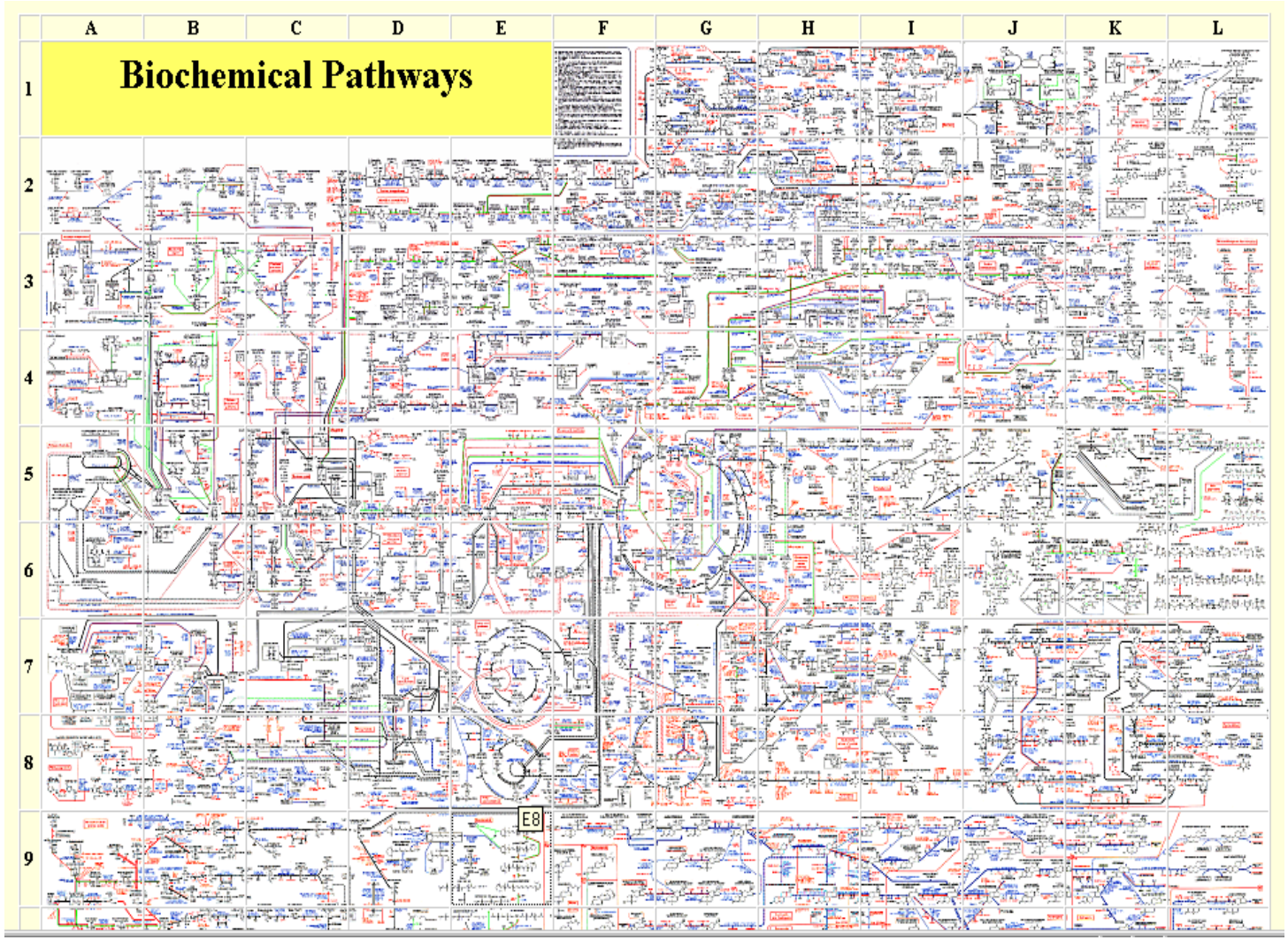
Metabolic Networks



https://www.unibo.it/en/teaching/course-unit-catalogue/course-unit/2018/433591

we can also share slides in english

Another book is the Laszlo Barabasi book: Linked

http://barabasi.com/book/linked

Network science

http://barabasi.com/book/network-science

network medicine

http://barabasi.com/book/network-medicine-complex-systems-in-human-disease-and-therapeutics

Graph

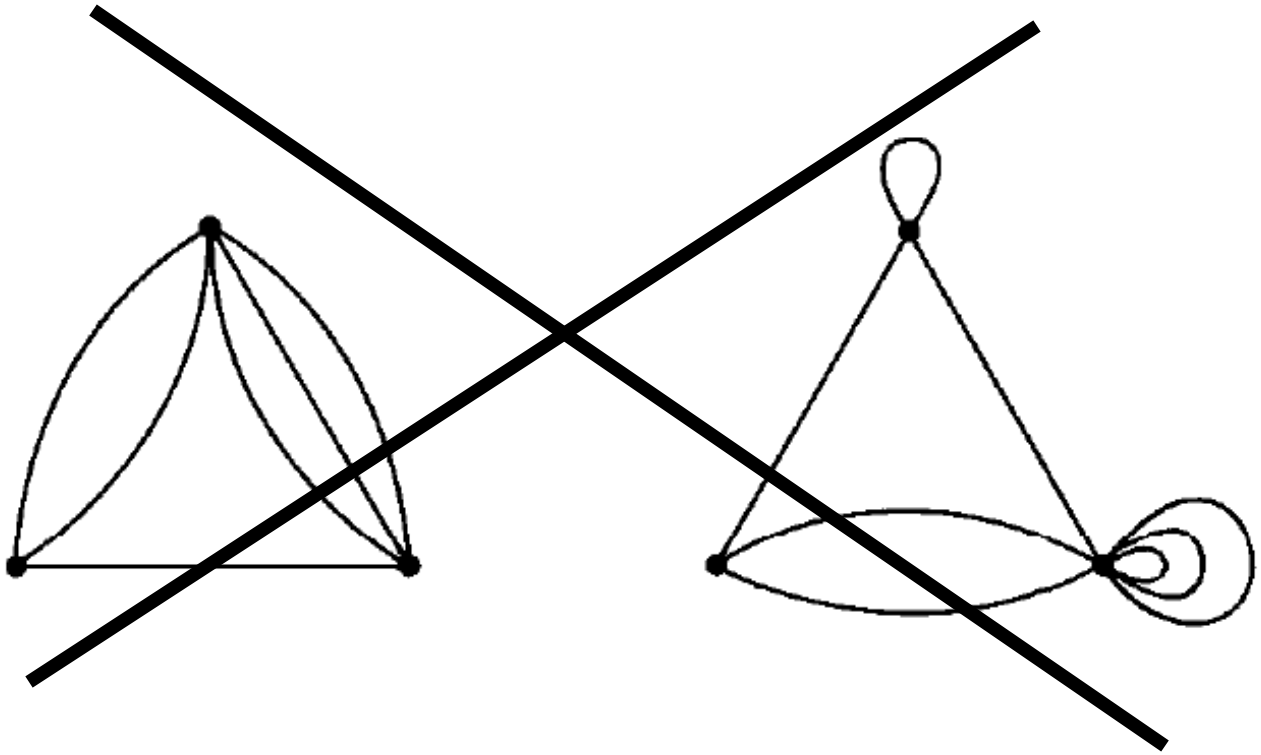
Graph G(V,L):

Finite nonempty set V of p points (vertexes, nodes) and set of k couples of distinct points L (links, lines, edges)

NOTES

In general: no loops (self-connections)

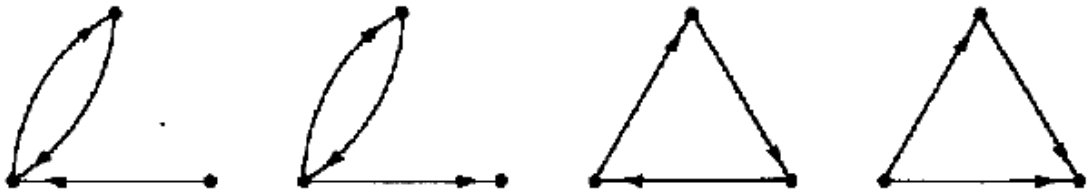
No multigraphs (multiple links between same nodes) only as weights



DiGraphs –directed graphs

Graphs with directed edges (from-to)
Edges: *arcs*
Oriented graph: without symmetric arcs

Hierarchical relationships: regulation, fluxes
neurons
genic transcription



Undirected graphs

Undirected graph: only symmetric links

Describe symmetrical relations:

friendship

chemical bond

statistical (Pearson's) correlation

can approximate a directed graph (…)

Network representation 1

Adjacency matrix A: NxN matrix (N nodes)

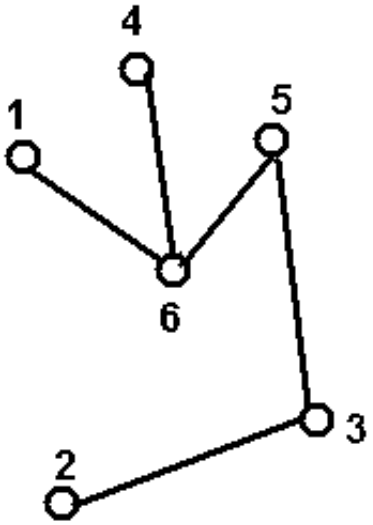
For unweighted networks:

A_{iJ} = 1 link from node i to node j

(1 link = 1 matrix element, no multigraph)

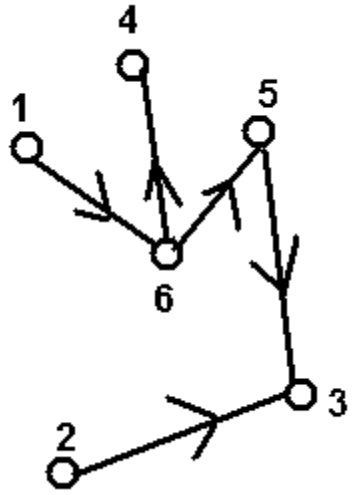
A_{ii} = 0 (no *loops*)

Examples



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

$$\mathrm{A} = \mathrm{A}^{\mathrm{T}}$$



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

$$\mathrm{A} \neq \mathrm{A}^{\mathrm{T}}$$

Adjacency matrix: weighted network

$A_{ij} = w_{ij}$ link from node i to node j with weight w_{ij}

Multigraph with N multi-edges: $w_{ij} = N$

If all weights are positive, we have useful theorems for matrices with positive values (Perron-Frobenius theorem) and we can connect to "physical" situations (transition probabilities)

Negative weights can be meaningful too: friend-enemy; Attraction-repulsion; ferromagnetic-antiferromagnetic (spin glasses)

Other representations 1.1

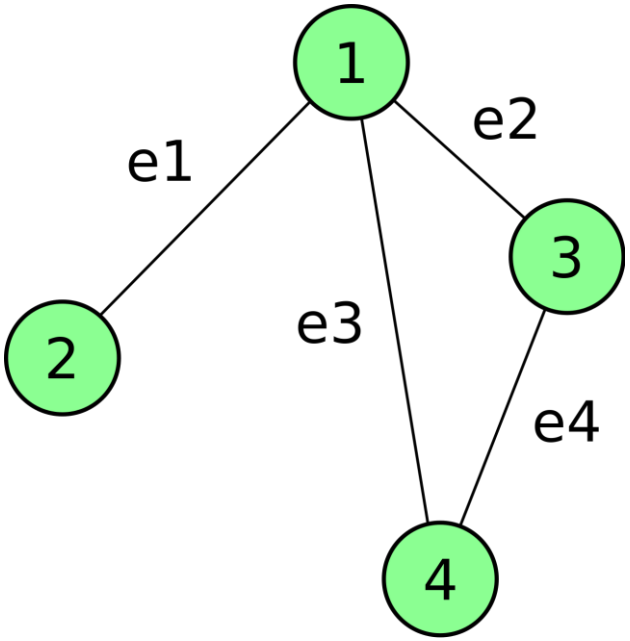
Incidence matrix I of G(V,L)

= matrix LxV (number L of links x number V of vertexes)

(one matrix line for each link: –1 outgoing +1 ingoing)

$$I_{NI} = -1; \quad I_{NJ} = +1; \quad \text{se } A_{IJ} \neq 0$$

Operatorial interpretation: “generalized derivative” on graph = oriented difference between values on adjacent nodes



The unoriented incidence matrix (or simply incidence matrix) of an undirected graph is a n × m matrix B, where n and m are the numbers of vertices and edges respectively: Bi,j = 1 if the vertex vi and edge ej are incident and 0 otherwise.

	e_1	e_2	e_3	e_4
1	1	1	1	0
2	1	0	0	0
3	0	1	0	1
4	0	0	1	1

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

The incidence matrix consists of 4 rows (corresponding to the four vertices, 1–4) and 4 columns (corresponding to the four edges, e1–e4)

Other representation 1.2

Link matrix (edge matrix) = matrix Lx2

(one line for each link)

3 5

3 8

12 34

22 32

Useful to store **sparse graphs** ($L=o(V^2)$): store single links instead of full matrix

$2xL << V^2$

Other representation 1.3

Useful for **diffusion processes on graph**

Each line describes a node:

Contains the indices of the arrival nodes

NOTE: same link can be represented in node lines (redundancy)

Es.:

3 4 2

1 5 12

1 4 5

…

Other representation 2

Laplacian Matrix L = D-A

$$L_{ij} = -A_{ij} \qquad i \neq j$$

$$L_{ii} = D_{ii} = k_i \qquad k_i = \sum_j A_{ij}$$

NOTE 1: theorems for L are not the same for A

NOTE 2: $I^T \cdot I = L$

symmetric matrix: loos information on link direction

Summary

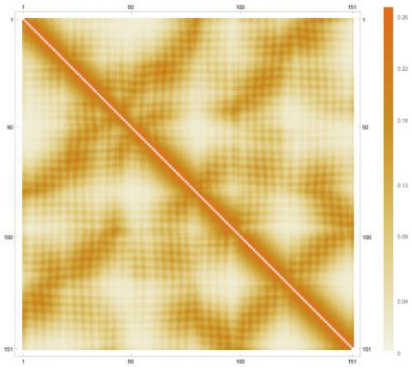
No loops -> $a_{ii}=0$
No multigraphs -> a_{mn} 1:1 link

Undirected graph -> symmetric adj. Matrix

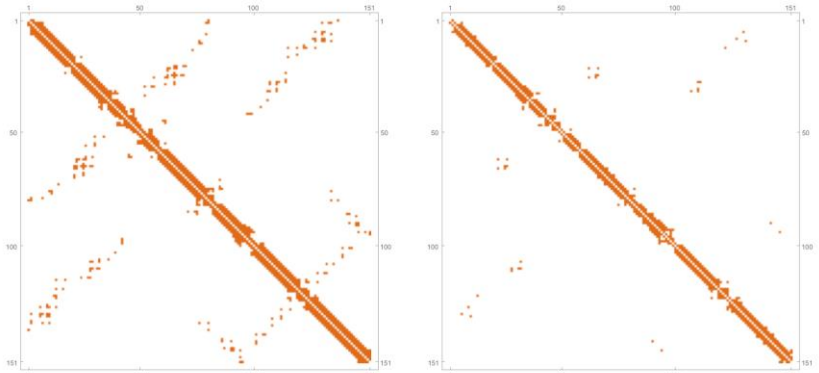
Directed graph -> asymmetric matrix

Protein contact maps

Represents the distance between all possible amino acid residue pairs of a three-dimensional protein structure using a binary two-dimensional matrix. The i j element of the matrix is 1 if the two residues are closer than a threshold, and 0 otherwise.



