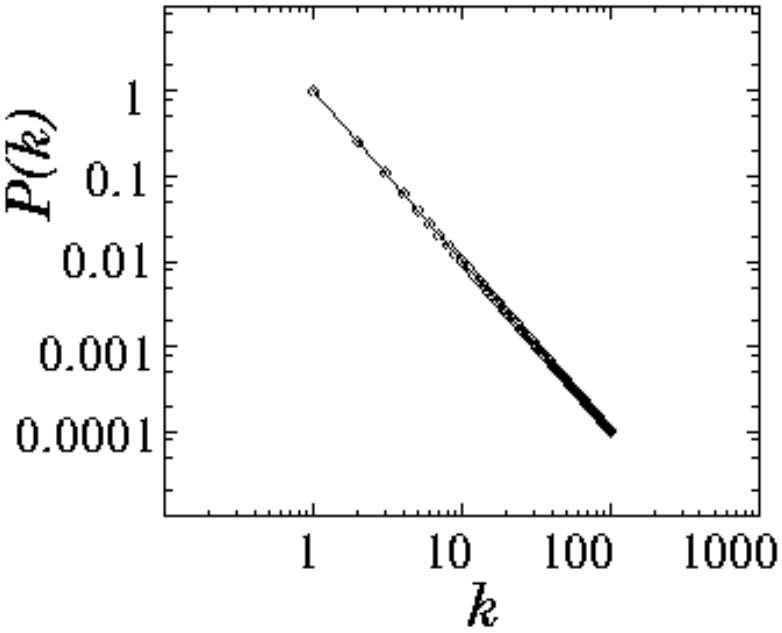
- with the exponent $\gamma \in (2,3]$

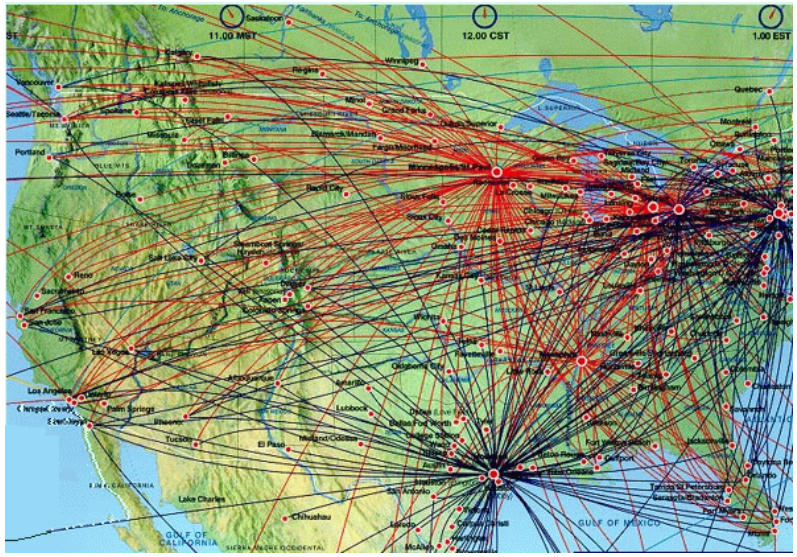
What does it mean?

Poisson distribution



Power-law distribution





Network with finite <k²>

Scale-free Network

Scale-free networks

Technological networks
Internet
World-Wide Web

Biological networks
Metabolic networks,
Protein-interaction networks,
Transcription networks

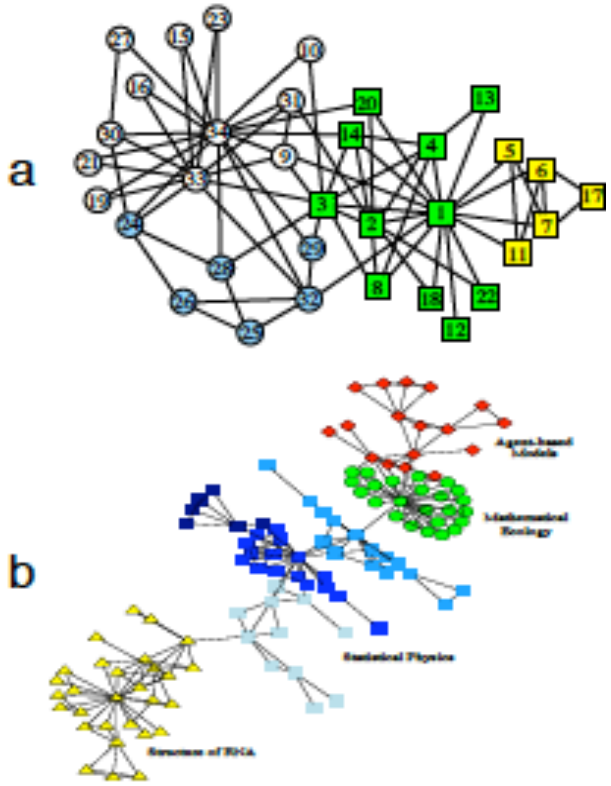
Transportation networks
Airport networks

