Lecture 12: Additional Topics in Networks

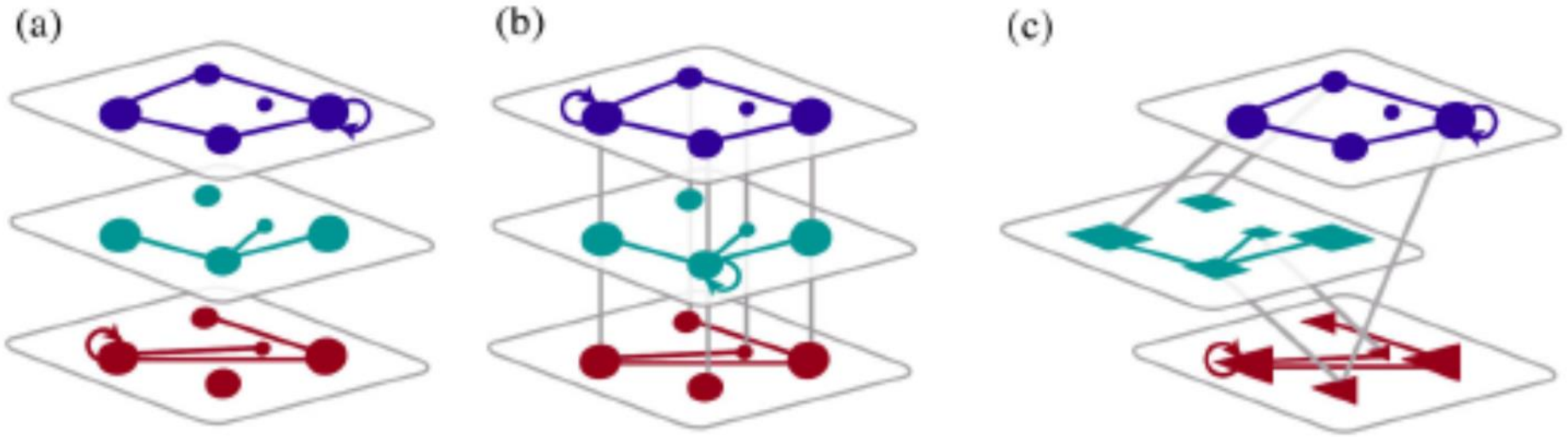
Multilayer Networks

Multilayer Networks

Two major classes of multilayer networks:

- ***Multiplex networks***
  - *Multiplex networks* are networks where the same set of nodes is represented in every layer, although the interaction between nodes might be different in each one.
- ***Network of networks***
  - This is instead formed by networks that are interlaced to each other but formed by different types of nodes

Multilayer Networks

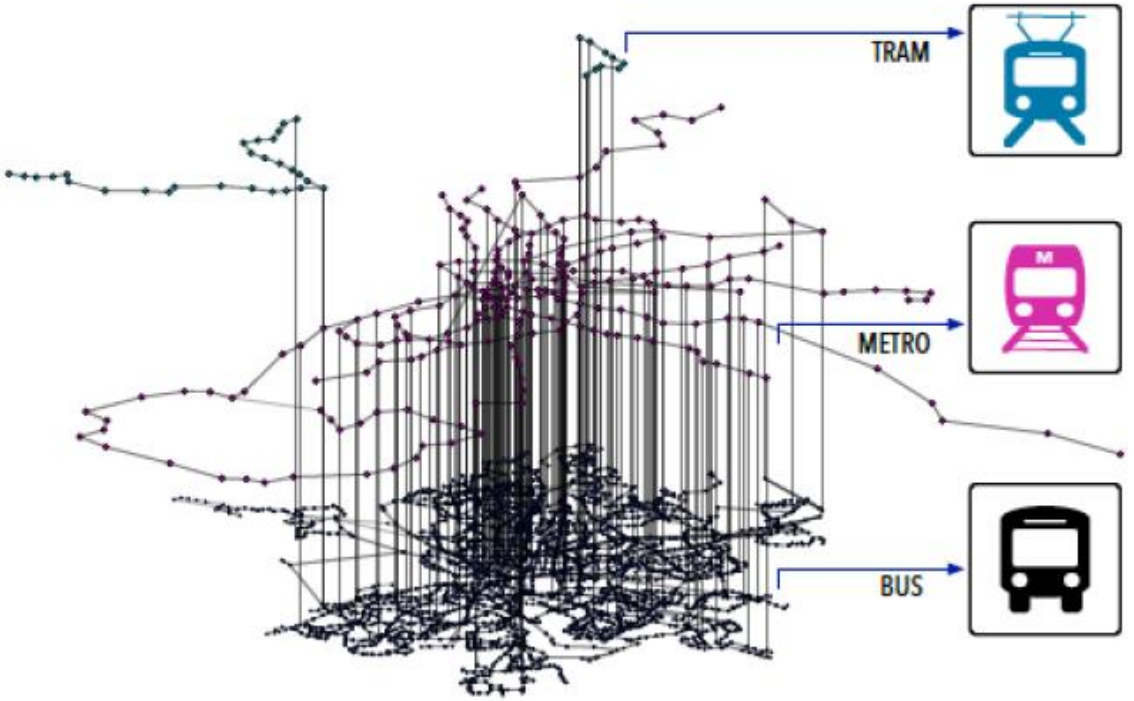


Multilayer Networks

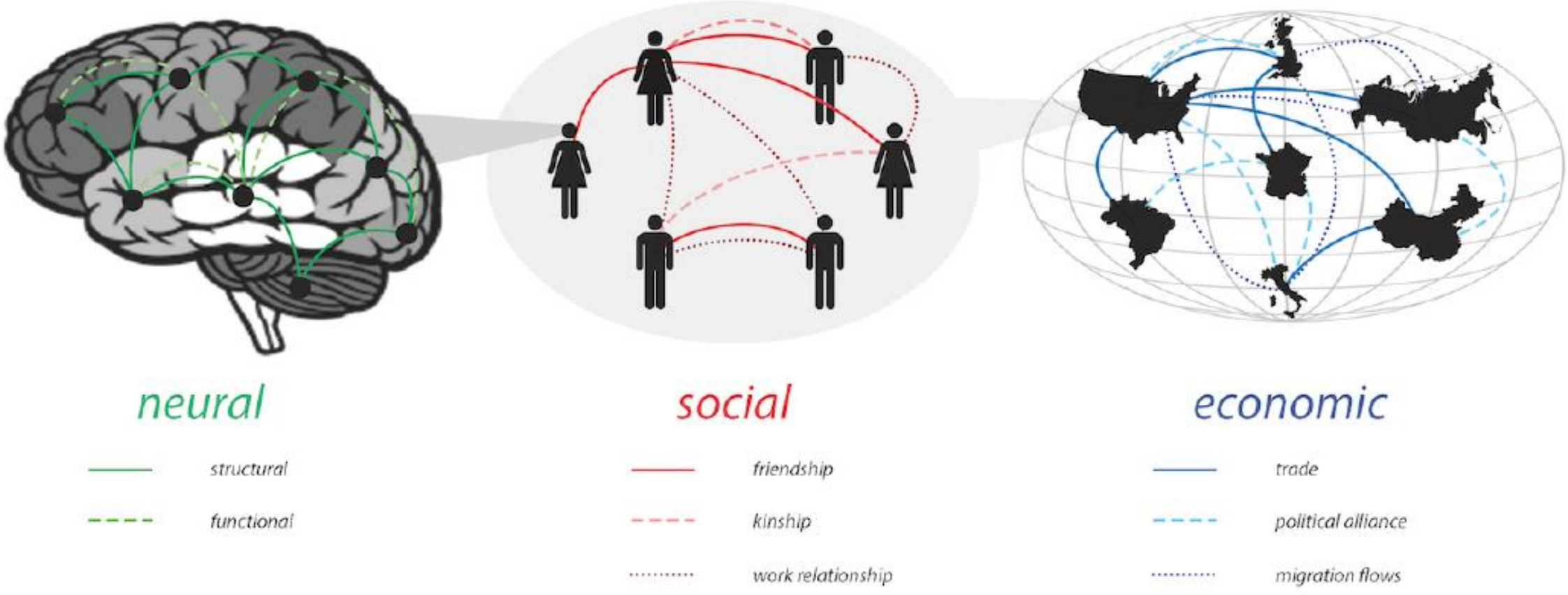
Relation between vertices can be of different kind. To represent networks with different types of links, one can consider structure divided into layers.