(a)



(b) (c)

Network representation of peptide bonds

or

Weighted network (intrapeptide distances)

- **Symmetric**
- **Spatial closeness**
- **Backbone**
- **”intrinsic” metric**

Connected graph (undirected)

A graph is connected if it is not divided into two or more non-communicating parts.

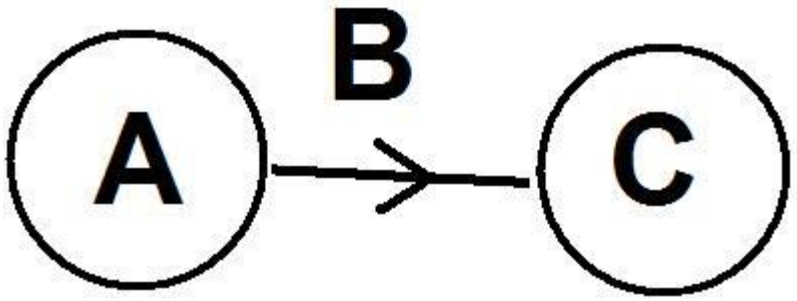
= there is no permutation that reduces the adjacency matrix in diagonal blocks.

$$PAP^{-1} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

If a graph is connected and undirected each node can be reached by any other

Directed case

$$PAP^{-1} = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$



In this case the blocks are not separated, but only one is reachable from the other ("sink", "absorbing state" in transition processes).

In a direct connected graph it is not guaranteed that all the nodes can be reached by all the others

Matrices, blocks, algebra

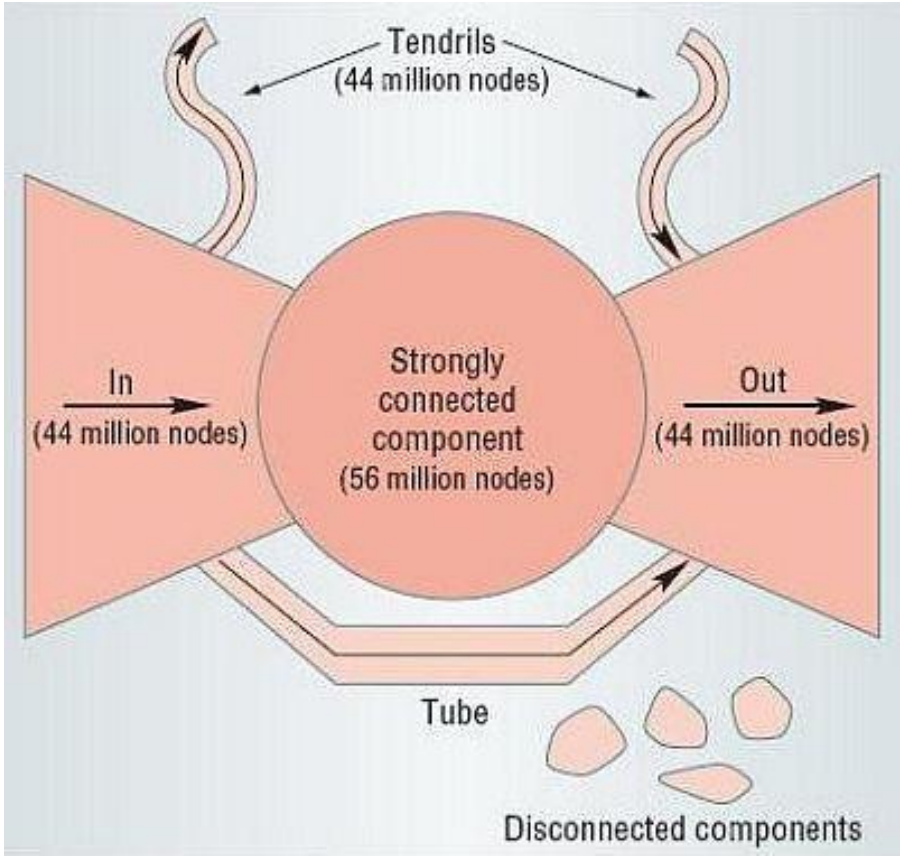
In cases (direct and non-direct) in which the adjacency matrix is divisible in blocks (diagonal or "triangular") it is said that the matrix is **reducible**

WWW (1999)

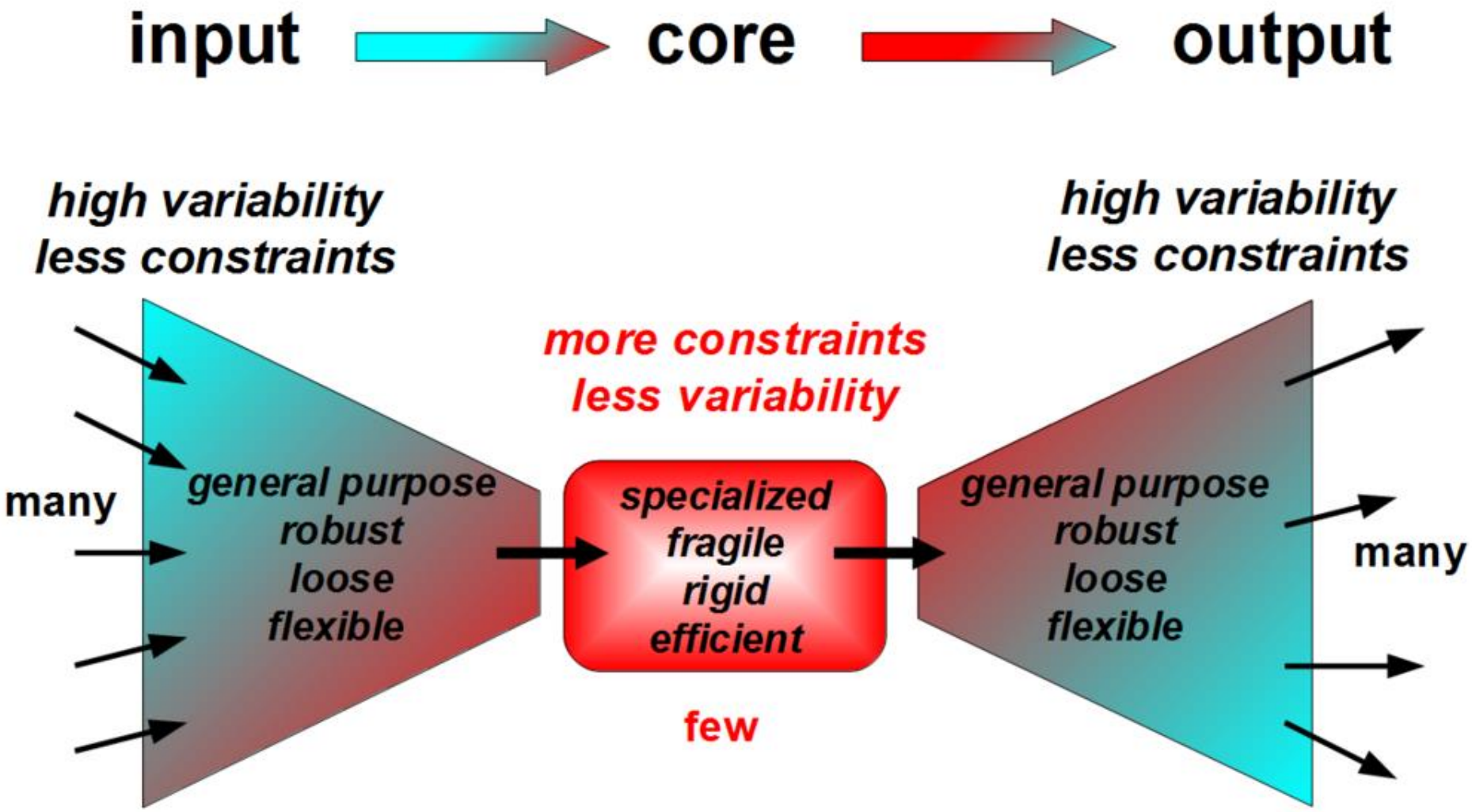
bow-tie structure:

$$\mathrm{PAP}^{-1} \approx \begin{pmatrix} 0 & In \\ Out & SCC \end{pmatrix}$$

$$\mathrm{PAP}^{-1} \approx \begin{pmatrix} SCC & Out \\ In & 0 \end{pmatrix}$$



Bowtie architecture



Connectedness (digraph)

1) strongly connected: **every** node is connected to another by a path -> cycles exist

2) unilaterally connected: for every two node there is **at least** one path

3) weakly connected: every node connected to another by at least a **semipath**

Theorem on connected graphs

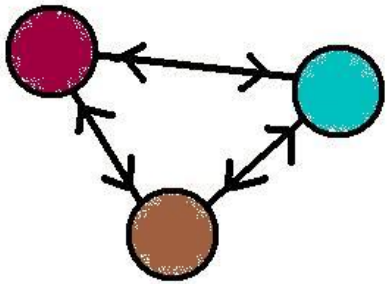
Strongly connected: exists one spanning closed walk

Unilaterally connected: exists one spanning walk

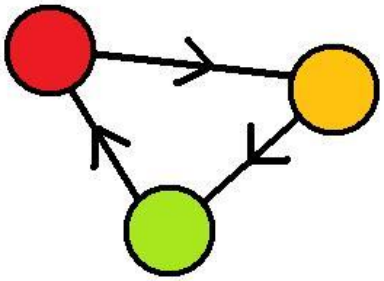
Weakly connected: exists one spanning semiwalk

Examples

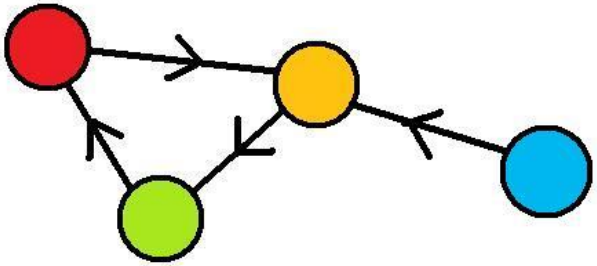
Strongly connected



Unilaterally connected



Weakly connected



Node Distance - unweighted

Distance = **shortest path** between 2 nodes

Isolated nodes: infinite distance

Adjacent nodes: distance = 1

D(u,v) = 0 if u=v

On a symmetric graph it is a metrics:

1) Semidefinite positive
2) Symmetric
3) Triangular inequality

Node Distance - weighted

Given the weights w_{ij}, if I consider closer nodes with a greater weight (highway with "w" lanes) I can use the harmonic composition (capacitors in series)

$$d_{ij} = \sum_{i \to j} \frac{1}{w_{kl}}$$

$$w = 0 \Rightarrow d = \infty$$

Example:

the Euclidean distance between elements (in a defined space) can be seen as a fully connected symmetric weighted network

(eg. protein contact maps)

big weight <-> small distance

$$d_{ij} = ||\vec{x}_i - \vec{x}_j||$$

$$w_{ij} = \frac{1}{d_{ij}}$$

Shortest path

Distance between 2 nodes = number of links that I have to cross (**geodesic**)

Diameter of a graph = maximum geodesic

Average path length = average on all paths

... how do you calculate? Dijkstra's algorithm

(Connectivity) Degree

Degree K = sum of incoming (outgoing) links

Undirected graph: $\Sigma K = 2 \cdot \#\text{links}$

DiGraph: $\Sigma K_{IN} = \Sigma K_{OUT} = \#\text{links}$

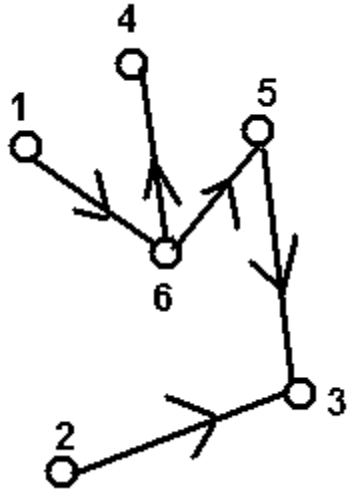
In a graph with N nodes: $0 \leq k \leq N-1$

Regular graphs: all nodes have same degree

Isolated node: K = 0

Endpoint (pendant node): K = 1

Example: directed



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

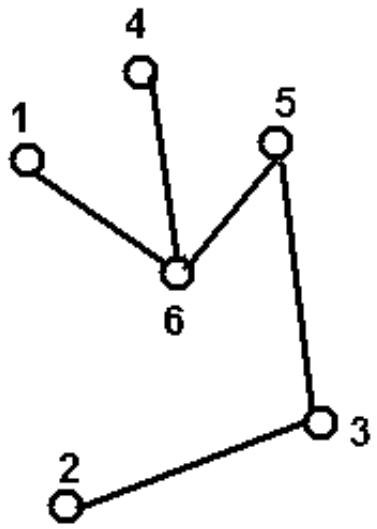
$$K_{IN} = \sum_j a_{ij} = \begin{pmatrix} 0 & 0 & 2 & 1 & 1 & 1 \end{pmatrix}$$

-> weakly connected

$$K_{OUT} = \sum_i a_{ij} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 2 \end{pmatrix}$$

$$\sum K_{IN} = \sum K_{OUT} = L$$

Example: undirected



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

$$\sum_i a_{ij} = \sum_j a_{ij}$$

$$K_{IN} = K_{OUT} = \begin{pmatrix} 1 & 1 & 2 & 1 & 2 & 3 \end{pmatrix} = 2 \cdot L$$