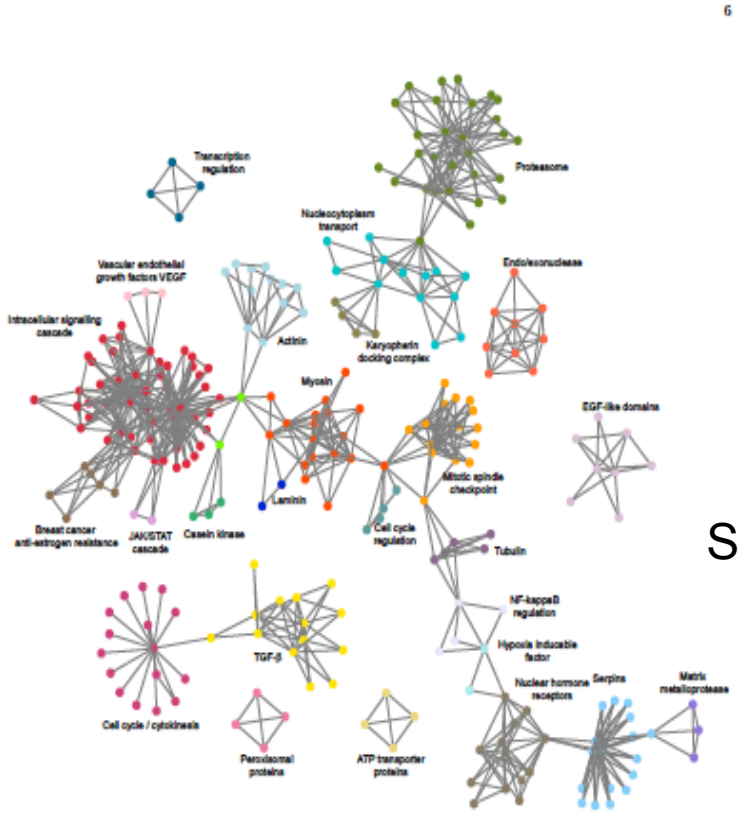
Social networks
Collaboration networks
Citation networks
Facebook

Economical networks
Networks of shareholders
The World Trade Web

Community structure

most complex networks
have a
mesoscale structure
which reveal densely connected
communities

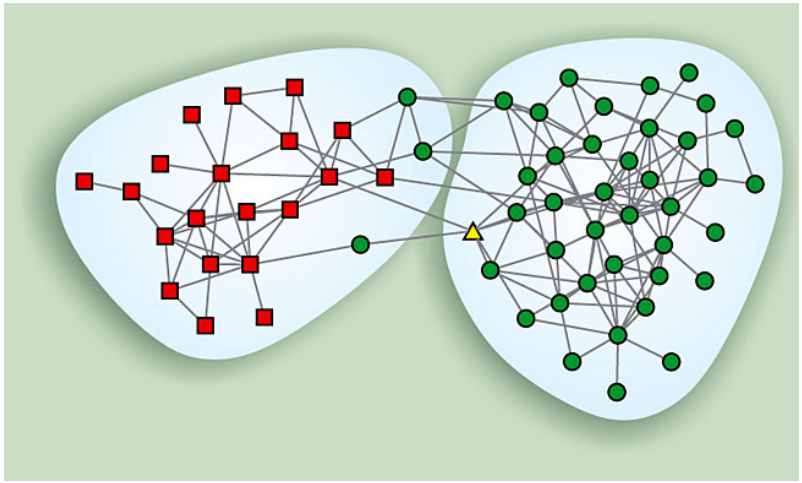




From
S. Fortunato
RMP

Communities

Dolphins social network



A community of a network define a set of nodes more likely to be connected to each other than to the other nodes of the network