"Layerization"

Network is represented by layers of vertices. Every layer has a link related to the particular kind of relationship,



Multilayer Networks

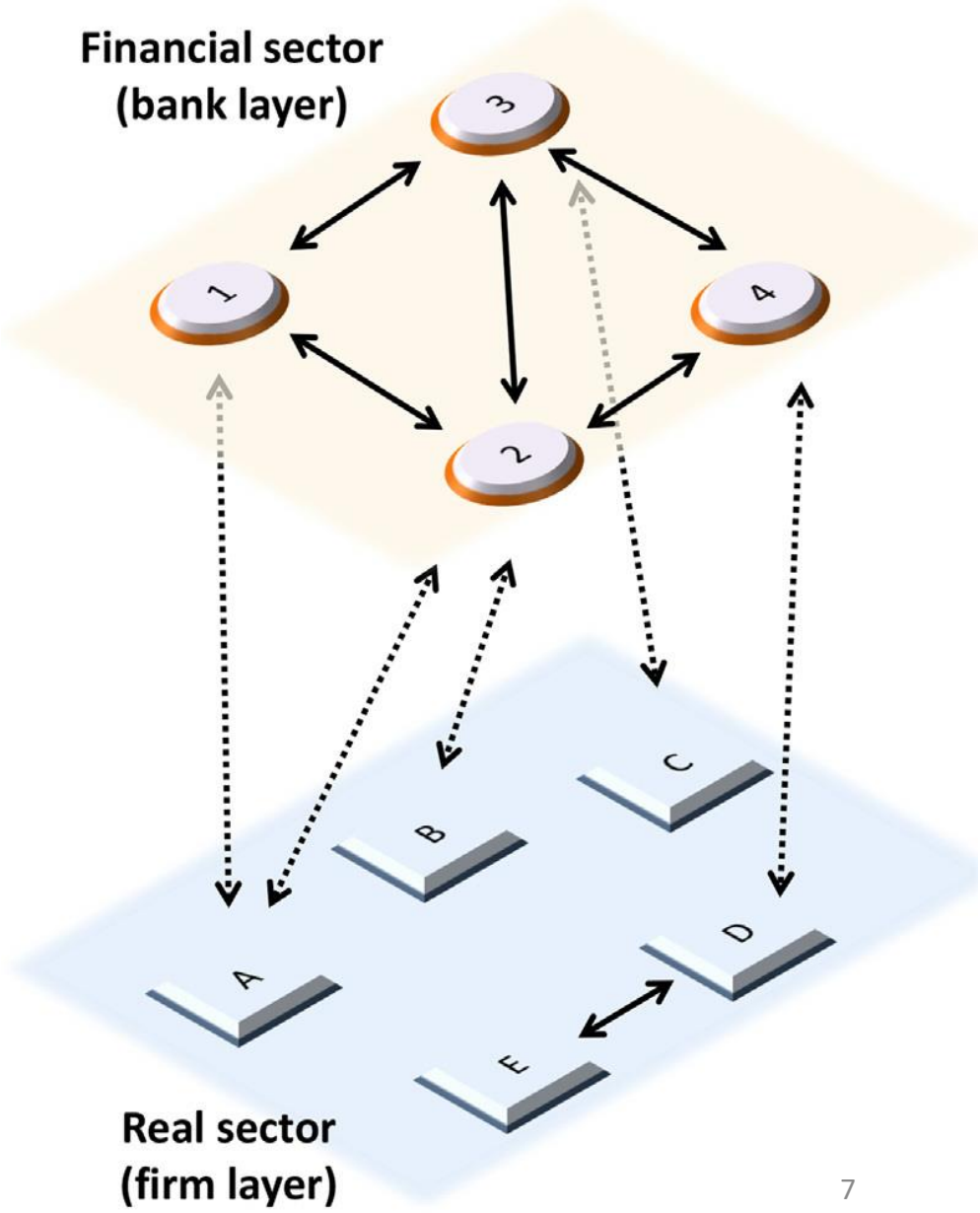


Multiplexity pervades many relational systems and may occur at very different spatio-temporal scales. At the scale of microns, our brain is composed of hundreds of billions of neurons coupled through diverse actions lasting milliseconds, which work together to make us move and take decisions (neural scale). The interactions among individuals, at distance of meters, ranging from seconds to weeks, regulated by a variety of social mechanisms such as social influence, homophily, etc (social scale). The largest one is that of countries, aggregates made of a large number of individuals, bond together by an intricate pattern of trade relations, political alliances and transportation links, thousands of kilometers away and which evolve over periods of months or years (economic scale).

Multilayer Networks

Example: Multilayer Networks in finance

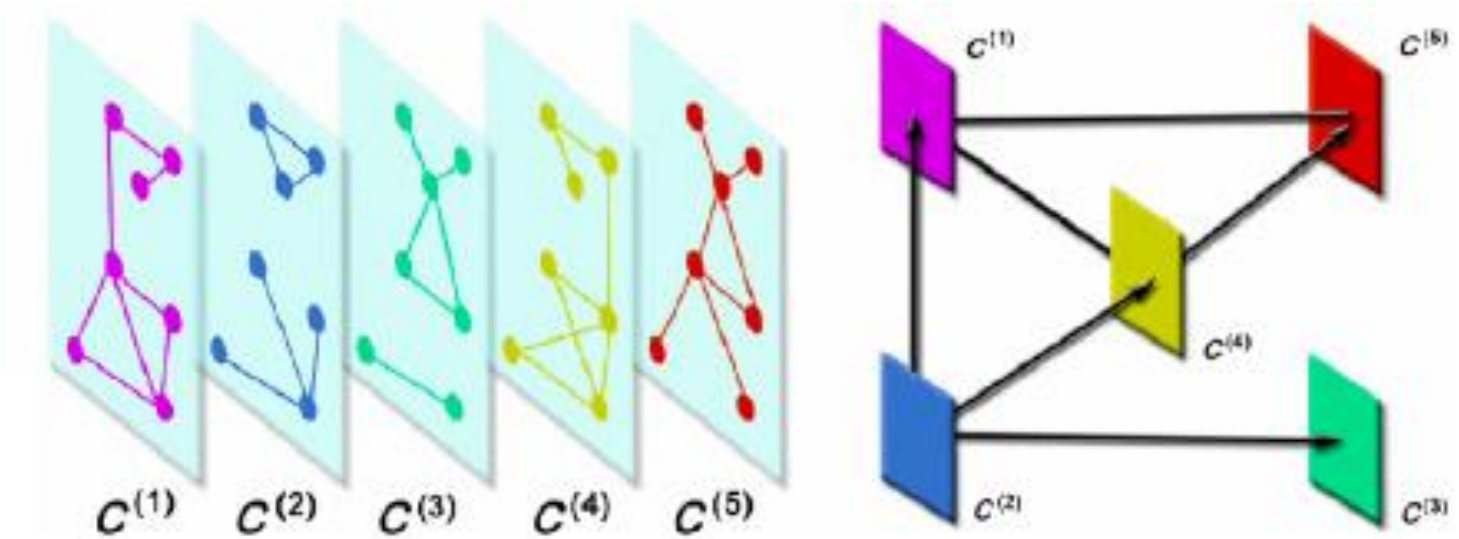
The multilayer formulation allows for novel approaches to the study of the dynamics of the economy. For example, researchers applied the multilayer approach to study the structure of financial markets by building a multilayer network where each layer represents different types of financial institutions, etc.