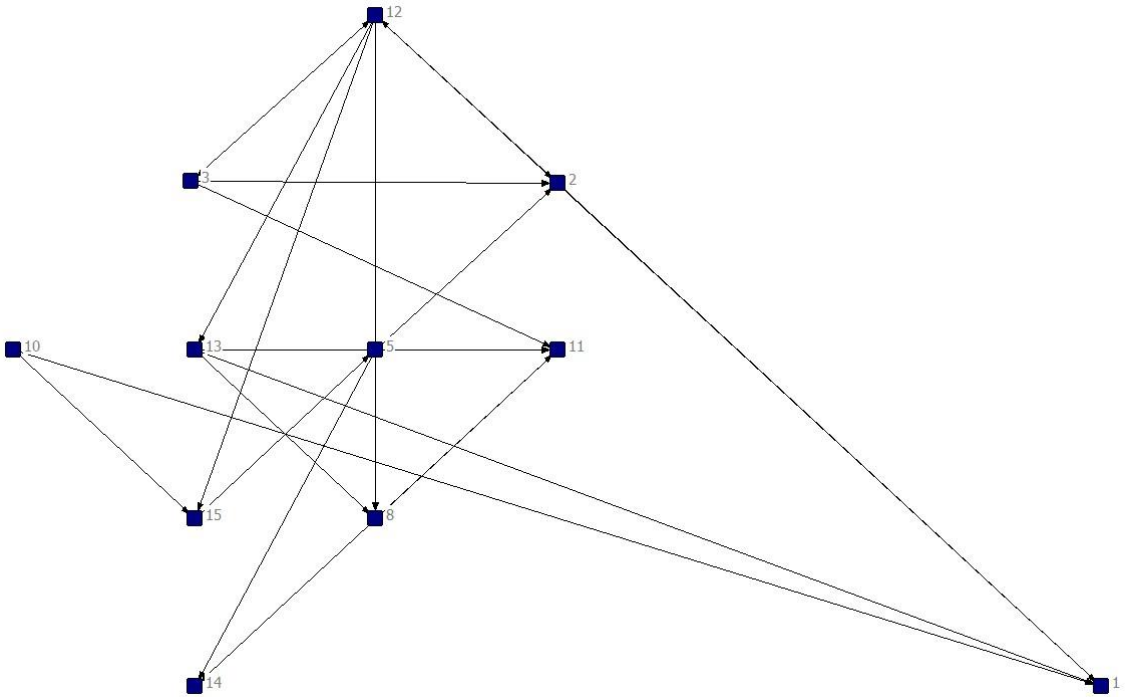
$$P_{2 \to 1} = P_{1 \to 2} = 4 = diameter$$

Directed graph visualization

Xy coordinates: x= incoming, y= outgoing

More nodes with same Kin Kout -> add noise

Hierarchical: node stratification (& 2d clustering)



Network software

Cytoscape
Pajek
Matlab
Mathematica
Python
R (Rgraphwiz etc)
Ucinet

http://science.sciencemag.org/content/323/5916/892
https://sites.google.com/site/ucinetsoftware/home
https://github.com/EnricoGiampieri/PLS2017BigDataNetworks/blob/master/Lezione%203%20-%20Network%20Biologici.ipynb

Graph

Completely connected graph K_N: has all links

Degree K=N-1 (if I have N nodes)

Link number = $$\binom{N}{2} = \frac{N!}{2!(N-2)!} = \frac{N(N-1)}{2}$$

Graph complement: same nodes, “opposite” links
(adjacency matrix+1 mod(2))

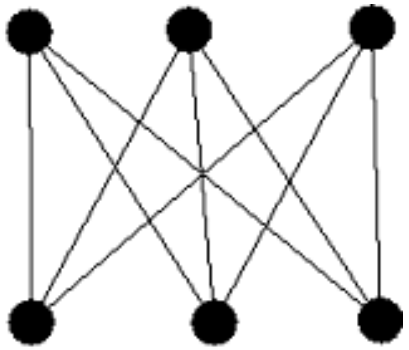
Self-complementary = isomorphic to its complement

Bipartite graph

Graph that connects **2 different types of nodes** to each other

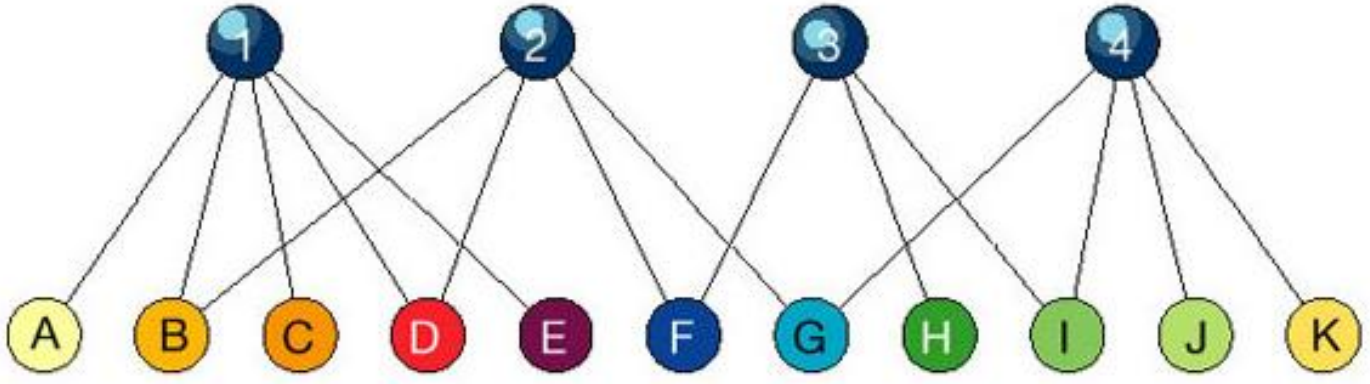
Eg: films/actors, papers/authors, gene/disease, noun/adjective

M nodes of type 1, N nodes of type 2



MxN rectangular adjacency matrix

Complete Bigraph KMN = M · N links



Bipartite example: diseasome

OMIM = Online Mendelian Inheritance in Man

Bipartite graph gene-pathology MxN

M = 1284 genes, N = 1777 pathologies

"To which pathology are one (or more) genes associated?"

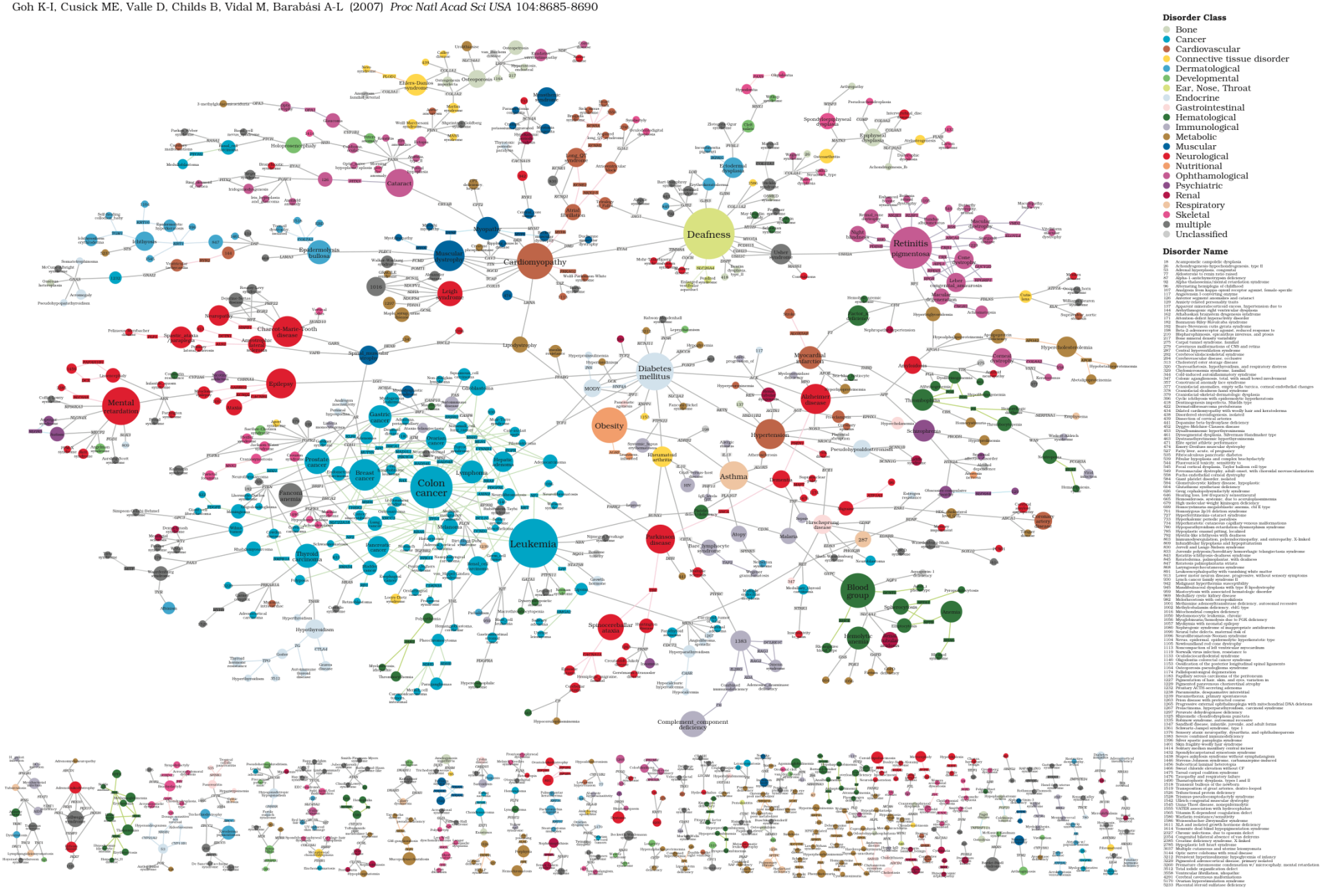
O_{ij} = i-gene j-pathology

I can get (by contraction)

1) a 1284x1284 gene-gene network
2) a pathology-pathology network 1777x1777

The human disease network

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) *Proc Natl Acad Sci USA* 104:8685-8690

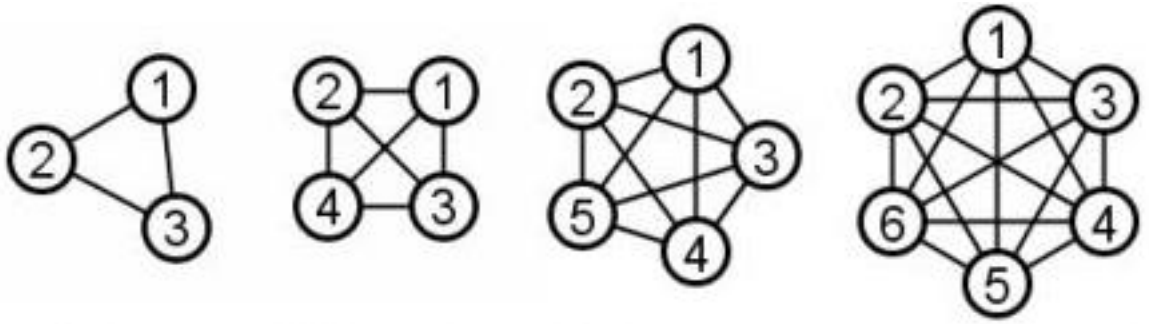


Supporting Information Figure 13 | Bipartite-graph representation of the diseasome. A disorder (circle) and a gene (rectangle) are connected if the gene is implicated in the disorder. The size of the circle represents the number of distinct genes associated with the disorder. Isolated disorders (disorders having no links to other disorders) are not shown. Also, only genes connecting disorders are shown.

K-Clique

Fully connected G subgraph (with degree K)

Finding a clique of a size in a graph (the clique problem) is NP-complete.



K-clique – community

A strongly connected subgraph can be seen as a community within the network, ie its nodes are likely to share similar properties.

A clique can represent a N-body relationship, when a link is a 1 to 1 relationship

example: protein complexes, chemical reactions

Clan, Plex, Core

N-cliques: nodes connected by a path of maximum length N (typically = 2).

N-clans: restriction of N-cliques, with the constraint that the diameter is less or equal to N.

K-plexes: at least N - K links with the other N-1 nodes, where N is the total number of nodes in the group.

K-cores: nodes connected with at least K other nodes of the group

NOTE: calculable for small networks (used in sociology)

Erdos-Renyi ”random” graphs

Networks whose interactions (links) between nodes are chosen randomly (without a priori correlations)

Null model for many problems (but not all)

Graph statistical mechanics - "perfect gas" with random impacts

Degree pdf in $G_{N,P}$

N nodes connected with probability P:

The probability of having a node with k links is:

$$p(k) = \begin{pmatrix} \text{N-1} \\ \text{k} \end{pmatrix} p^k \cdot (1-p)^{N-1-k}$$

The average connectivity is:

$$\langle k \rangle = \frac{\langle \# \, link \rangle}{N} = \frac{2 \cdot P \dfrac{N(N-1)}{2}}{N} = P(N-1) \approx P \cdot N$$

Network properties

- whole network properties
  - #nodes/edges, link density
  - diameter, girth, genus, clustering coefficient
  - spectral properties (eigenvalues/eigenvectors)
  - # cycles & motifs
  - modules, clusters
  - assortativity/mixing

- properties of single nodes/edges
  - centrality measures
  - statistical distribution of parameters (1-dim)
  - relationships between parameters (pdf N-dim)

nodes/links properties : centrality

There are parameters that can be associated to each single node or link, or to groups of them.

They provide information on the relevance of network elements, and are typically characteristic of the type of network considered

- useful for ranking and stratification of nodes

(Connectivity) Degree

LOCAL parameter for its calculation

Number of incoming / outgoing links from a node

Sum for rows / columns of the adj. mat.

$$K_{IN} = \sum_{j} a_{ij}; \qquad K_{OUT} = \sum_{i} a_{ij}$$

Very connected = very important

(eg website visited, book / article read / quoted)

Eigenvector centrality (from adjacency matrix)

This greatest eigenvalue results in the desired centrality measure. The v th component of the related eigenvector then gives the relative centrality score of the vertex v in the network.

Degree distribution

The distribution of connectivity of all nodes provides important information on network properties.