Introduction

to

Maximum Entropy Principle

Ensemble

Definition

An ensemble X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$ where the outcome x is the value of a random variable which takes on one of possible values $\mathcal{A}_X = \{a_1, a_2, \ldots a_M\}$ having probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots p_M\}$ with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{i \in \mathcal{A}_X} P(x = a_i) = 1$

Abbreviation

Briefer notation will be used. For example, $P(x = a_i)$ maybe written as $P(a_i)$ or $P(x)$

Joint ensemble

A *joint ensemble* XY is an ensemble in which each outcome is an ordered pair (x, y) with $x \in \mathcal{A}_X = \{a_1, a_2, \ldots a_M\}$ $y \in \mathcal{A}_Y = \{b_1, b_2, \ldots b_R\}$

We call $P(x, y)$ *the* ***joint probability*** *of* (x, y)

Marginal probability

We can obtain the marginal probability $P(x)$ from the joint probability $P(x, y)$ by summation $P(x) = \sum_{y \in \mathcal{A}_Y} P(x, y)$

Conditional probability

The conditional probability is defined as

$$P(x = a_i \,|\, y = b_j) = \frac{P(x = a_i, y = b_j)}{P(y = b_j)} \quad \textbf{if} \quad P(y = b_j) \neq 0$$

Shannon information content of an outcome

Definition

The *Shannon information content of an outcome* is defined to be

$$h(x) = -\log_c p(x)$$

Comment

The original definition is given in bits, i.e. the base of the logarithm is chosen to be $c = 2$. However a popular choice is also $c = e$. The Shannon information content calculated in base $c = e$ and the one calculated in base $c = 2$ differ only by a multiplicative constant. If not explicitly stated here we take $c = e$

Shannon information content of an outcome

The smaller is the probability of an outcome, the larger is its Shannon information content

$$h(x) = -\ln p(x) = \ln \frac{1}{p(x)}$$

If the Shannon information content of a constant outcome is zero

$$p(x) = 1 \textbf{ then } h(x) = 0$$

Shannon information content of a joint ensemble

The Shannon information content of an outcome of a joint ensemble is given by

$$h(x, y) = -\ln p(x, y)$$

In the case in which x and y are independent we have that the Shannon information content of (x,y) is given by the sum of the information content of x and y

$$h(x, y) = -\ln p(x, y) = -\ln[p(x)p(y)] = -h(x) - h(y)$$

Entropy of an ensemble

Definition

The **entropy of an ensemble** is defined to be the average Shannon information of an outcome

$$S = -\sum_{x \in \mathcal{A}_X} P(x) \ln P(x)$$

where the following convention is adopted,

$$0 \ln 0 = 0$$

Therefore we can also write

$$S = -\sum_{x \in \mathcal{A}_X | P(x) > 0} P(x) \ln P(x)$$

Properties of the Entropy

The entropy is non negative and is zero only for deterministic outcomes

$$S \geq 0 \quad \textbf{with } S = 0 \textbf{ iff } P(x) = 1 \textbf{ for one } x$$

- **Proof:** Given the expression for the entropy

$$S = -\sum_{x \in \mathcal{A}_X | P(x) > 0} P(x) \ln P(x)$$

- If we have a non deterministic variable the

$$P(x) \in (0,1) \forall x \textbf{ therefore } h(x) = -\ln P(x) > 0 \textbf{ it follows that } S > 0$$

- If we have a deterministic outcome

If $P(x) > 0$ **then** $P(x) = 1$ **with** $h(x) = -\ln P(x) = 0$ **it follows that** $S = 0$

Properties of the Entropy

The entropy is maximised for uniform distribution

- If the random variable can take M distinct values, i.e.

$$\textbf{If} \, |\mathcal{A}_X| = M$$

- then the maximum entropy over all possible distributions is

$$\max_{P(x)} S[P(x)] = S[P_U(x)] = \ln M$$

- where $P_U(x)$ is the uniform distribution

$$P_U(x) = \frac{1}{M}$$

Proof

Let us assume that our variable can take M possible values $|\mathcal{A}_X| = M$

The entropy of any distribution $P(x)$ which is naturally normalised

$$\sum_{x \in \mathcal{A}_X} P(x) = 1$$

is given by

$$S = -\sum_{x \in \mathcal{A}_X} P(x) \ln P(x)$$

In order to maximise the entropy over all normalised distributions consider the functional

$$\mathcal{F} = S - \nu \left(\sum_{x \in \mathcal{A}_X} P(x) - 1 \right) = -\sum_{x \in \mathcal{A}_X} P(x) \ln P(x) - \nu \left(\sum_{x \in \mathcal{A}_X} P(x) - 1 \right)$$

where ν is a Lagrangian multiplier.
By differentiating respect to $P(x)$ and putting the derivative to zero we get

$$\frac{\partial \mathcal{F}}{\partial P(x)} = -\ln P(x) - 1 - \nu = 0$$

Proof (continuation)