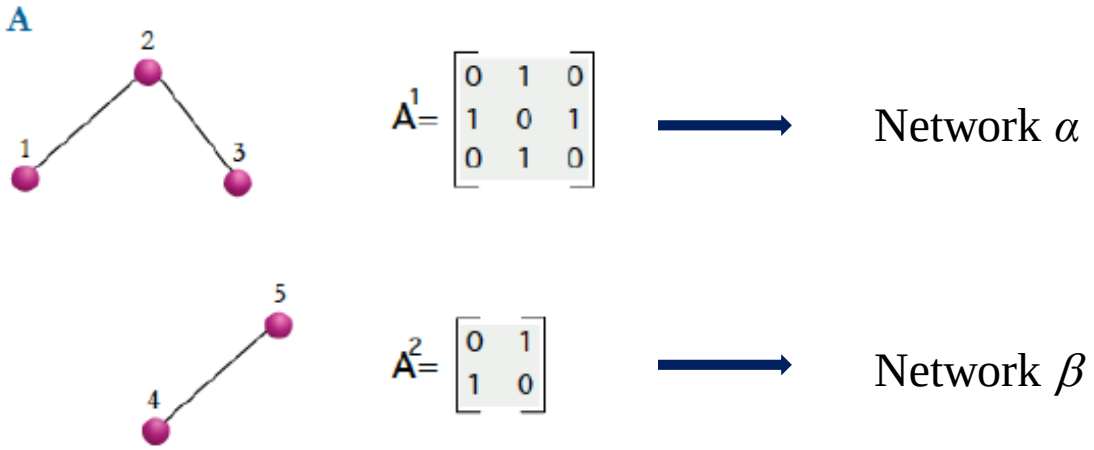
Multilayer Networks

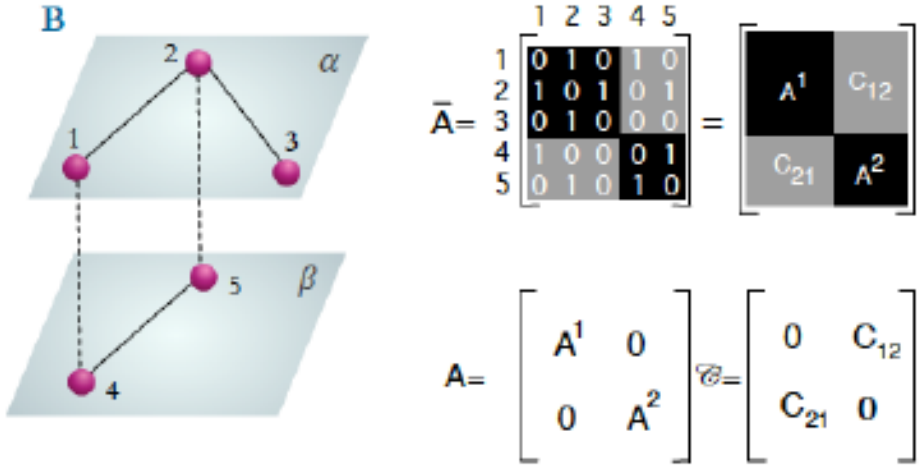
Intra-layer and Inter-layer

For this special case of networks we could in principle indicate the adjacency matrix for every layer (intra-layer) and the inter-layer connections.



Multilayer Networks





Adjacency Matrix

The structure of each layer is represented by an adjacency matrix A_i, where $i = 1(\alpha), = 2(\beta)$. C_{ab} stores the connections between layers α and β. Note that the number of nodes in each layer is not the same.

Multilayer Networks

Adjacency Matrix

The most general way to store the information on the multilayer networks is to use the generalization of the Adjacency matrix.

Adjacency Tensor

We now consider the quantity

$$A_{uv\alpha\beta} = \begin{cases} 1, & \text{if node } u \in \alpha \text{ is connected to node } v \in \beta \\ 0, & \text{otherwise} \end{cases}$$

This quantity is obviously redundant and one only uses this representation in special cases.

Multilayer Networks

Degree

We can define the degree of node *i* in layer α as

$$k_i^{\alpha} = \sum_j a_{ij}^{\alpha}$$

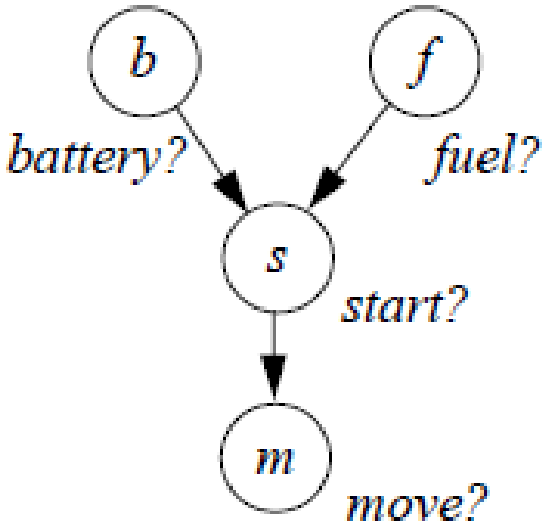
Consequently, its degree in the multilayer network is no longer a scalar but the vector

$$k_i = \{k_i^1, \ldots, k_i^L\}$$

which results in a total degree or degree overlap of $o_i = \sum_{\alpha} k_i^{\alpha}$.