In particular, the heterogeneity of K values can characterize the type of network: (stochastic processes)

PPI network

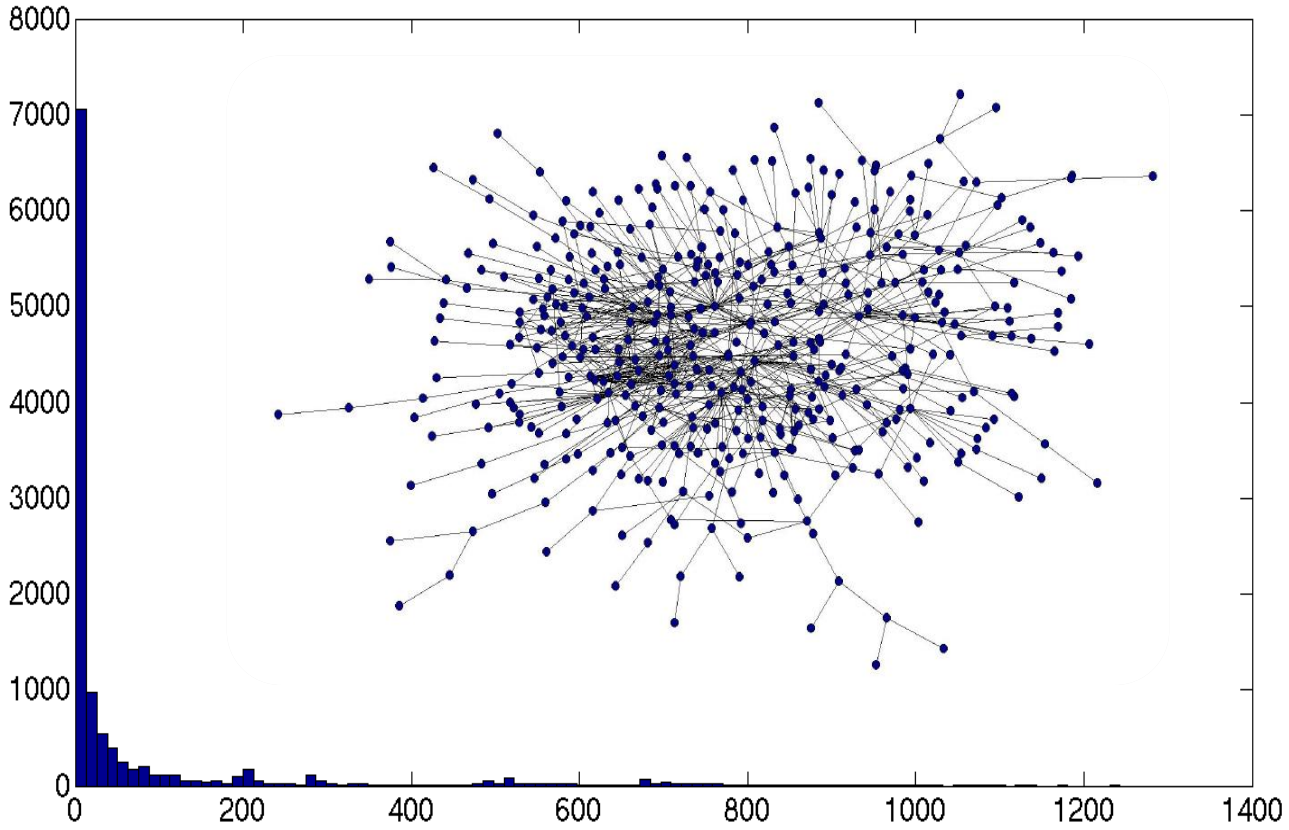
Human Protein Interaction: PathwayCommons

hierarchical network

>11000 nodes

>420000 links

fat-tailed p(k)



Scale-free network

Networks with strongly heterogeneous p (K) connectivity distribution (heavy tail pdfs)

$$p(k) \propto k^{-\alpha}$$

it is not a single class (like E-R)

various scale-free network models may differ for other parameters (eg C, BC vs K & assortativity)

the network building process changes

“Scale-free” network: examples

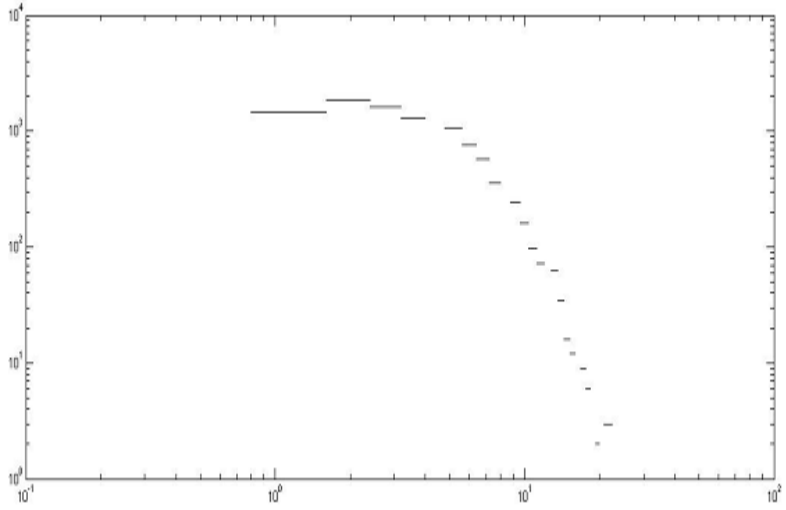
Important because observed in real systems:

technological: WWW, air transports

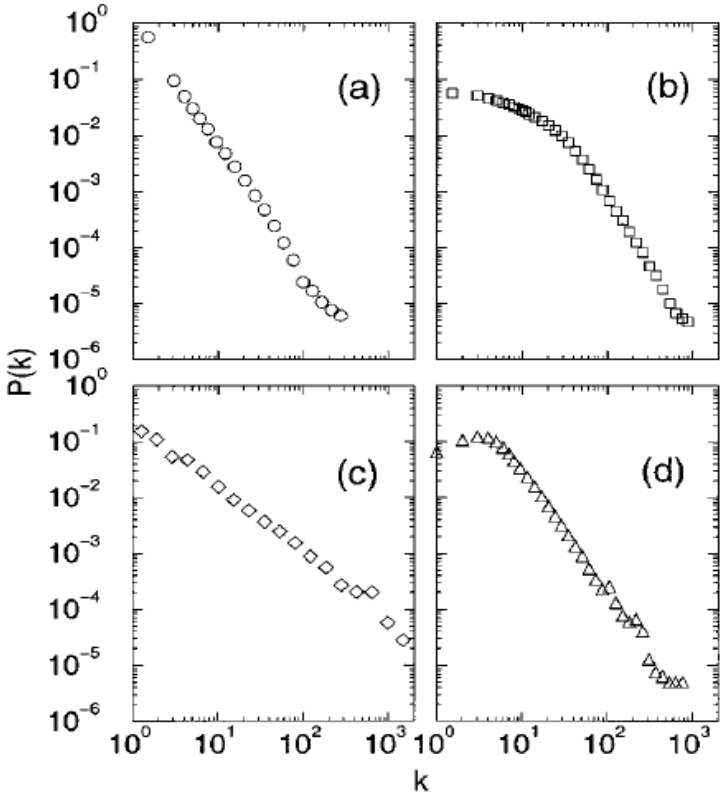
social: citation network, sexual contacts, calls/email exchange, facebook ”clicks”

biological: “omic” networks (protein, gene, metabolite, DNA contacts)

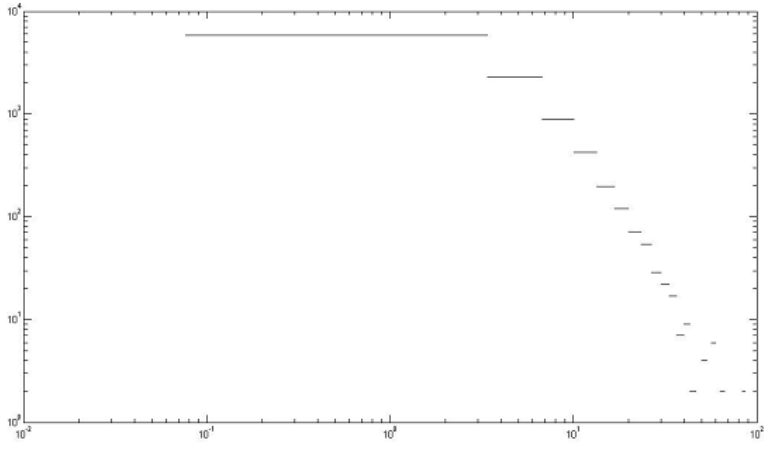
Chi square: ν=4



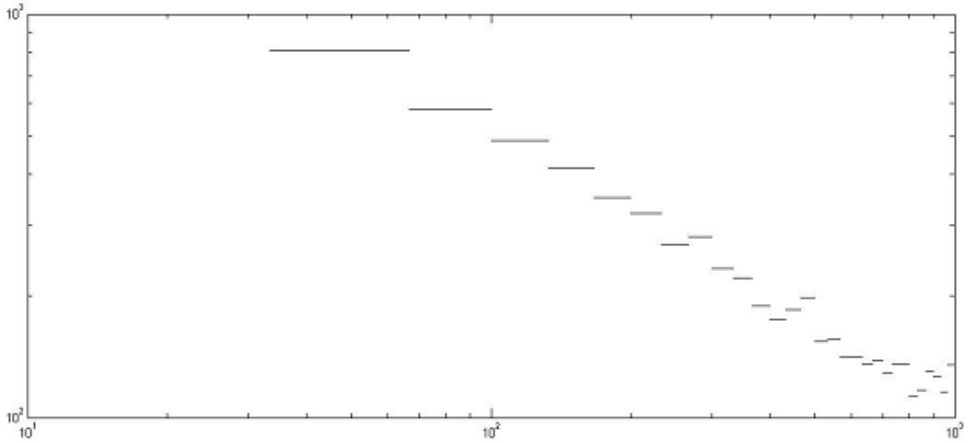
Real



Lognormal



Power-law



Heavy tail distributions

Lognormal

$$p(x) = \frac{1}{x\sqrt{2ps^2}} e^{-\frac{(\log x - m)^2}{2s^2}}$$

Gamma

$$p(x) = \frac{1}{G(k)q^k} x^{k-1} e^{-\frac{x}{q}}$$

Weibull / Stretched_exponential

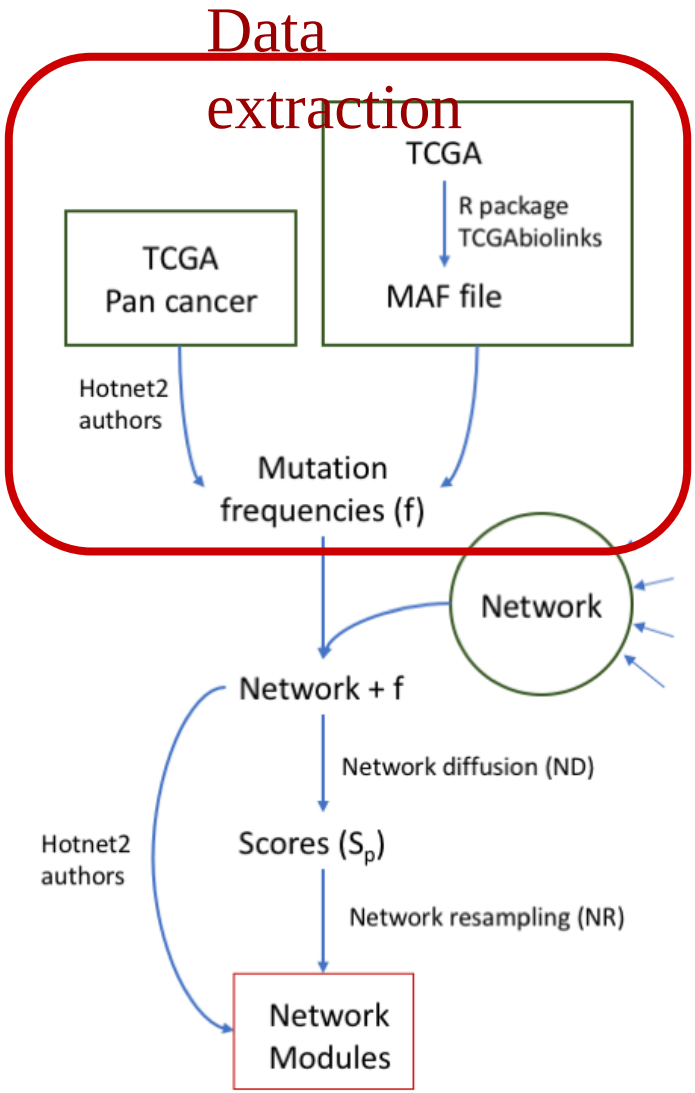
$$p(x) = \frac{k}{l}\left(\frac{x}{l}\right)^{k-1} e^{-\left(\frac{x}{l}\right)^k}$$

Focus on mutational data, as obtained fron SNP or WES

MAF files can be downloaded from TCGA in an automatized way by using a R software

Aims

- Explore possible analysis pipelines using network-based gene module finding tools, using publicly available genomic datasets
- Compare different approaches and different protein-protein interaction network