From the equations

$$\frac{\partial \mathcal{F}}{\partial P(x)} = -\ln P(x) - 1 - \nu = 0 \ \forall x \in \mathcal{A}_X$$

we get

$$P(x) = e^{-1-\nu}$$

By extremising $\mathcal{F}$ with respect to ν we get the normalization condition

$$\frac{\partial \mathcal{F}}{\partial \nu} = -\left(\sum_{x \in \mathcal{A}_X} P(x) - 1 \right) = 0$$

Since we have $|\mathcal{A}_X| = M$ the normalisation condition reads

$$\sum_{x \in \mathcal{A}_X} P(x) = e^{-1-\nu} M = 1 \ \textbf{or equivalently} \ e^{-1-\nu} = \frac{1}{M}$$

It follows that the distribution $P(x)$ that maximised the entropy is uniform

$$P(x) = P_U(x) = \frac{1}{M}$$ and that $$S[P_U(x)] = -\sum_{x \in \mathcal{A}_X} \frac{1}{M} \ln \frac{1}{M} = \ln M$$

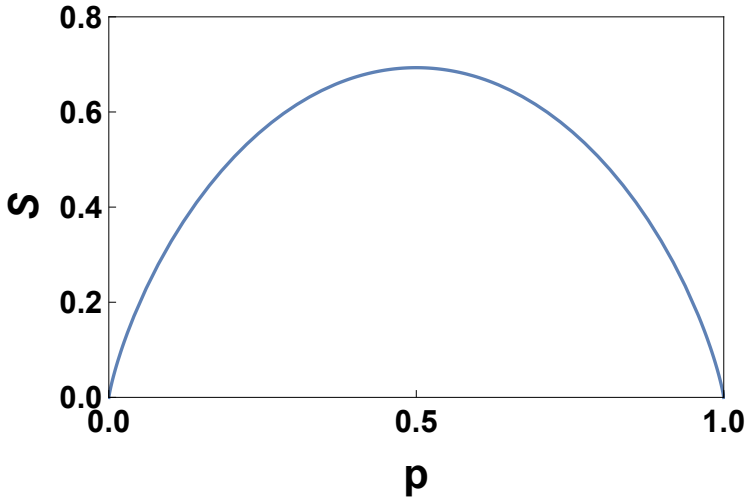
Entropy of a Bernoulli variable

Given a Bernoulli variable $x \in \{0,1\}$

with distribution $P(x) = p^x(1-p)^{1-x}$

the entropy is given by

$$S = -p \ln p - (1-p)\ln(1-p)$$



The entropy is zero for p=0 or p=1 (deterministic variable) and is maximised for p=1/2, i.e.

$$S = 0 \textbf{ for } p = 0 \textbf{ or } p = 1$$

$$S = \ln M = \ln 2 \textbf{ for } p = \frac{1}{2}$$

The entropy is a concave function

Entropy of a joint ensemble

Defintion

The entropy of a joint ensemble is defined as

$$S = -\sum_{(x,y)\in \mathscr{A}_X Y} P(x,y)\ln P(x,y)$$

with the usual convention $0\ln 0 = 0$

Uncorrelated joint ensembles

For uncorrelated variables, i.e. if $P(x,y) = P(x)P(y)$

The entropy is given by $S = -\sum_{(x,y)\in \mathscr{A}_X Y} P(x)P(y)\ln[P(x)P(y)]$

therefore we have $S = S_X + S_Y$

Quote

Everything should be made

as simple as possible, but not simpler

Einstein

Maximum entropy principle

The least biased ensemble

that satisfies a set of constraints

if the ensemble that maximises the entropy

(under the imposed constraints)

Maximum entropy principle

- Typically the constraints come from observations (data) or from previous knowledge about the ensemble.
- The maximum entropy principle is a very powerful tool to construct ensemble starting from partial information

Examples of Maximum entropy ensembles

Let us construct a maximum entropy ensemble in which we fix the expectations of some observables

$$f_\mu(x) \textbf{ for } \mu = 1,2\ldots,P$$

i.e. our constraints will be

$$\sum_{x\in\mathcal{A}_X} P(x)f_\mu(x) = C_\mu \;\; \mu = 1,2\ldots,P$$

with $C_\mu,\ \mu = 1,2\ldots,P$ being P constants.