Causal Networks

Causal relationships

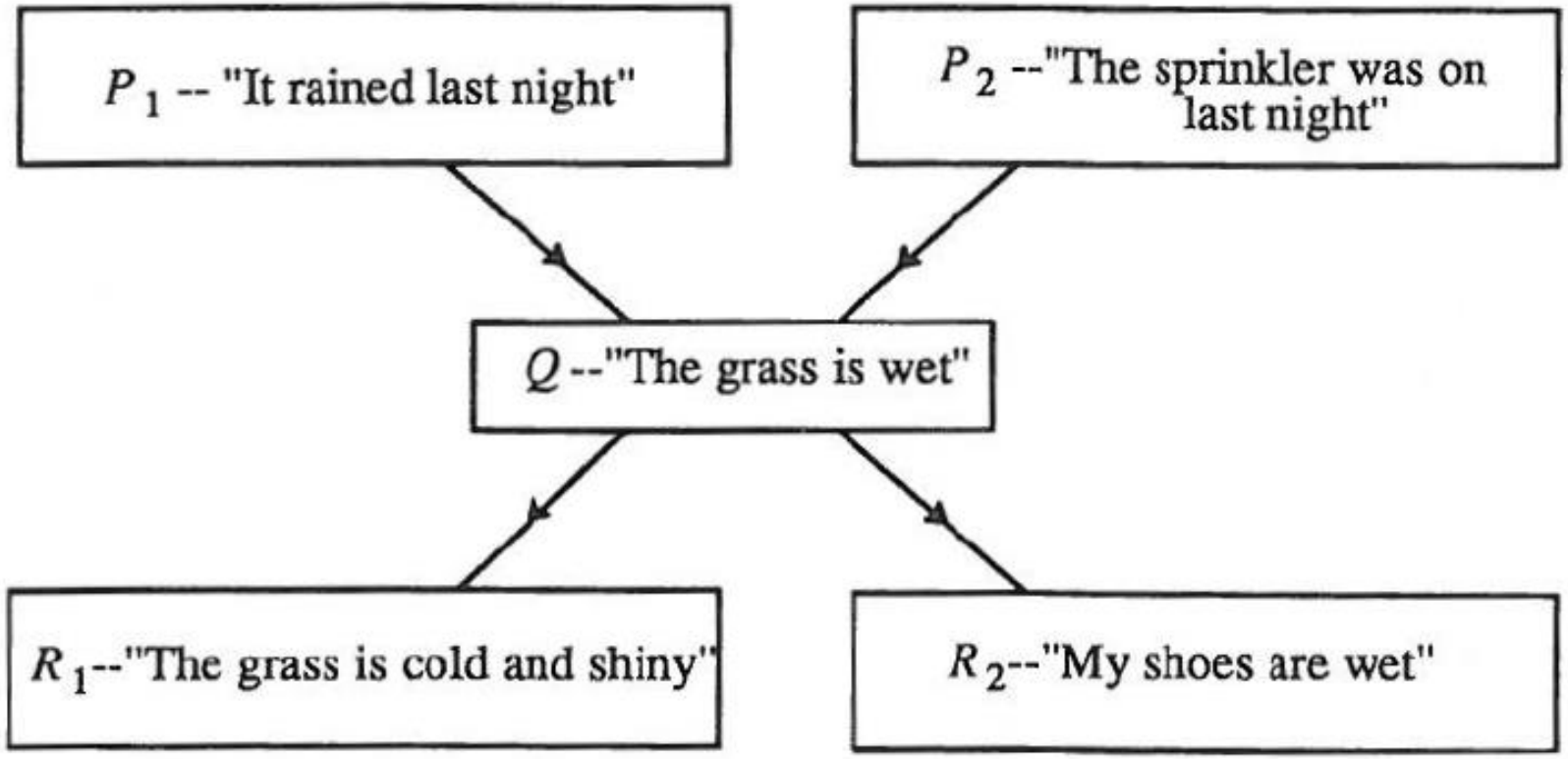


A causal network

Fire causes smoke

Smoke is evidence for fire

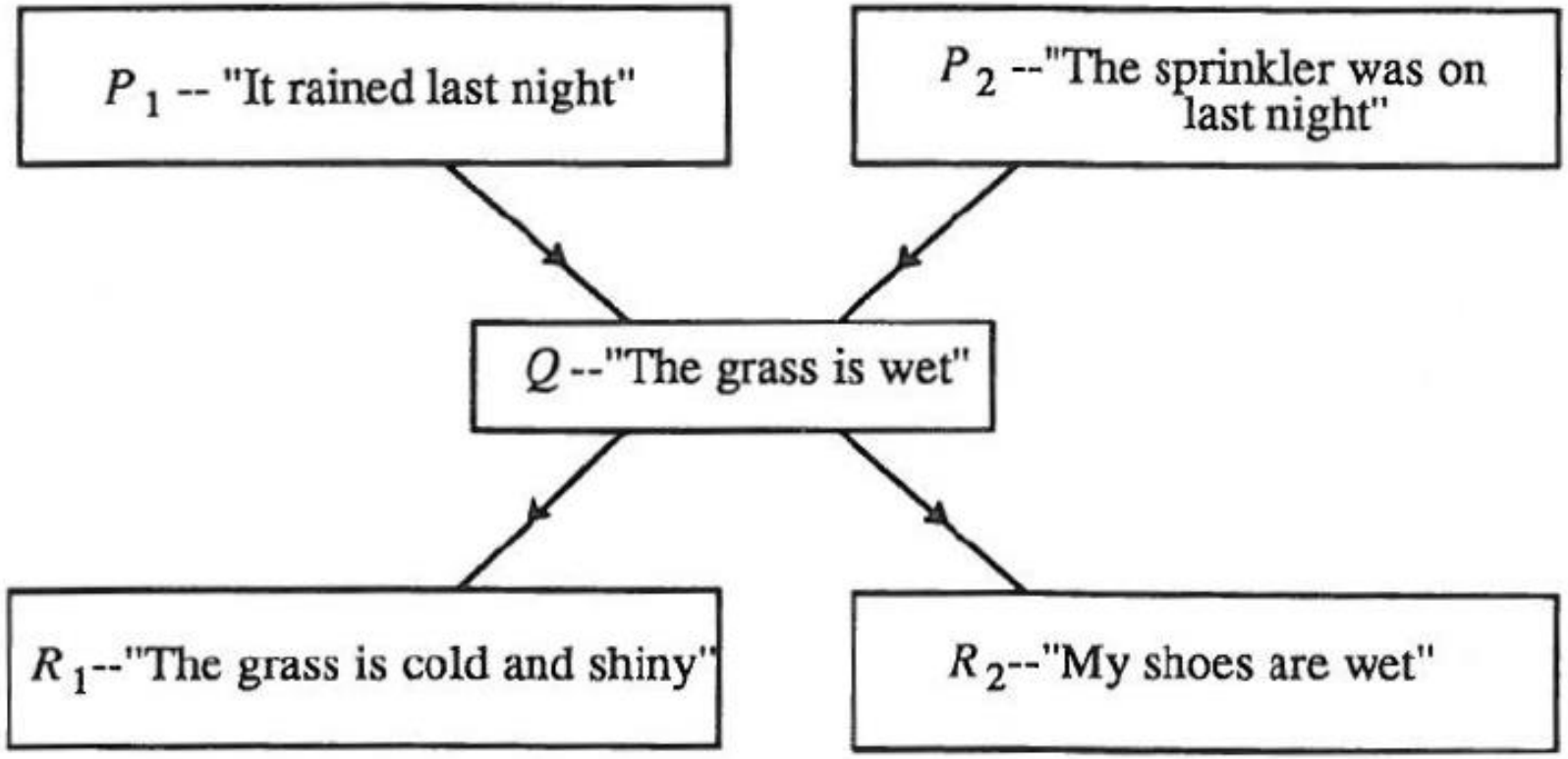
Causal relationships



P_1, P_2, Q, R_1 and R_2 stand for the propositions:

P_1 : "It rained last night,"
P_2 : "The sprinkler was on last night,"
Q : "The grass is wet,"
R_1 : "The grass is cold and shiny,"
R_2 : "My shoes are wet."

Causal relationships



The *causal* (*C*) and evidential (*E*) relationships between these propositions would be:

$$P_1 \rightarrow_C Q \;;\quad Q \rightarrow_E P_1$$
$$P_2 \rightarrow_C Q \;;\quad Q \rightarrow_E P_2$$
$$Q \rightarrow_C R_1 \;;\quad R_1 \rightarrow_E Q$$
$$Q \rightarrow_C R_2 \;;\quad R_2 \rightarrow_E Q$$

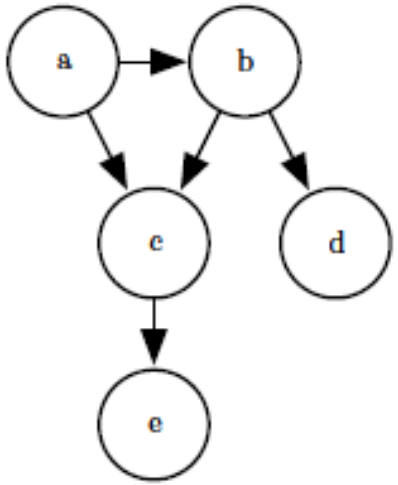
Directed graphical models are a way of encoding causal relationships between variables, whereas ***probabilistic graphical models*** are a way of encoding causality in a probabilistic manner.

A causal network is a ***Bayesian network*** with the requirement that the relationships be causal. A ***Bayesian network*** (also known as a Bayes network, Bayes net, belief network, or decision network) is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks are ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor.

Bayesian Networks

A Bayesian network is made up of two parts:

- *A directed acyclic graph*
  - *The nodes are random variables (which can be discrete or continuous)*
  - *Arrows connect pairs of nodes (X is a parent of Y if there is an arrow from node X to node Y).*
- *A set of parameters*



A directed graphical model over random variables *a*, *b*, *c*, *d* and *e*. This graph corresponds to probability distributions that can be factored as

$$p(a, b, c, d, e) = p(a)p(b \mid a)p(c \mid a, b)p(d \mid b)p(e \mid c).$$

This graphical model enables us to quickly see some properties of the distribution. For example, *a* and *c* interact directly, but *a* and *e* interact only indirectly via *c*.

Bayesian Networks

- Provide a way to represent knowledge in an uncertain domain and a way to reason about this knowledge
- Many applications: medicine, factories, help desks, spam filtering, etc.
- Bayesian networks have been the most important contribution to the field of ***AI*** in the last 10 years

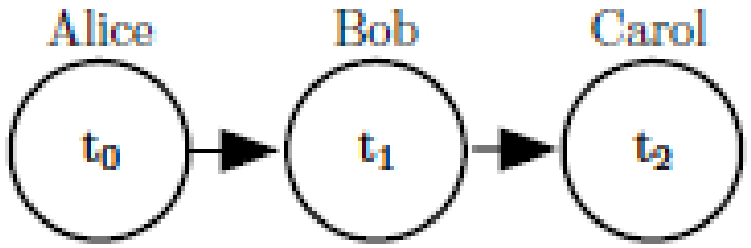
Inference Using Bayes Theorem

- The general probabilistic inference problem is to find the probability of an event given a set of evidence;
- This can be done in Bayesian nets with sequential applications of Bayes Theorem;
- In 1986 Judea Pearl published an innovative algorithm for performing inference in Bayesian nets.