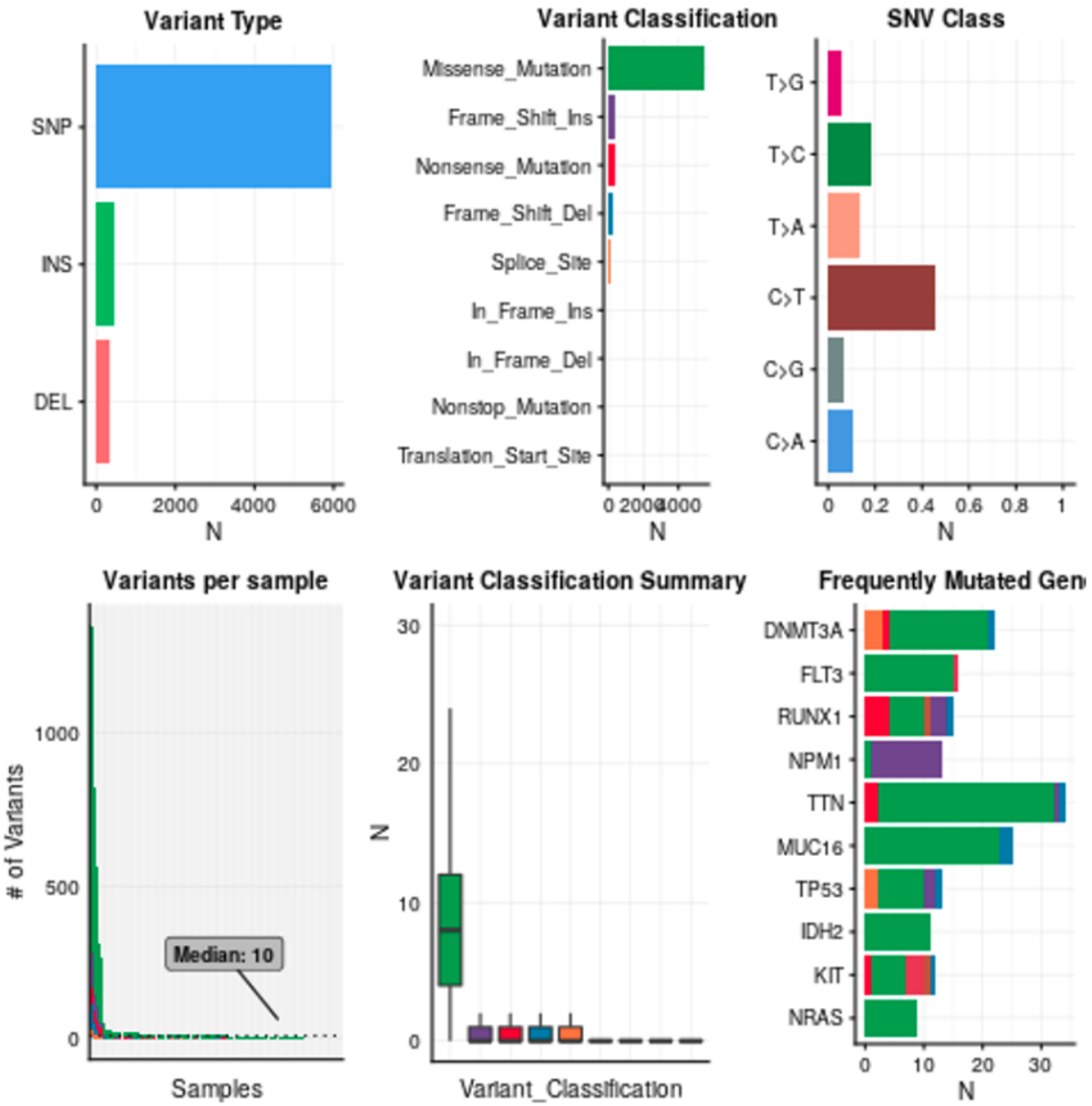


Methods: Data Preprocessing

1. Download MAF file from TCGA database

1. Filtering outliers

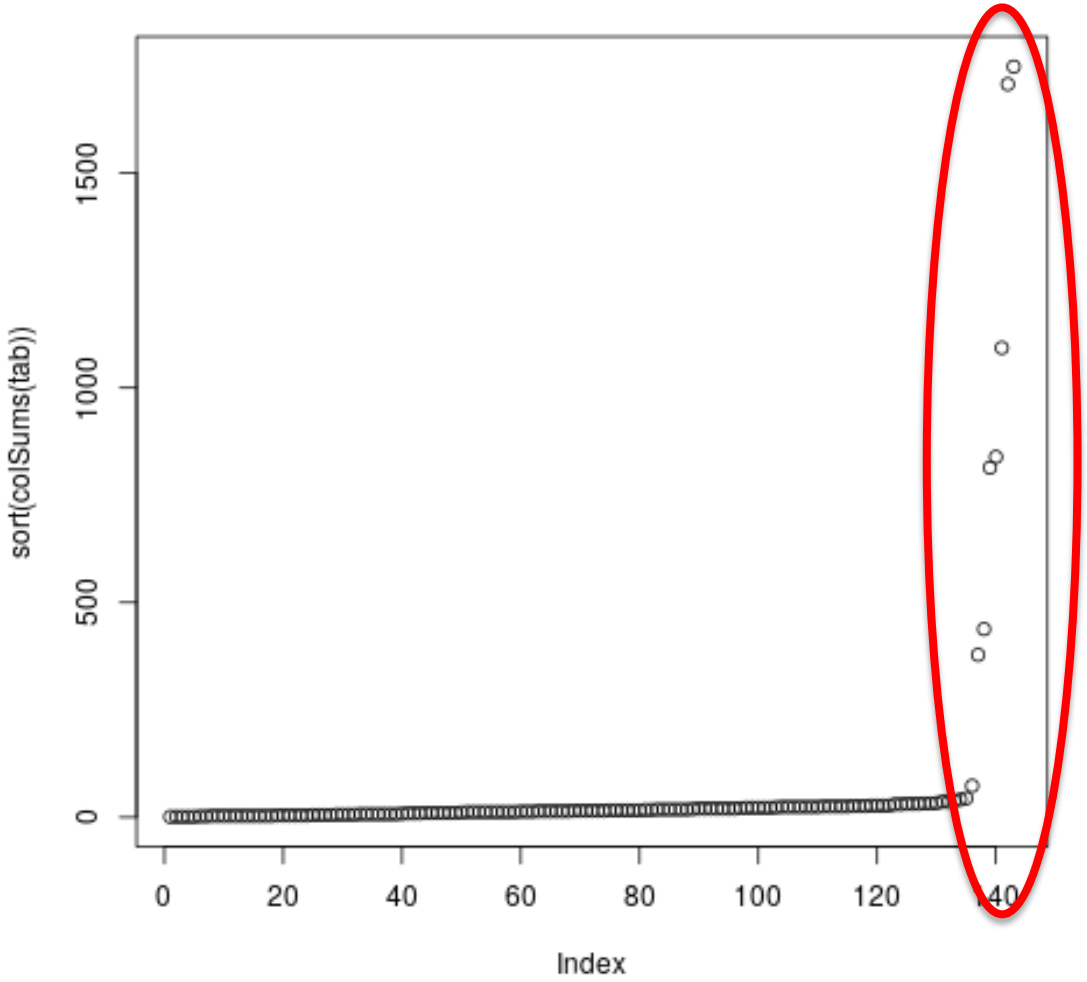
1. Creating vector of mutation frequencies per gene within all samples



Methods: Filtering

1. Filtering off the outlier samples

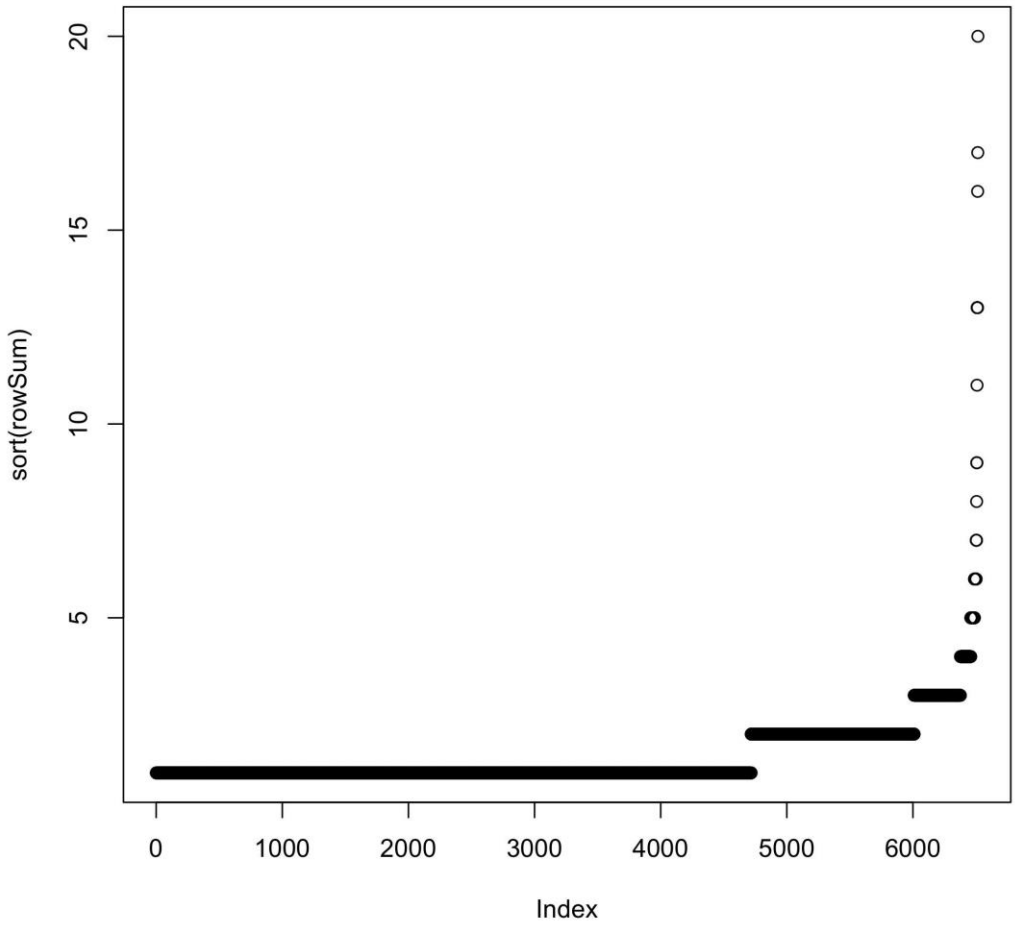
1. Filtering off *intergenic* mutations



Methods: Input

Frequency:

Number of times that the gene is mutated across the samples / number of samples



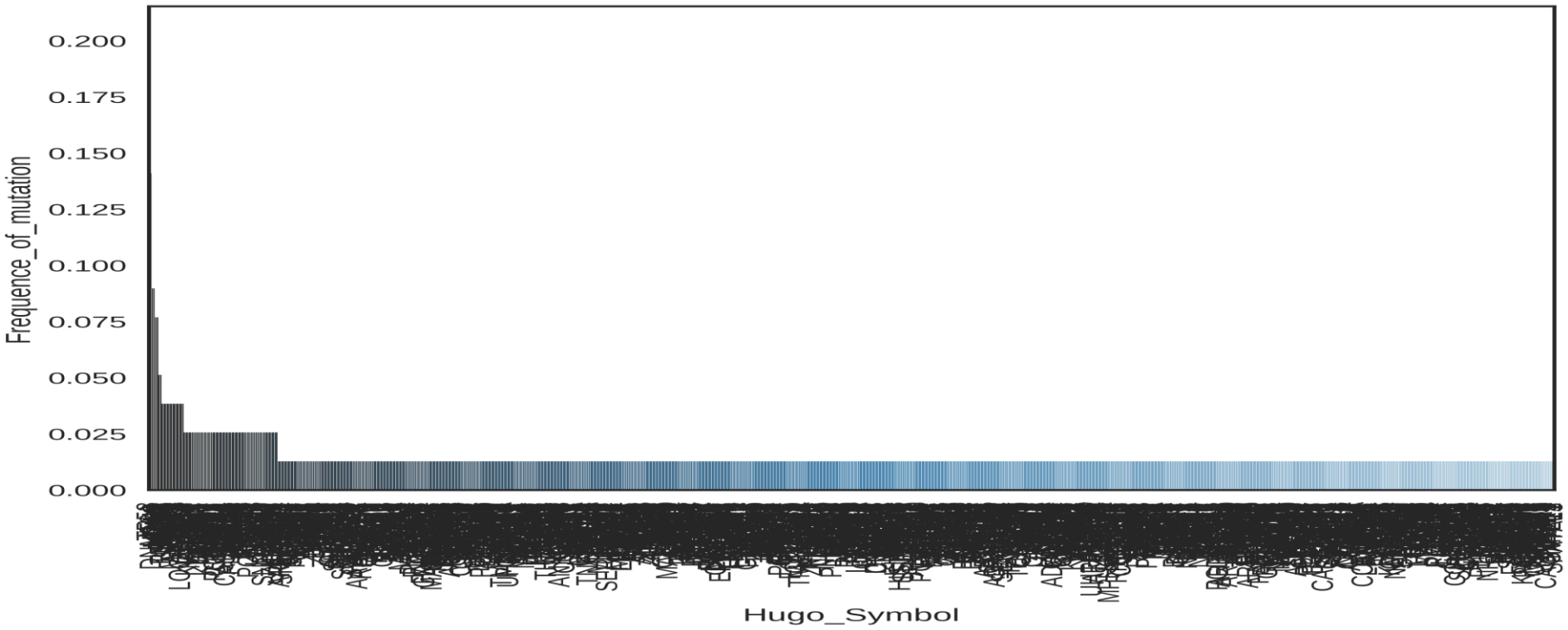
G_id	freq	Symbol
1788	0,14	DNMT3A
2322	0,12	FLT3
4869	0,09	NPM1
7157	0,09	TP53
3418	0,08	IDH2
861	0,08	RUNX1
94025	0,07	MUC16
3815	0,07	KIT
4893	0,06	NRAS
3417	0,05	IDH1
7490	0,05	WT1
3845	0,04	KRAS
7273	0,04	TTN
1050	0,04	CEBPA
:	:	:
:	:	:

The input can be **generic**
Mutation frequency, gene expression values, or in general molecular alterations: (genomic, proteomic, metabolic, etc)

Mol_id	Val	Symbol
M_1	V_1	S_1
M_2	V_2	S_2
M_3	V_3	S_3
:	:	:
:	:	:
M_n	V_3	S_n

The mutation frequency vectors suffer of a well-known pathology (Pareto distribution):

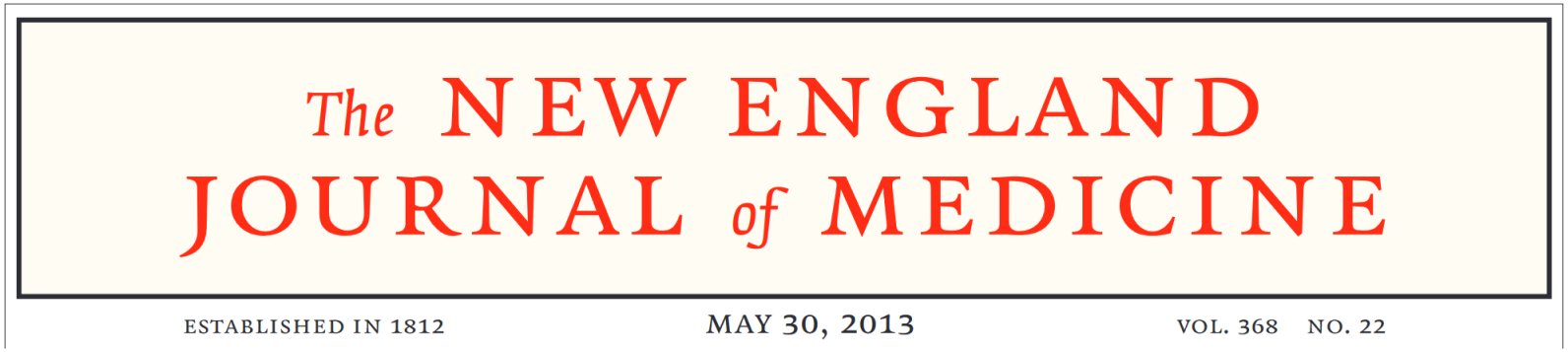
Few genes have a high mutation frequency, while a lot of genes have one very low



Genome Res. 2012 Feb;22(2):375-85.