Examples of Maximum entropy ensembles

The maximum entropy ensemble satisfying these constraints is given by the Gibbs measure

$$P(x) = \frac{e^{-\sum_{\mu=1}^{P} \lambda_\mu f_\mu(x)}}{Z}$$

where Z is the normalisation constant also called partition function

$$Z = \sum_{x \in \mathcal{A}_X} e^{-\sum_{\mu=1}^{P} \lambda_\mu f_\mu(x)}$$

and λ_μ are the Lagrangian multipliers fixed by the constraints or equivalently

$$-\frac{\partial \ln Z}{\partial \lambda_\mu} = C_\mu$$

Proof

We consider the maximum entropy ensemble of distribution $P(x)$ satisfying the constraints

$$\sum_{x\in\mathscr{A}_X} P(x)f_\mu(x) = C_\mu \;\; \mu = 1,2\ldots,P$$

and the normalisation constraint

$$\sum_{x\in\mathscr{A}_X} P(x) = 1$$

Therefore we need to maximise the entropy $S = -\sum_{x\in\mathscr{A}_X} P(x)\ln P(x)$

Under this constraints.
To this end we consider the functional

$$\mathscr{F} = -\sum_{x\in\mathscr{A}_X} P(x)\ln P(x) - \sum_{\mu=1}^{P}\lambda_\mu\left(\sum_{x\in\mathscr{A}_X} P(x)f_\mu(x) - C_\mu\right) - \nu\left(\sum_{x\in\mathscr{A}_X} P(x) - 1\right)$$

where $\{\lambda_\mu\},\nu$ are Lagrangian multipliers.
By differentiating respect to $P(x)$ and to each Lagrangian multiplier putting the derivative to zero we can determine the maximum entropy ensemble distribution.

Proof (continuation)

These equations read

$$\frac{\partial \mathscr{F}}{\partial P(x)} = -\ln P(x) - \sum_{\mu=1}^{P} \lambda_\mu f_\mu(x) - 1 - \nu = 0$$

$$\frac{\partial \mathscr{F}}{\partial \lambda_\mu} = -\left(\sum_{x \in \mathscr{A}_X} P(x) f_\mu(x) - C_\mu \right) = 0$$

$$\frac{\partial \mathscr{F}}{\partial \nu} = -\left(\sum_{x \in \mathscr{A}_X} P(x) - 1 \right) = 0$$

From the first equation we get

$$P(x) = e^{-1-\nu} e^{-\sum_{\mu=1}^{P} \lambda_\mu f_\mu(x)}$$

From the normalisation condition we get

$$e^{\nu+1} = Z = \sum_{x \in \mathscr{A}_X} e^{-\lambda_\mu f_\mu(x)}$$

Finally $\{\lambda_\mu\}$ are fixed by the conditions

$$C_\mu = \sum_{x \in \mathscr{A}_X} f_\mu(x) P(x) = \frac{1}{Z} \sum_{x \in \mathscr{A}_X} f_\mu(x) e^{-\sum_{\tilde{\mu}=1}^{P} \lambda_{\tilde{\mu}} f_{\tilde{\mu}}(x)} = -\frac{\partial \ln Z}{\partial \lambda_\mu}$$

Entropy of the ensemble

- The entropy of this ensemble is given by

$$S = \sum_{\mu=1}^{P} \lambda_\mu C_\mu + \ln Z$$

- (left as an exercise)

Log-likelihood of an outcome

Consider an outcome x of a random variable with unknown distribution $P(x)$

We assume that the unknown distribution is coming from a family

of distributions $P_{\vec{\lambda}}(x)$ dependent on the parameters $\vec{\lambda}$