Inference Using Bayes Theorem

Example:

Consider modeling the finishing times of a team in a relay race. Suppose the team consists of three runners: Alice, Bob and Carol. At the start of the race, Alice carries a baton and begins running around a track. After completing her lap around the track, she hands the baton to Bob. Bob then runs his own lap and hands the baton to Carol, who runs the final lap.



A directed graphical model depicting the relay race example. Alice's finishing time t_0 influences Bob's finishing time t_0, because Bob does not get to start running until Alice finishes. Likewise, Carol only gets to start running after Bob finishes, so Bob's finishing time t_1 directly influences Carol's finishing time t_2.

An Application: Reconstruction of Gene Regulatory Networks

Organism	Number of predicted genes	Part of the genome that encodes proteins (exons)
E. Coli (bacteria)	5000	90%
Yeast	6000	70%
Worm	18,000	27%
Fly	14,000	20%
Weed	25,500	20%
Human	30,000	< 5%

Chimpanzee at work

Human at work

Only **1%** difference between human & chimp

Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios

Andrew G. Clark,[1] Stephen Glanowski,[3] Rasmus Nielsen,[2] Paul D. Thomas,[4] Anish Kejariwal,[4] Melissa A. Todd,[2] David M. Tanenbaum,[5] Daniel Civello,[6] Fu Lu,[5] Brian Murphy,[3] Steve Ferriera,[3] Gary Wang,[3] Xianqgun Zheng,[5] Thomas J. White,[6] John J. Sninsky,[6] Mark D. Adams,[5]* Michele Cargill[6]†

Even though human and chimpanzee gene sequences are nearly 99% identical, sequence comparisons can nevertheless be highly informative in identifying biologically important changes that have occurred since our ancestral lineages diverged. We analyzed alignments of 7645 chimpanzee gene sequences to their human and mouse orthologs. These three-species sequence alignments allowed us to identify genes undergoing natural selection along the human and chimp lineage by fitting models that include parameters specifying rates of synonymous and nonsynonymous nucleotide substitution. This evolutionary approach revealed an informative set of genes with significantly different patterns of substitution on the human lineage compared with the chimpanzee and mouse lineages. Partitions of genes into inferred biological classes identified accelerated evolution in several functional classes, including olfaction and nuclear transport. In addition to suggesting adaptive physiological differences between chimps and humans, human-accelerated genes are significantly more likely to underlie major known Mendelian disorders.

Although the human genome project will allow us to compare our genome to that of other primates and discover features that are uniquely human, there is no guarantee that such features are responsible for any of our unique biological attributes. To identify genes and biological processes that have been most altered by our recent evolutionary divergence from other primates, we need to fit the data to models of sequence divergence that allow us to distinguish between divergence caused by random drift and divergence driven by natural selection. Early observations of unexpectedly low levels of protein divergence between humans and chimpanzees led to the hypothesis that most of the evolutionary changes must have occurred at the level of gene regulation (*1*). Recently, much more extensive efforts at DNA sequencing in nonhuman primates has confirmed the very close evolutionary relationship between humans and chimps (*2*), with an average nucleotide divergence of just 1.2% (*3–5*). The role of protein divergence in causing morphological, physiological, and behavioral differences between these two species, however, remains unknown.

Here we apply evolutionary tests to identify genes and pathways from a new collection of more than 200,000 chimpanzee exonic sequences that show patterns of divergence consistent with natural selection along the human and chimpanzee lineages.

To construct the human-chimp-mouse alignments, we sequenced PCR amplifications using primers designed to essentially all human exons from one male chimpanzee, resulting in more than 20,000 human-chimp gene alignments spanning 18.5 Mb (*6–8*). To identify changes that are specific to the divergence in the human lineage, we compared the human-chimp aligned genes to their mouse ortholog. Inference of orthology involved a combination of reciprocal best matches and syntenic evidence between human and mouse gene annotations (*9*, *10*). This genome-wide set of orthologs underwent a series of filtering steps to remove ambiguities, orthologs with little sequence data, and genes with suspect annotation (*6*). The filtered ortholog set was compared to

[1]Molecular Biology and Genetics, [2]Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA. [3]Applied Biosystems, 45 West Gude Drive, Rockville, MD 20850, USA. [4]Protein Informatics, Celera Genomics, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [5]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [6]Celera Diagnostics, 1401 Harbor Bay Parkway, Alameda, CA 94502, USA.

*Present address: Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA.

†To whom correspondence should be addressed. E-mail: michele_cargill@celeradiagnostics.com

Question:

How are genes regulated?

Basic molecular biology

DNA, the main information carrier molecule in a cell, is constructed by 2 strands of *polynucleotide* which is a chain of small molecules called *nucleotide*.

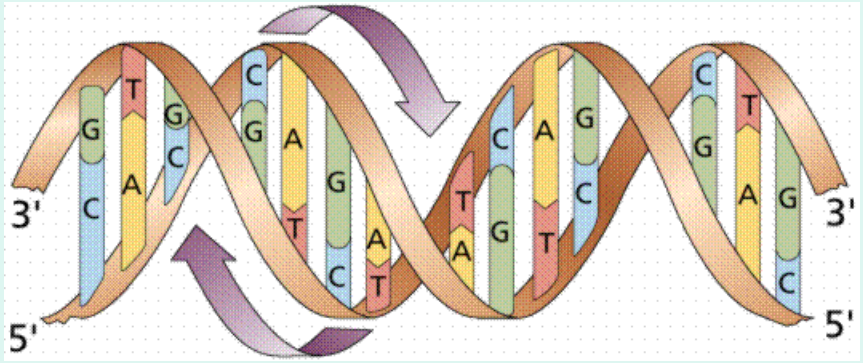
There are 4 basic *nucleotide* :

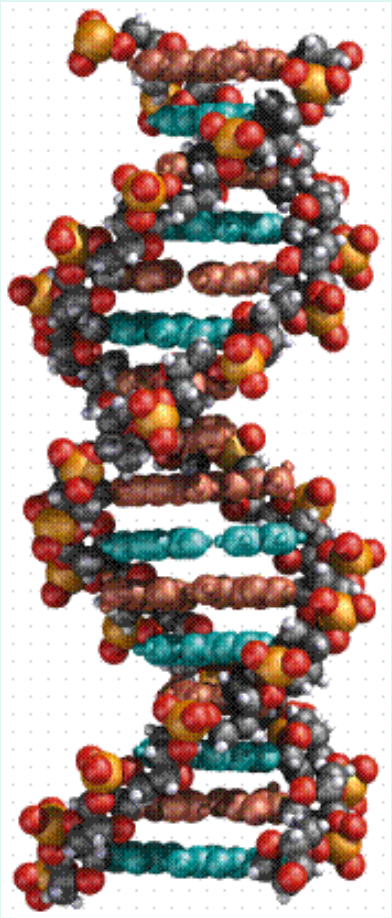
A (*adenosine*)

T (*thymine*)

C (*cytosine*)

G (*guanine*)





- ***RNA***, like ***DNA*** is constructed from ***nucleotides***. But instead of the *thymine* (**T**), it has an alternative *uracil* (**U**) .

There are 3 types of *RNA*:

m*RNA* (messenger *RNA*)
r*RNA* (ribosomal *RNA*)
t*RNA* (transfer *RNA*)

- ***Gene***, a specific segment of the ***DNA*** that codes ***proteins***.

- ***Protein***, the main building blocks and functional molecules of the organism.

•Genome, the set of genes carried by an individual.