De novo discovery of mutated driver pathways in cancer.

Vandin F[1], Upfal E, Raphael BJ.



Nat Genet. 2015 Feb;47(2):106-14.

Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.

Leiserson MD[1], Vandin F[1], Wu HT[1], Dobson JR[2], Eldridge JV[3], Thomas JL[3], Papoutsaki A[3], Kim Y[3], Niu B[4], McLellan M[4], Lawrence MS[5], Gonzalez-Perez A[6], Tamborero D[6], Cheng Y[7],

www.nature.com/scientificreports

OPEN

Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules

Received: 07 January 2016
Accepted: 19 August 2016
Published: 12 October 2016

Matteo Bersanelli[1,*], Ettore Mosca[2,*], Daniel Remondini[1], Gastone Castellani[1] & Luciano Milanesi[2]

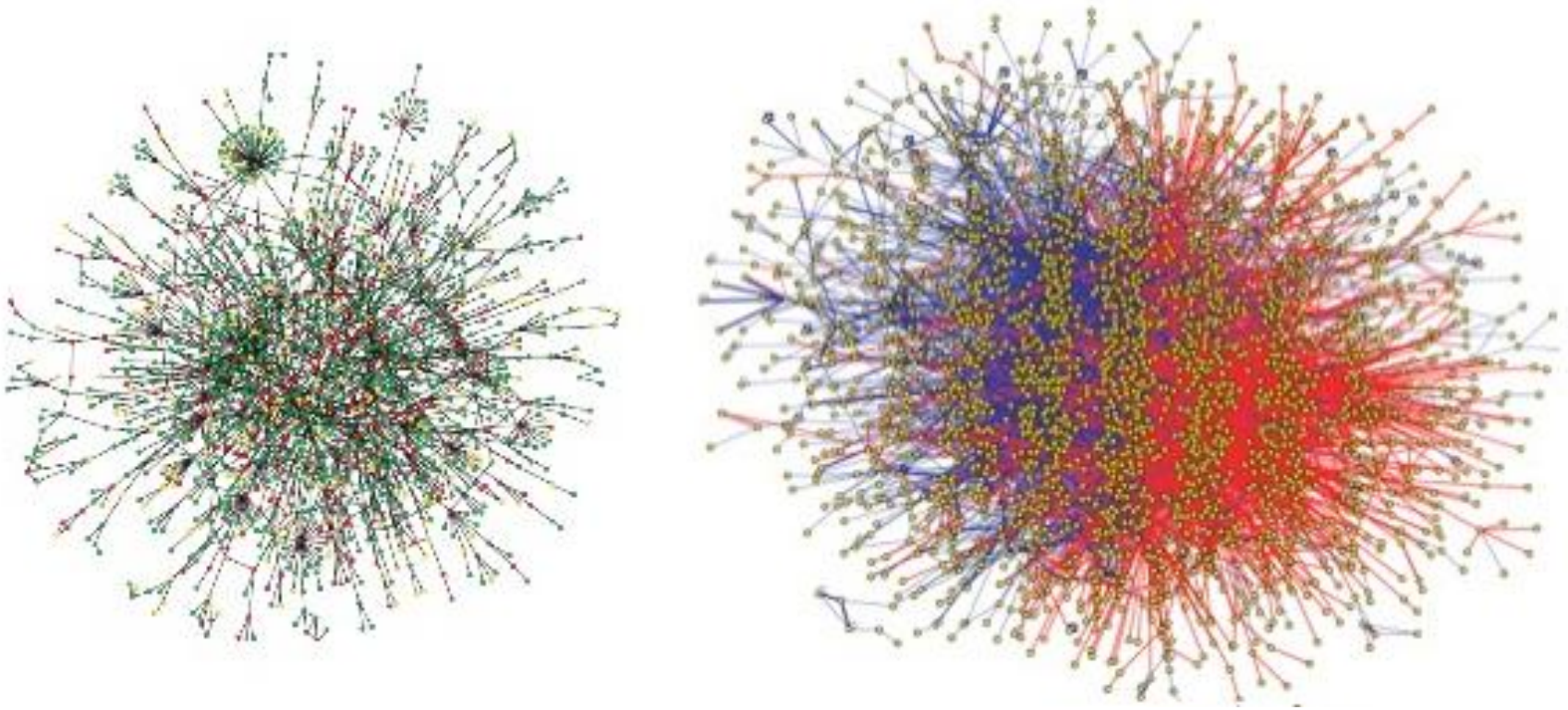
ORIGINAL RESEARCH
published: 25 September 2017
doi: 10.3389/fgene.2017.00129

Network Diffusion-Based Prioritization of Autism Risk Genes Identifies Significantly Connected Gene Modules

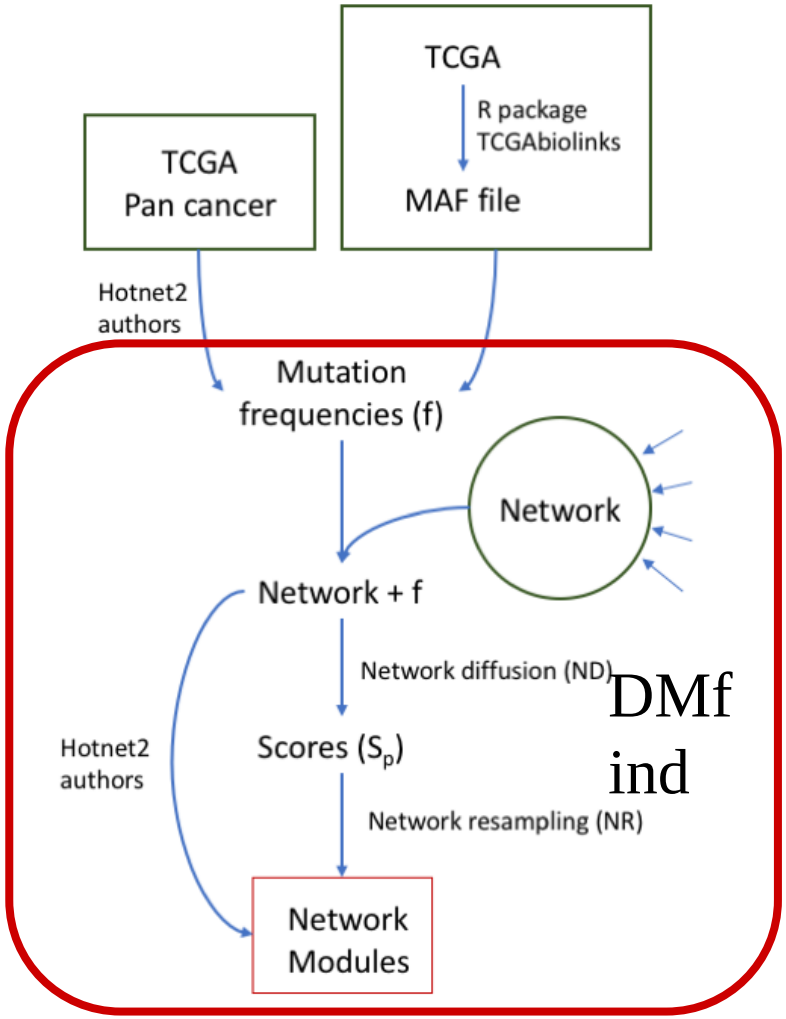
Ettore Mosca[1], Matteo Bersanelli[2], Matteo Gnocchi[1], Marco Moscatelli[1], Gastone Castellani[2], Luciano Milanesi[1] and Alessandra Mezzelani[1]*

[1] Bioinformatics Group, Institute of Biomedical Technologies, National Research Council of Italy, Segrate, Italy, [2] Applied Physics Group, Department of Physics and Astronomy, University of Bologna, Bologna, Italy

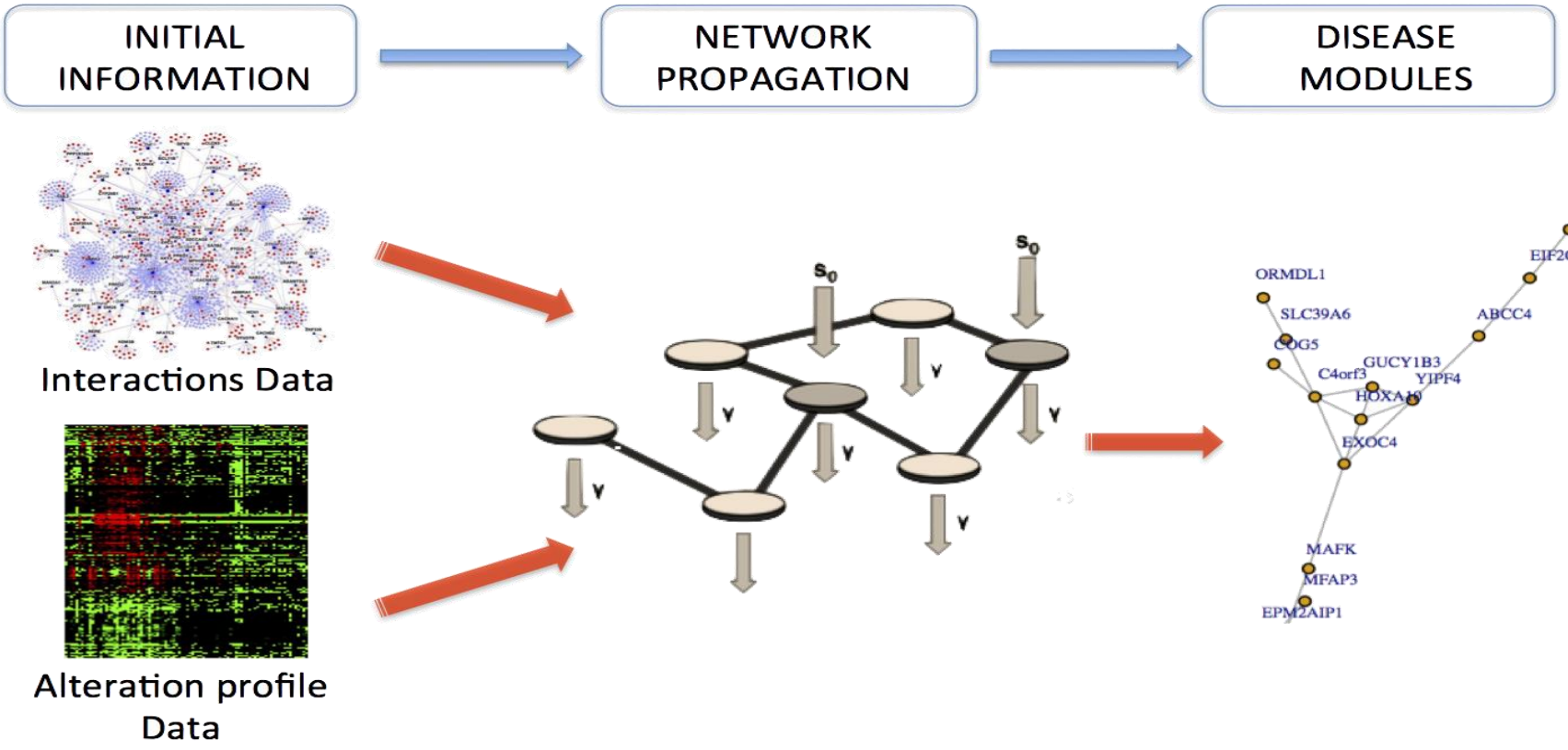
Protein–protein interactions (PPIs) are the physical contacts of high specificity established between two or more **protein** molecules as a result of biochemical events. PPI modulation and properties such as allosteric sites and hotspots, are crucial for drug-design strategies, expecially for cancer treatment. Examples are Titrobifan, inhibitor of the glycoprotein IIb/IIIa, used as a cardiovascular drug, and Maraviroc, inhibitor of the CCR5-gp120 interaction, used as anti-HIV drug.



Workflow



Module identification



TOOLS

- DMfind
- Hotnet2
  - Modified diffusion process similar to DMfind estimates edge weights
  - Threshold edge weights to identify true edges
  - Permutation test to identify significant subnetworks

Network Diffusion

$$\frac{d\vec{P}(t)}{dt} = -\beta_0 \boldsymbol{L}' \vec{P}(t)$$

Network Diffusive Laplacian

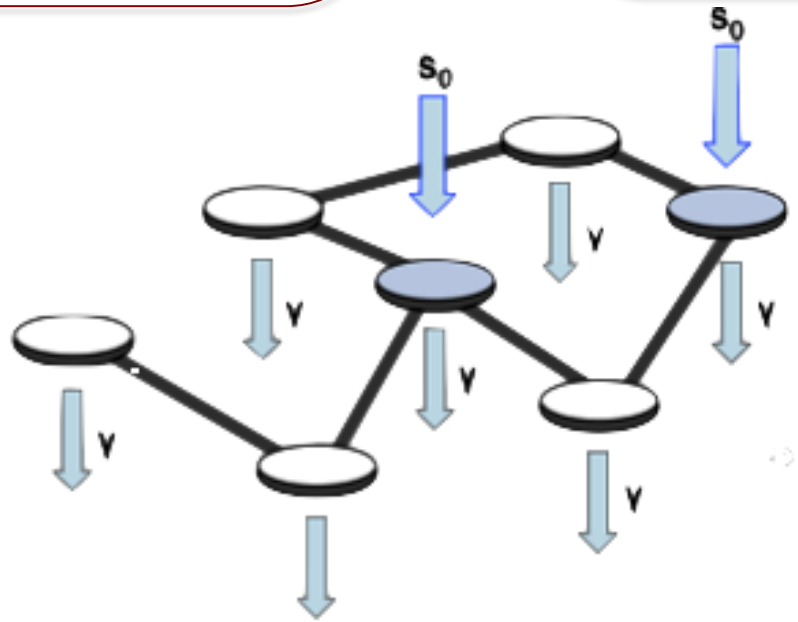
$$\boldsymbol{L}' = (\boldsymbol{D} - \boldsymbol{A})\boldsymbol{D}^{-1} = \boldsymbol{I} - \boldsymbol{\pi}$$

Open Source – Sink Model

$$\frac{d\vec{\phi}(t)}{dt} + \boldsymbol{L}'\vec{\phi}(t) = \gamma_1 \vec{\phi}^0 - \gamma_2 \boldsymbol{\pi}^{out} \vec{\phi}$$