Definition

The *log-likelihood* of a parameters $\vec{\lambda}$ is defined as

$$\mathscr{L}(\vec{\lambda}\,|\,x) = \ln P_{\vec{\lambda}}(x)$$

Likelihood of a set of data

- Consider a set of data formed by independent outcomes of the random variable $\mathbf{x}$

$$\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$$

- The log-likelihood of this set of data is

$$\mathscr{L}(\vec{\lambda} \,|\, \mathbf{x}) = \sum_{i=1}^{N} \ln P_{\vec{\lambda}}(x_i)$$

Maximum likelihood estimation

The maximum likelihood estimation of the parameters $\overrightarrow{\lambda^\star}$

corresponding to the distribution $P_{\overrightarrow{\lambda^\star}}(x)$

that best approximate the data

(according to maximum likelihood estimation) takes the form

$$\overrightarrow{\lambda}^\star = \textbf{argmax}_{\overrightarrow{\lambda}} \mathscr{L}(\overrightarrow{\lambda} \,|\, \mathbf{x}) = \textbf{argmax}_{\overrightarrow{\lambda}} \left[\sum_{i=1}^{N} \ln P_{\overrightarrow{\lambda}}(x_i) \right]$$

Relation between maximum entropy and maximum likelihood

Assuming that $P_{\vec{\lambda}}(x)$ is the Gibbs measures of the type

$$P_{\vec{\lambda}}(x) = \frac{e^{-\sum_{\mu=1}^{P} \lambda_\mu f_\mu(x)}}{Z}$$

Maximum likelihood estimation of the parameters $\vec{\lambda}^{\star}$

$$\vec{\lambda}^{\star} = \textbf{argmax}_{\vec{\lambda}} \mathscr{L}(\vec{\lambda} \,|\, \mathbf{x})$$

Implies that $P_{\vec{\lambda}}(x)$ is the maximum entropy ensemble with constraints fixed by the data

$$\langle f_\mu(x) \rangle_{DATA} = \langle f_\mu(x) \rangle_{ENSEMBLE} = \sum_{x \in \mathscr{A}_X} P_{\vec{\lambda}} f_\mu(x)$$

Proof

Consider a set of data formed by independent outcomes of the random variable X

$$D = \{x_1, x_2, \ldots, x_N\}$$

The log-likelihood of this set of data is

$$\mathscr{L}(\overrightarrow{\lambda}\,|\,\mathbf{x}) = \sum_{i=1}^{N} \ln P_{\overrightarrow{\lambda}}(x_i)$$

assuming

$$P_{\overrightarrow{\lambda}}(x) = \frac{e^{-\sum_{\mu=1}^{P} \lambda_\mu f_\mu(x)}}{Z}$$

We have

$$\mathscr{L}(\overrightarrow{\lambda}\,|\,\mathbf{x}) = \sum_{i=1}^{N} \ln P_{\overrightarrow{\lambda}}(x_i) = -\sum_{\mu} \lambda_\mu \sum_{i=1}^{N} f_\mu(x_i) - N \ln Z$$

Proof

Maximising the log-likelihood

$$\mathscr{L}(\overrightarrow{\lambda}\,|\,\mathbf{x}) = \sum_{i=1}^{N} \ln P_{\overrightarrow{\lambda}}(x_i) = -\sum_{\mu} \lambda_\mu \sum_{i=1}^{N} f_\mu(x_i) - N \ln Z$$

The log-likelihood of this set of data is

$$0 = \frac{\partial \mathscr{L}(\overrightarrow{\lambda}\,|\,\mathbf{x})}{\partial \lambda_\mu} = -\sum_{i=1}^{N} f_\mu(x_i) - N \frac{\partial \ln Z}{\partial \lambda_\mu} \quad \textbf{for } \mu = 1,2\ldots,P$$

We get

$$\frac{1}{N}\sum_{i=1}^{N} f_\mu(x_i) = -\frac{\partial \ln Z}{\partial \lambda_\mu} = \sum_{x \in \mathscr{A}_X} P_{\overrightarrow{\lambda}}(x) f_\mu(x) \quad \textbf{for } \mu = 1,2,\ldots,P$$

Therefore we have

$$\langle f_\mu(x) \rangle_{DATA} = \langle f_\mu(x) \rangle_{ENSEMBLE} = \sum_{x \in \mathscr{A}_X} P_{\overrightarrow{\lambda}} f_\mu(x) \quad \textbf{for } \mu - 1,2\ldots,P$$

Final remarks

In this first lesson we have covered

A. *Introduction to networks*

B. *Maximum entropy principle*

In the next lesson we will introduce

maximum entropy ensembles of networks