•gene –transcription/splicing--> mRNA –translation--> protein

•gene expression

•gene expression level

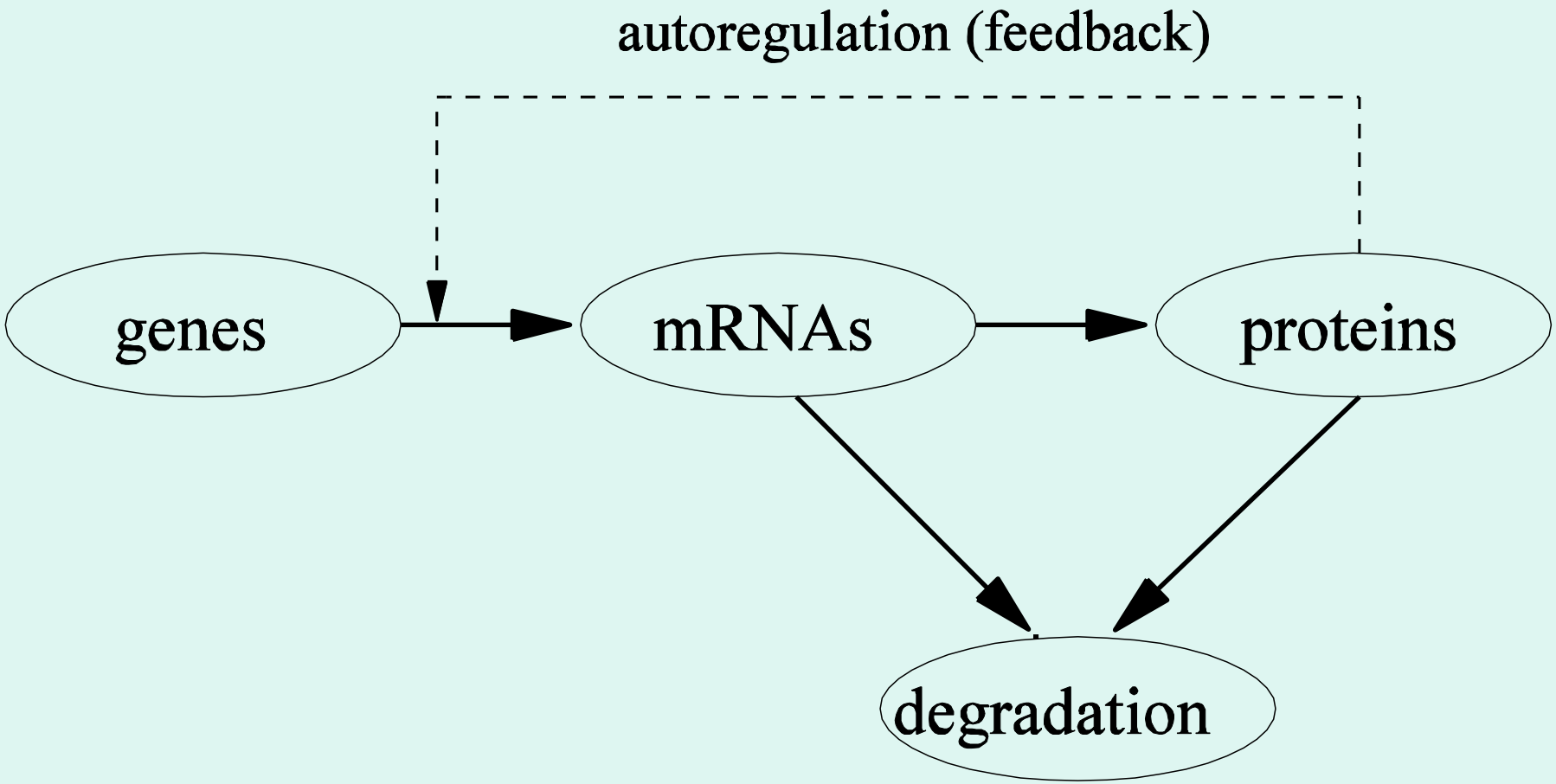
•Chromosomes, long double stranded DNA molecules

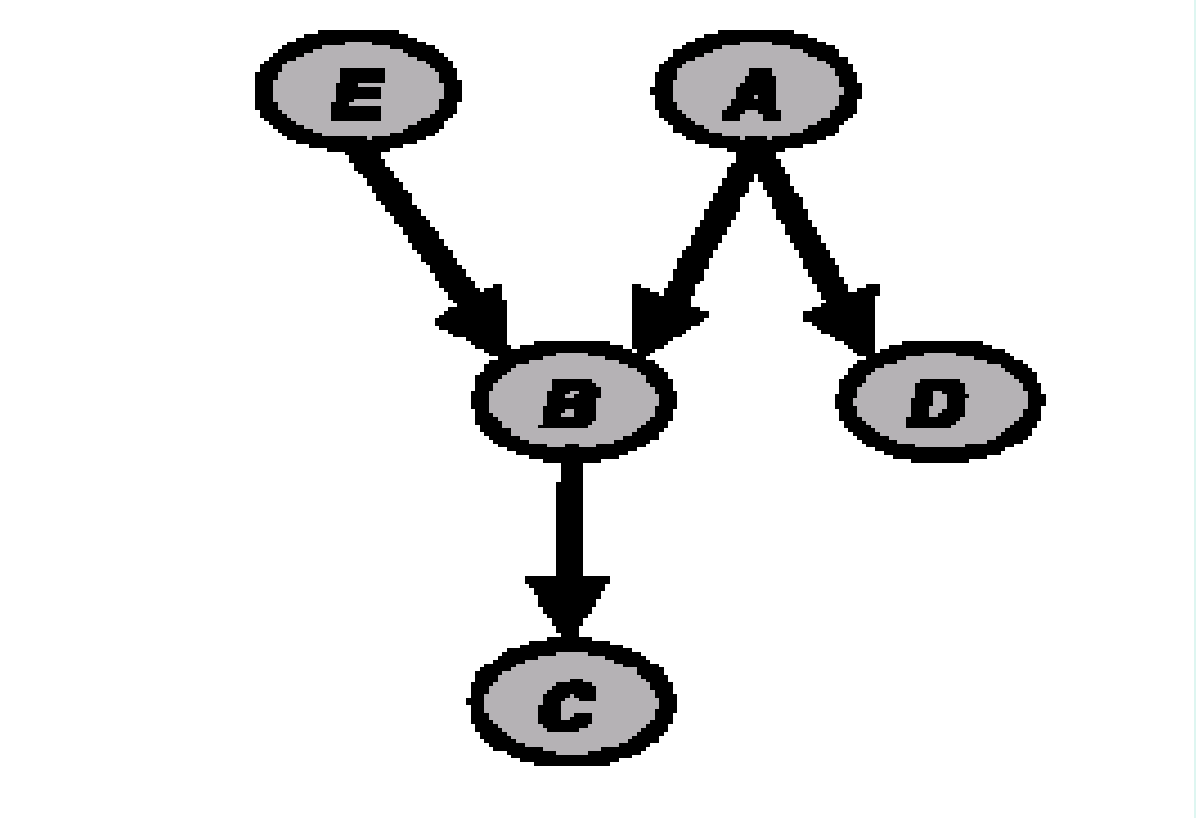
The total genome size differs quite considerably in different organisms, and the following table shows a few examples.