Stationary Distribution

$$\vec{\phi}_{st} = (\boldsymbol{L}' + \gamma_2 \boldsymbol{\pi}^{out})^{-1} \gamma_1 \vec{\phi}^0$$



$$\frac{d\vec{\phi}(t)}{dt} + \boldsymbol{L'}\vec{\phi}(t) = \gamma_1\vec{\phi^0} - \gamma_2\boldsymbol{\pi}^{out}\vec{\phi}$$

$$\boldsymbol{\pi}^{out} = \boldsymbol{I} \Downarrow \gamma_1 = \gamma 2 = \frac{1-\alpha}{\alpha}$$

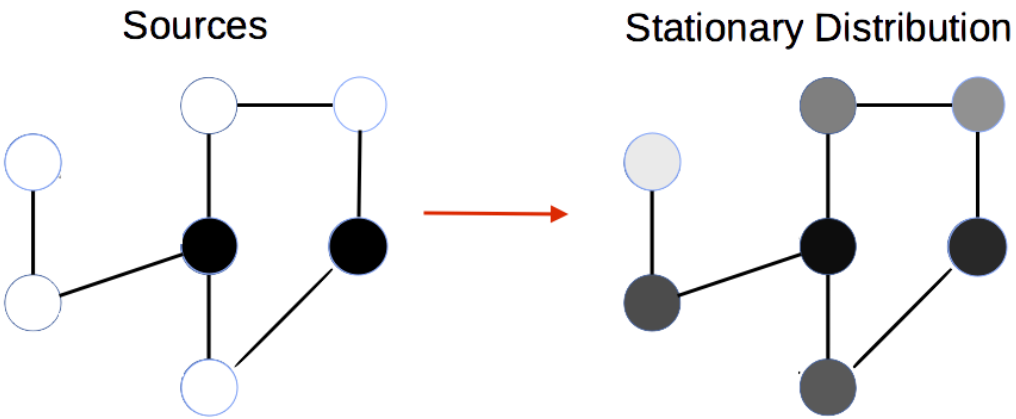
$$\vec{\phi}(t+\Delta t) = \alpha\boldsymbol{\pi}\vec{\phi}(t) + (1-\alpha)\vec{\phi^0}$$

Network Smoothing Index

$$\vec{\phi}(t) \rightarrow \vec{\phi^*}$$

$$S_i(\vec{\phi^0}) = \frac{\phi_i^*}{\phi_i^0 + \epsilon}$$

In order to mitigate the effect of hubs the NSI is adjusted as follows:



$$Sp_i(\vec{\phi^0}) = -log_{10}(p_i)S_i(\vec{\phi^0})$$

where pj is the fraction of times an Sj obtained from the smoothing of a randomised source vector is equal or greater than the real Sj .

Materials

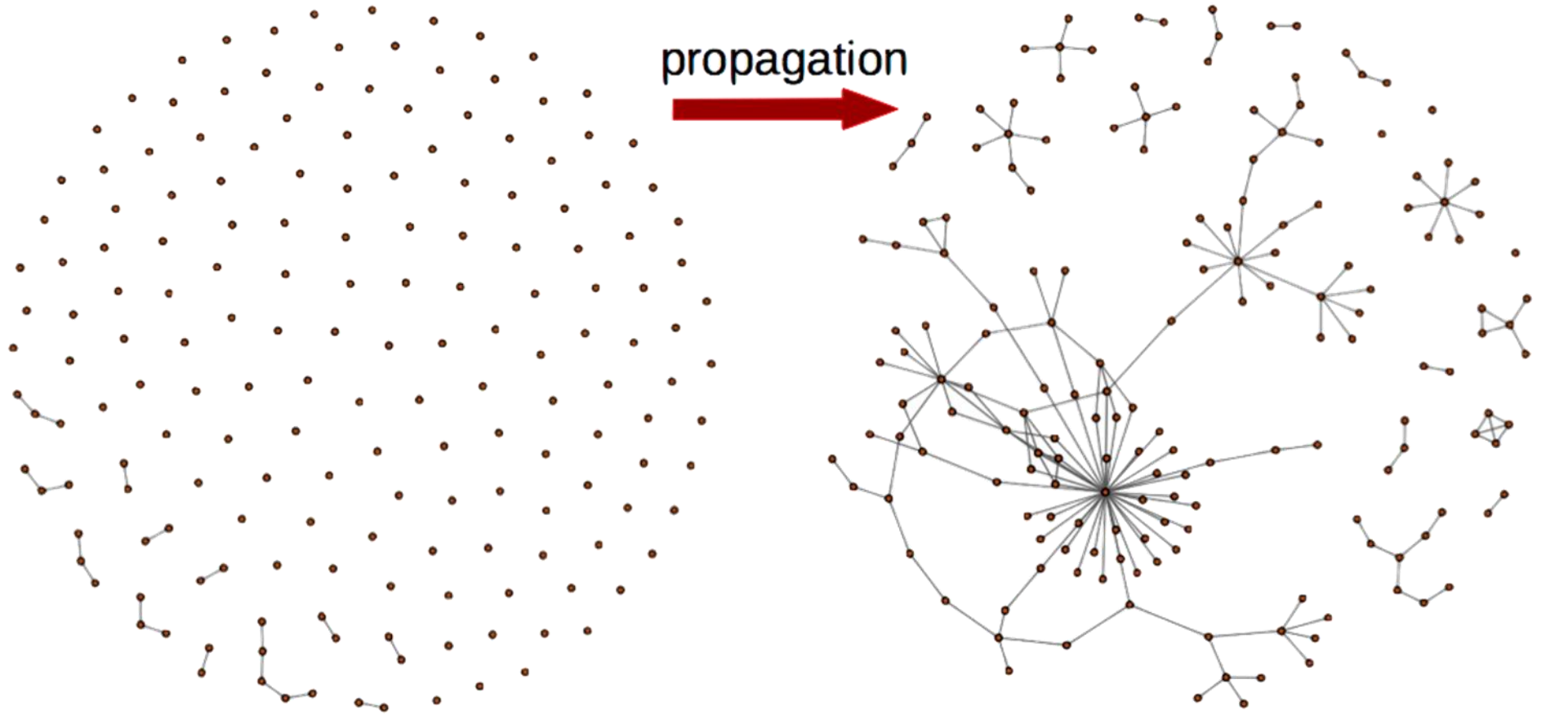
DATASET 1 (TCGA)

- List of genes with Somatic Mutations in Acute Myeloid Leukemia (LAML)
- Consider only mutations from primary tumors
- 6516 genes
- 143 samples

DATASET 2 (UNIBO)

- List of genes with Somatic Mutations in Acute Myeloid Leukemia (LAML) and Gene expression data
- 11572 genes
- 78 samples

THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute


propagation

G_id	Symbol	X0	Sp	mod
5625	TP53	0,21	0,46	1
4746	HNF4A	0,00	0,15	1
4105	FAM111A	0,00	0,13	1
7809	MYCBP2	0,04	0,13	1
3938	PCDHA12	0,03	0,09	1
1276	CUL9	0,00	0,06	1
10694	RORC	0,00	0,06	1
155	EPHA7	0,01	0,05	1
6208	FBXW7	0,00	0,05	1
4745	PCDHA10	0,01	0,05	1
2616	CTNNA3	0,01	0,05	1
2945	AGAP1	0,01	0,04	1
8276	PCDHA3	0,01	0,04	1
8927	EPHA10	0,01	0,04	1
8081	CELA2B	0,00	0,04	1

X0 is the initial mutation frequency

Sp is the score after the propagation

The important mutations are those with small X0 and large Sp

The new list can be validated by queriyng the web with machine learning and intelligent algorithms
To verify if these genes are associated with AML or other HMs.
Moreover is possible to query the web for identify therapies and drugs.
Among the algorithms we are using we can mention:

- Deep learning methods with text analysis
- Network centrality methods (the pagerank algorithm, as used by Google)
- Unsupervised Neural Network

New Results

Network integration of multi-tumour omics data suggests novel targeting strategies

Italo Faria do Valle, Giulia Menichetti, Giorgia Simonetti, Samantha Bruno, Isabella Zironi, Danielle Fernandes Durso, Jose C. M. Mombach, Giovanni Martinelli, Gastone Castellani, Daniel Remondini

doi: https://doi.org/10.1101/146209

Nature Communications 2018 – in press

Public resources: ”virtual” experiments

Transcriptome, Epigenomics, Drugs, Clinical trials, protein structure, …

National Human Genome Research Institute
ENCODE

NCBI

EMBL-EBI
ArrayExpress

RCSB PDB
PROTEIN DATA BANK

NIH
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

THE HUMAN PROTEIN ATLAS

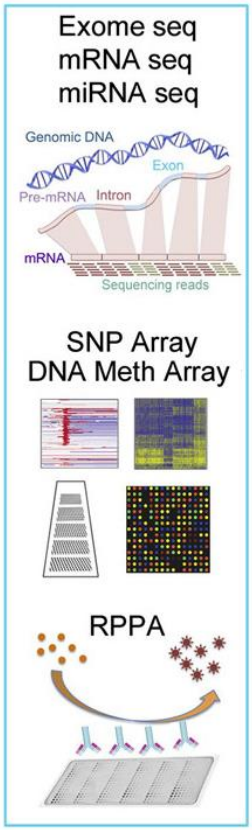
Genomics of Drug Sensitivity in Cancer

1000 Genomes
A Deep Catalog of Human Genetic Variation

EMBL-EBI
IntAct

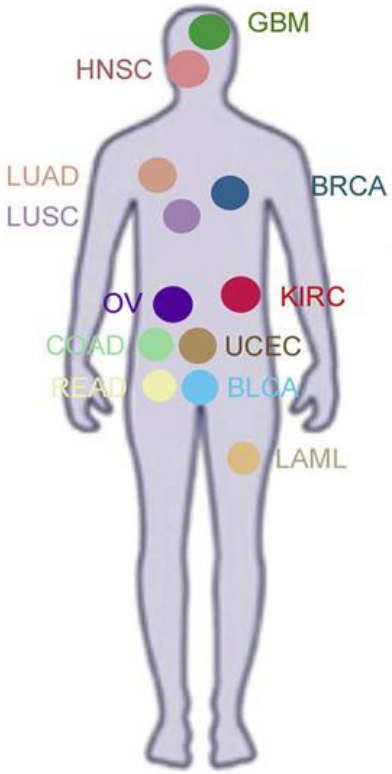
Public data: TCGA database

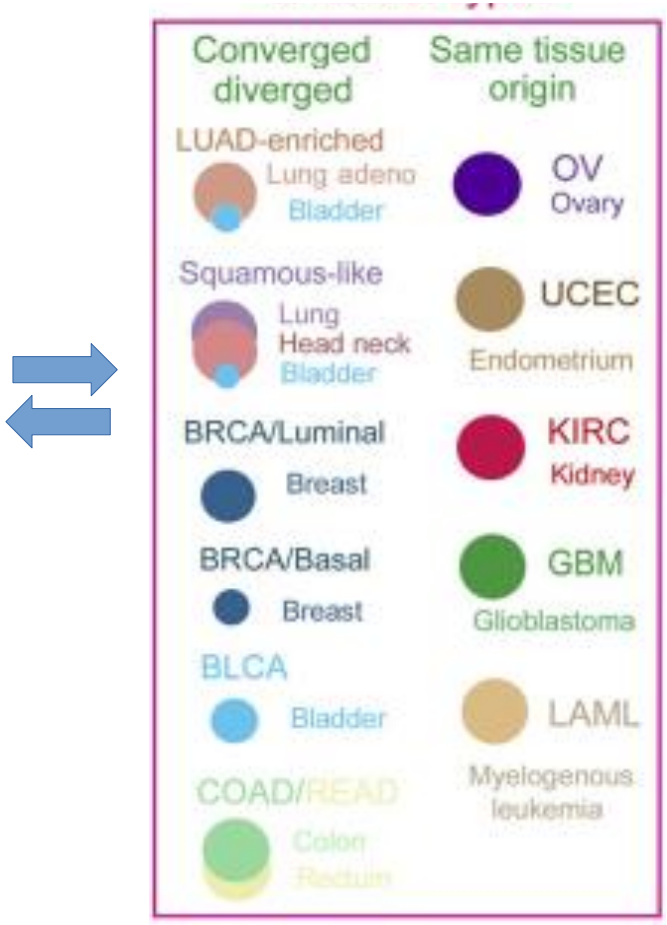
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute



. Cancer types

2000 samples

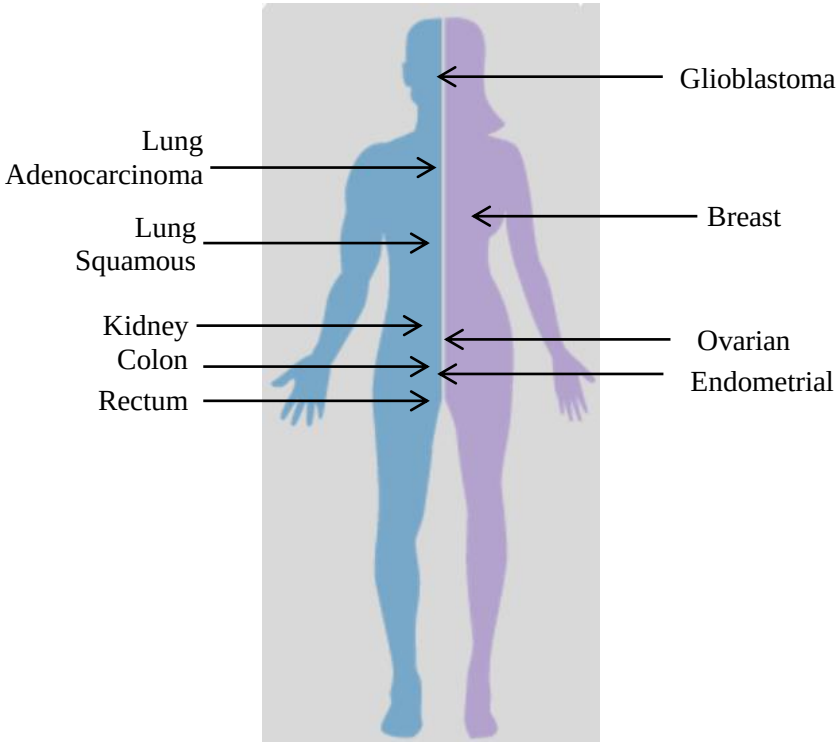




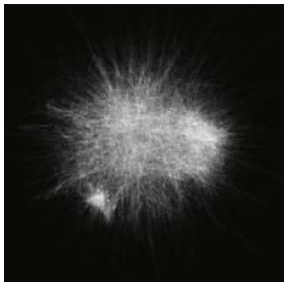
Adapted from Hoadley et. al. (Cell, 2013)

Omics data + Network combination

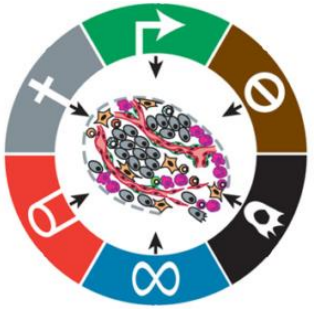
Gene Expression Profiling



PPI network



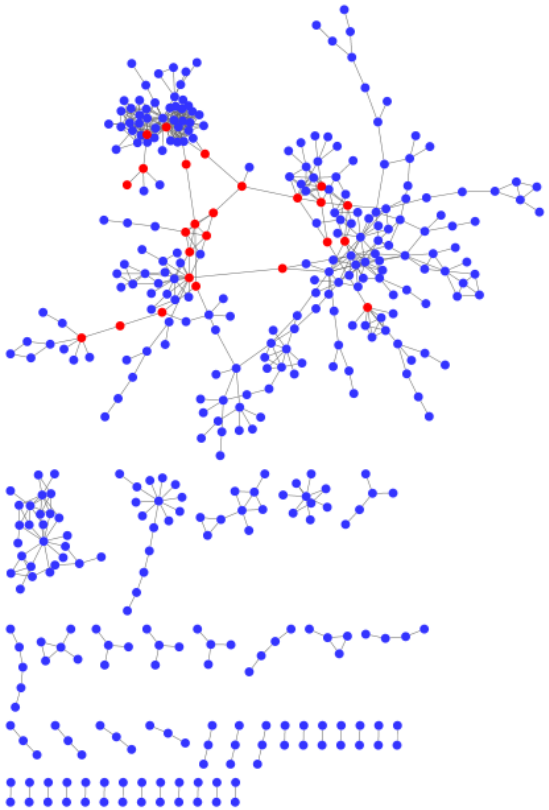
BioPlex
Protein-Protein Interaction Network



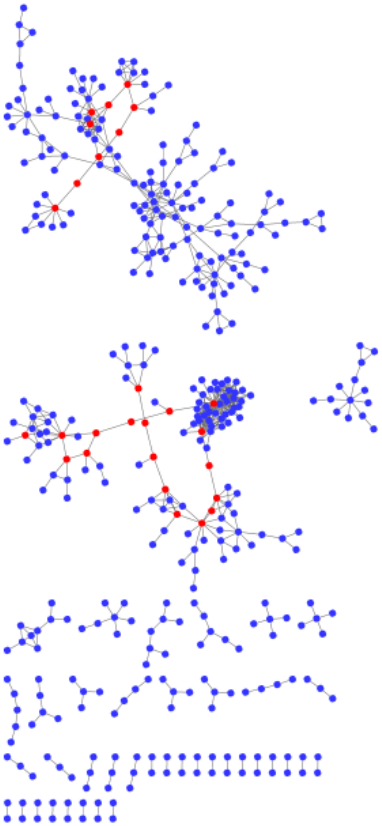
Ontocancro
Genes annotated in cancer-related pathways

AIM: embed transcriptome information into cellular network structure

Network analysis: 3 gene signatures

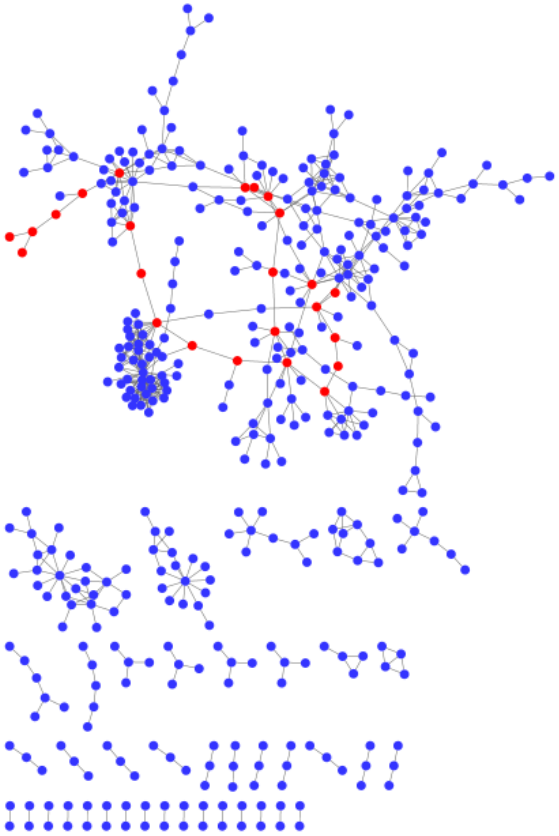


COAD READ



BRCA UCEC OV

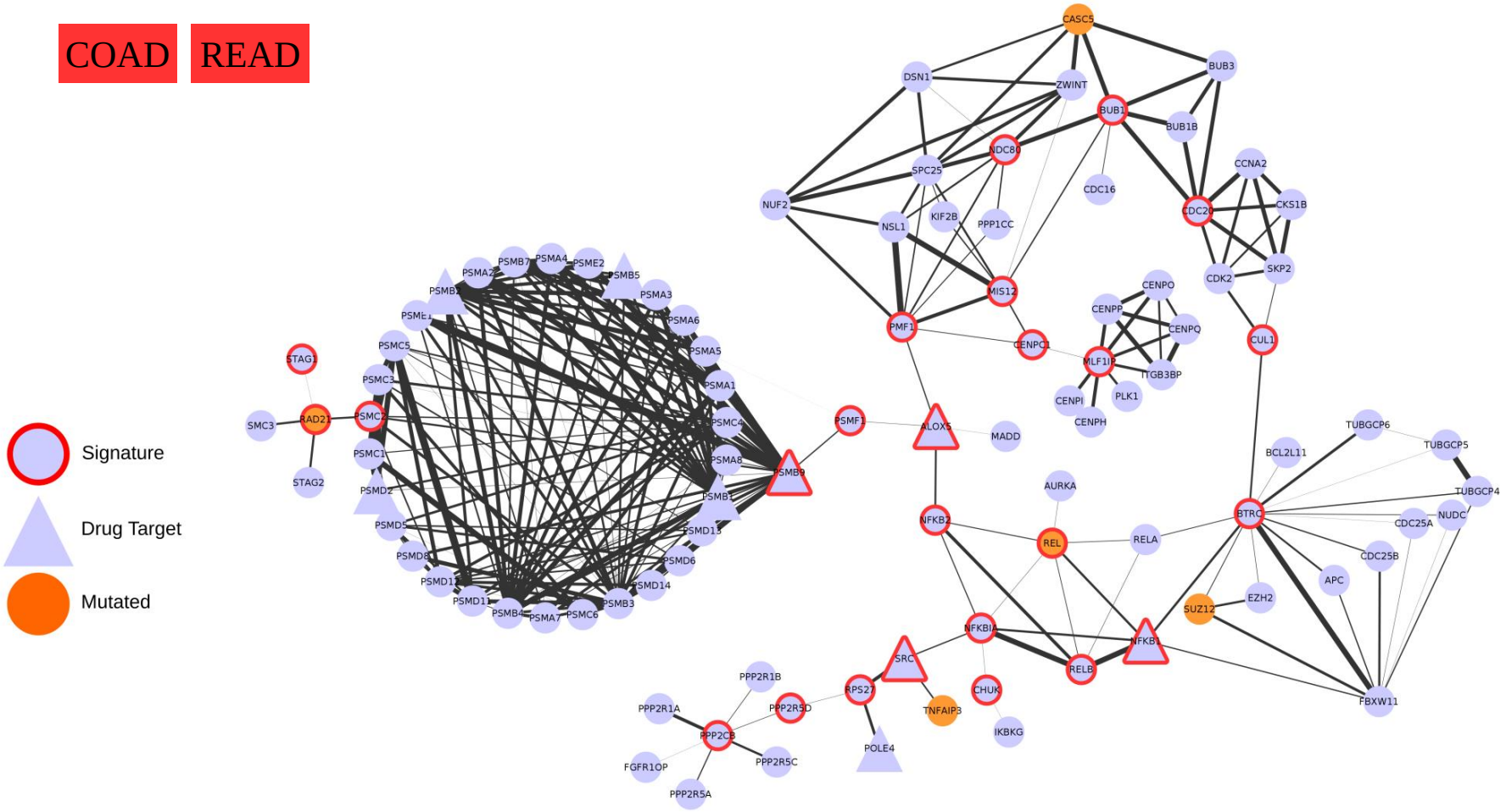
GBM LUAD LUSC



LGG KIRP KIRC

Signature genes, mutations, drug targets

COAD READ



Patient clustering: Survival Outcome

Tumor Clusters

COAD READ

BRCA UCEC OV

GBM LUAD LUSC

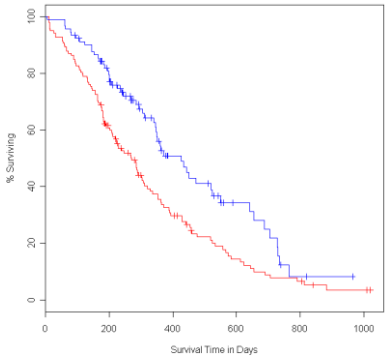
LGG KIRP KIRC

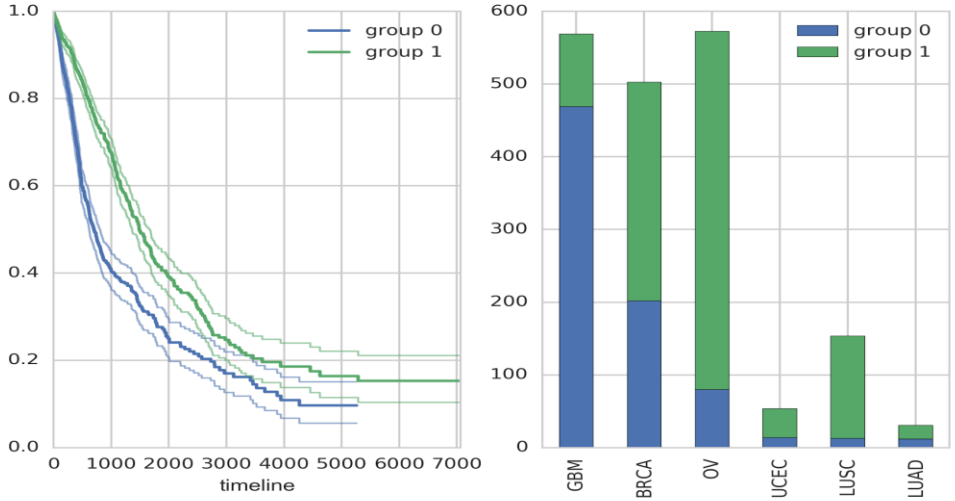
Gene Signatures

Clustering Patients

K-means (k=2)

Survival outcome





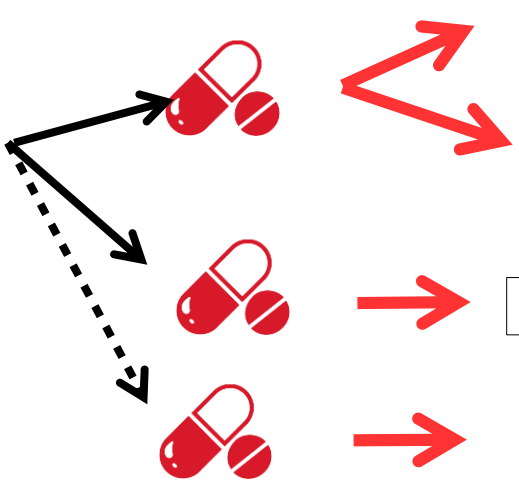
Log-rank test
$P = 7 \times 10^{-18}$

BRCA UCEC OV

GBM LUAD LUSC

Drug targeting: in vitro experiments

BRCA	UCEC
OV	LUAD
LUSC	GBM



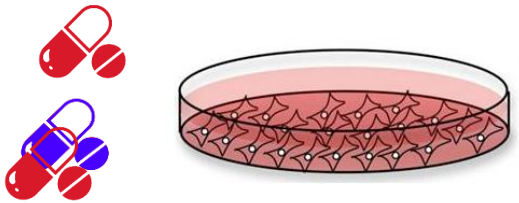
Ubiquitin-Proteasome System — Bortezomib

NFKB Signaling

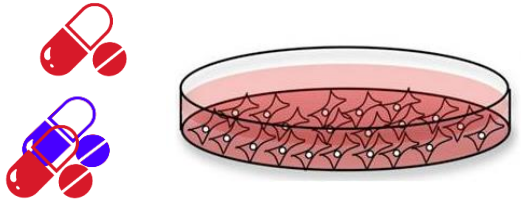
Chromosomal Instability — BI6727

DNA Repair — PF-00477736

3 drugs
2 cell lines



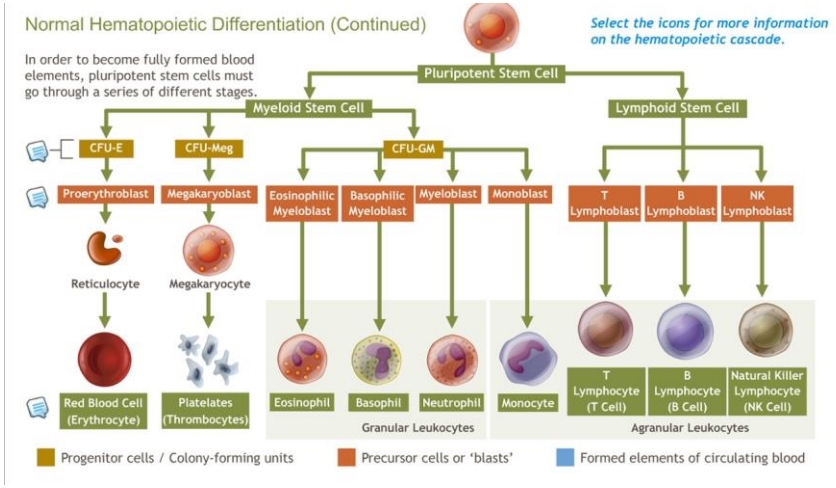
TG98
Glioblastoma Cell Line



MCF-7
Breast Adenocarcinoma

Next

- Extend the analysis to other tumours (e.g. hematological)



- Consider individual profiling for personalized targeting