Organism	Number of chromosomes	Genome size in base pairs
Bacteria	1	~400,000 - ~10,000,000
Yeast	12	14,000,000
Worm	6	100,000,000
Fly	4	300,000,000
Weed	5	125,000,000
Human	23	3,000,000,000

Gene Regulation and Networks

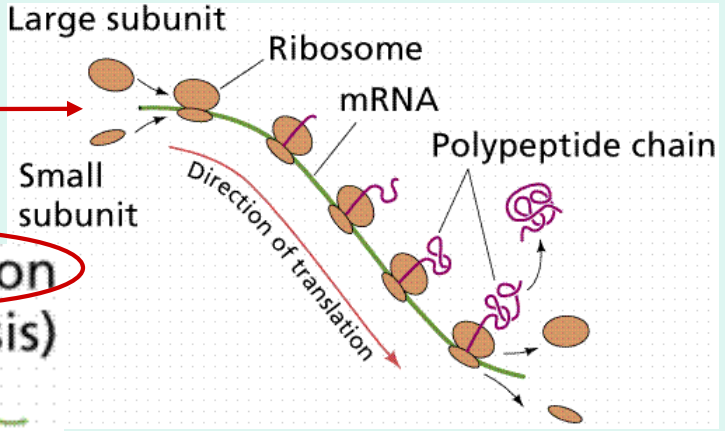
- ~320 different types of human cells
- In different cells at any time, different genes are expressed to different levels
- when/how genes are regulated by other genes
- gene regulatory network (circuit) <- key to an understanding of life!
- Gene profiling: monitoring the state of an organism by DNA micro-array

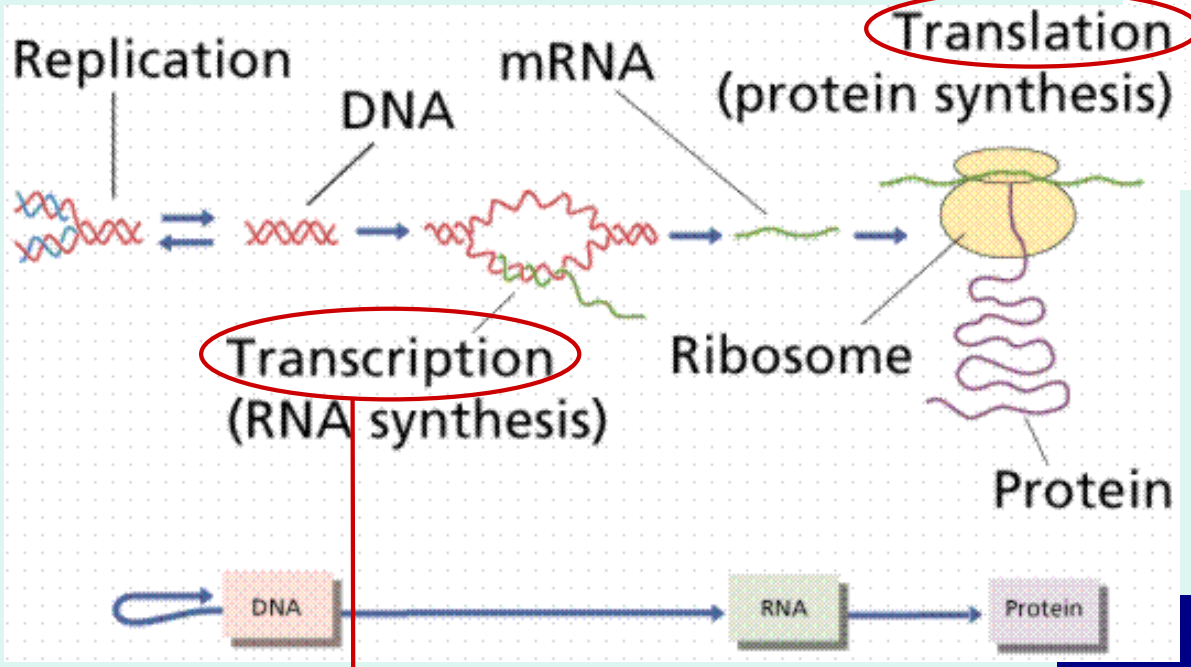

autoregulation (feedback)
genes
mRNAs
proteins
degradation

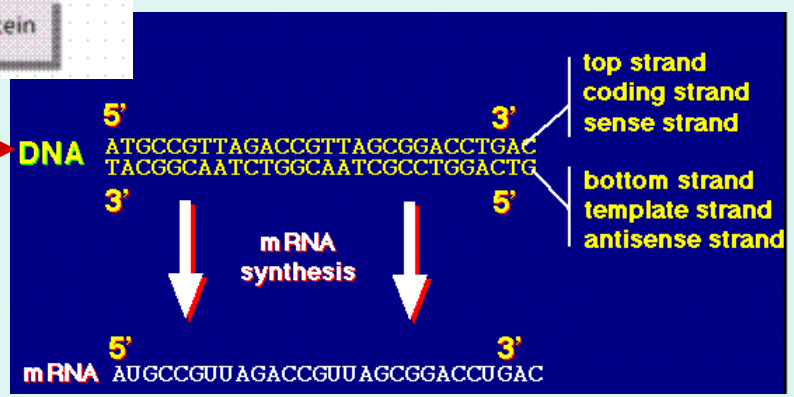


A gene regulatory network of 5 genes

Protein Synthesis





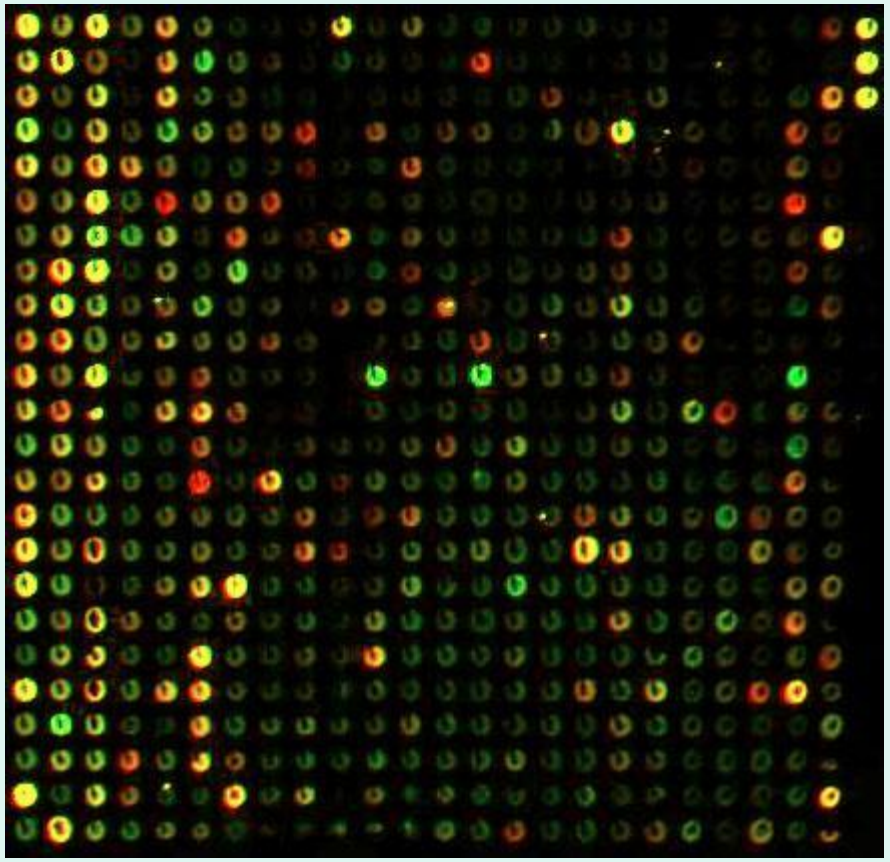


Traditional analysis:

- one gene in one experiment
- small data sets
- simple analytical methods

High-throughput genomic analysis:

- simultaneous measurements of thousands of gene expression levels → massive data sets
- statistical methods
- machine learning approach



※ A typical microarray is about 1 sq. in., the spot diameter is of the order of 0.1 mm, but for some types can be even smaller.

The treated samples are labeled with red fluorescein and wild with green. The red/green intensity reflects the m*RNA* abundance in the mixture.

Issues

Number of replication is low

$$\frac{dx_i}{dt} = \sum_j w_{ij} x_j + b_i$$

Microarray data are noisy:

-> instrumental errors

-> stochastic nature of biochemical processes

Model & Simulation

$$\frac{dx_i}{dt} = \alpha_i \prod_j x_j^{w_{ij}} - \beta_i x_i$$

An example:

$$x_1' - x_1 = -0.2x_1$$

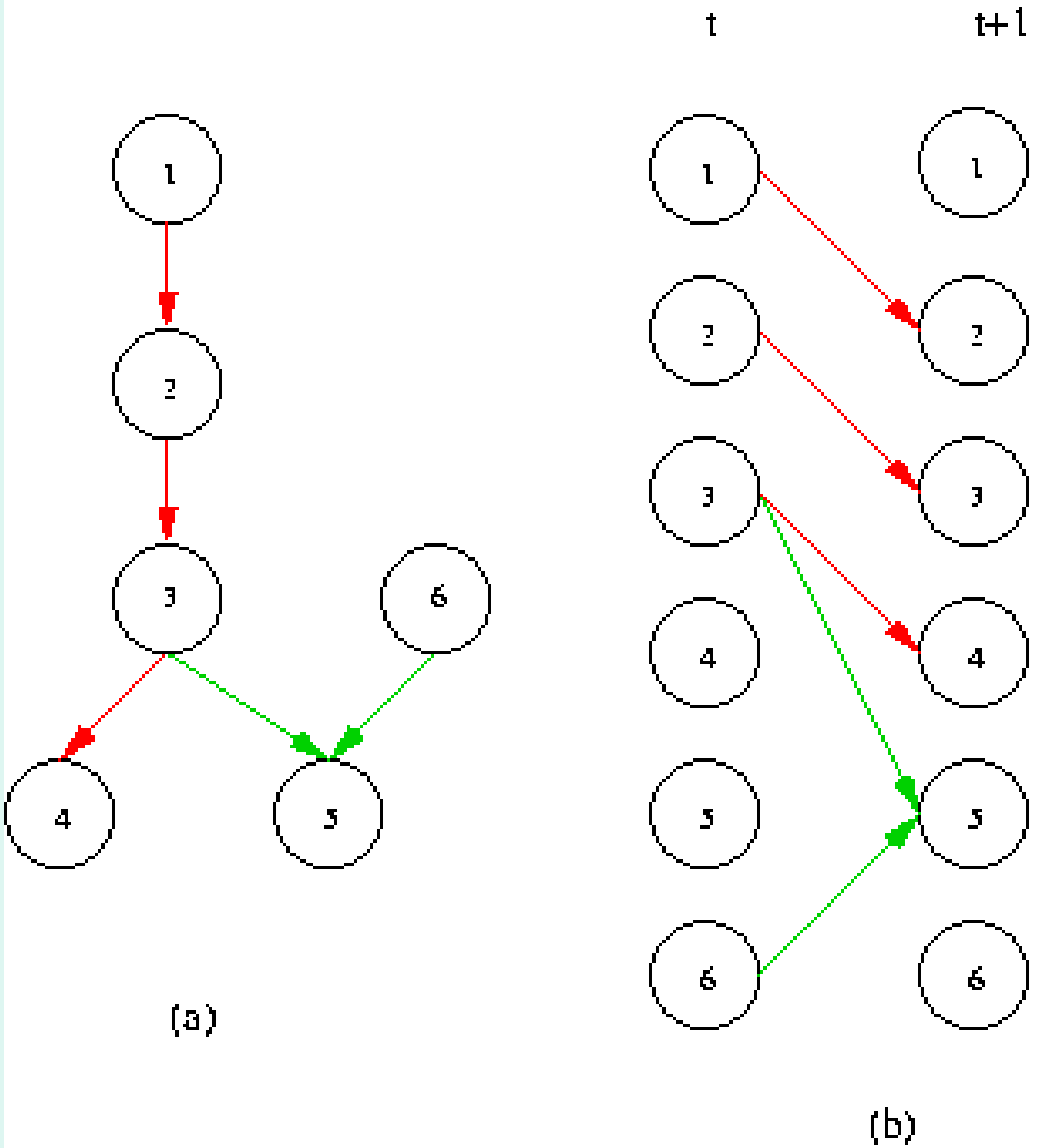
$$x_2' - x_2 = 1.2x_1^{2.0} - 0.6x_2$$

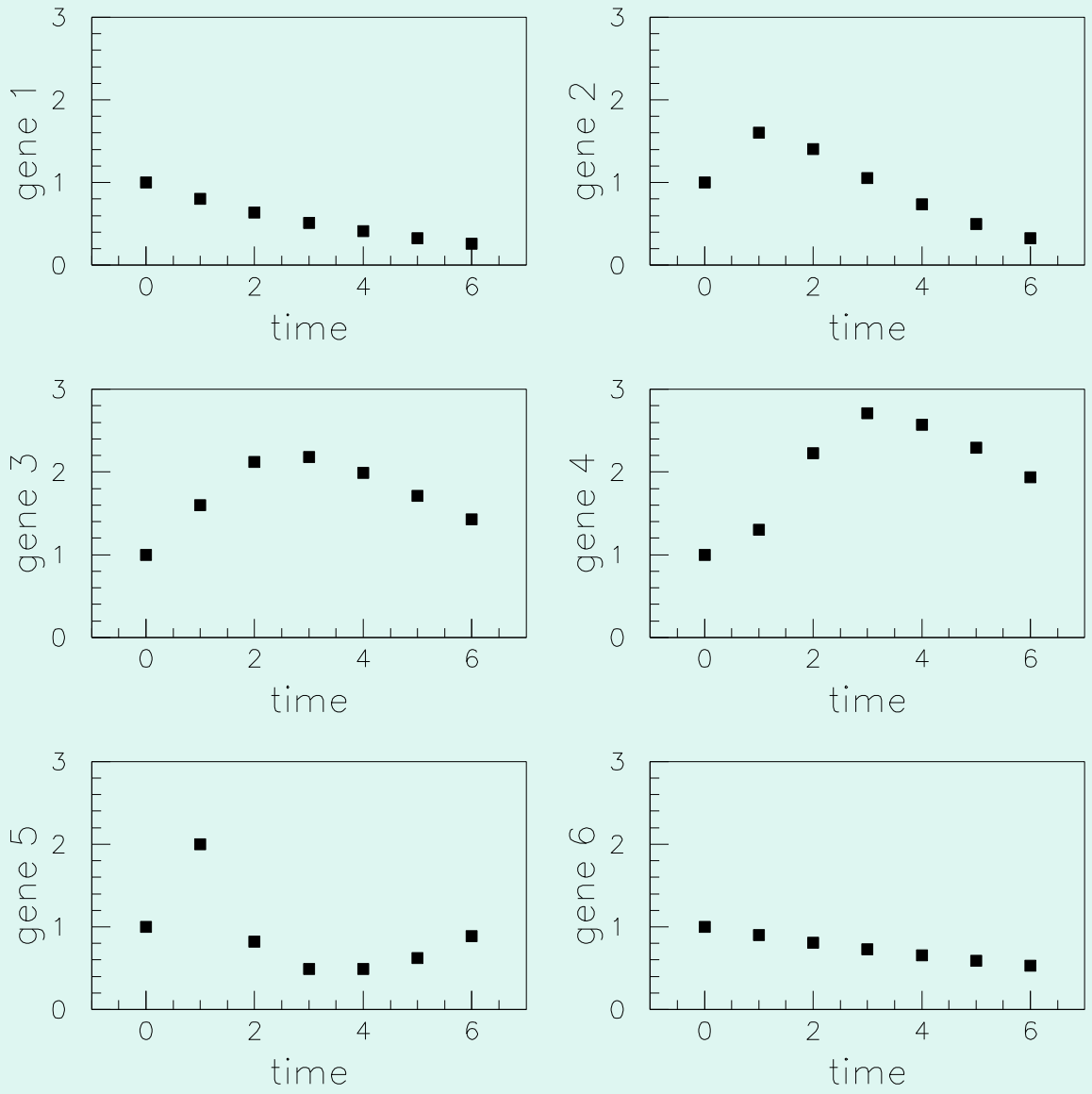
$$x_3' - x_3 = 1.3x_2^{0.5} - 0.7x_3$$

$$x_4' - x_4 = 1.8x_3^{1.0} - 1.5x_4$$

$$x_5' - x_5 = 2.0x_3^{-2.0}x_6^{-0.5} - 1.0x_5$$

$$x_6' - x_6 = -0.1x_6$$


t
t+1
1
2
3
4
5
6
1
2
3
4
5
6
(a)
(b)


gene 1
gene 2
gene 3
gene 4
gene 5
gene 6
time
0
1
2
3
0
2
4
6

Optimization Algorithms

- Monte Carlo Simulation
- Simulated Annealing
- Genetic Algorithm
- Guided Simulated Annealing
- Others -- Machine learning, etc

Computational algorithms

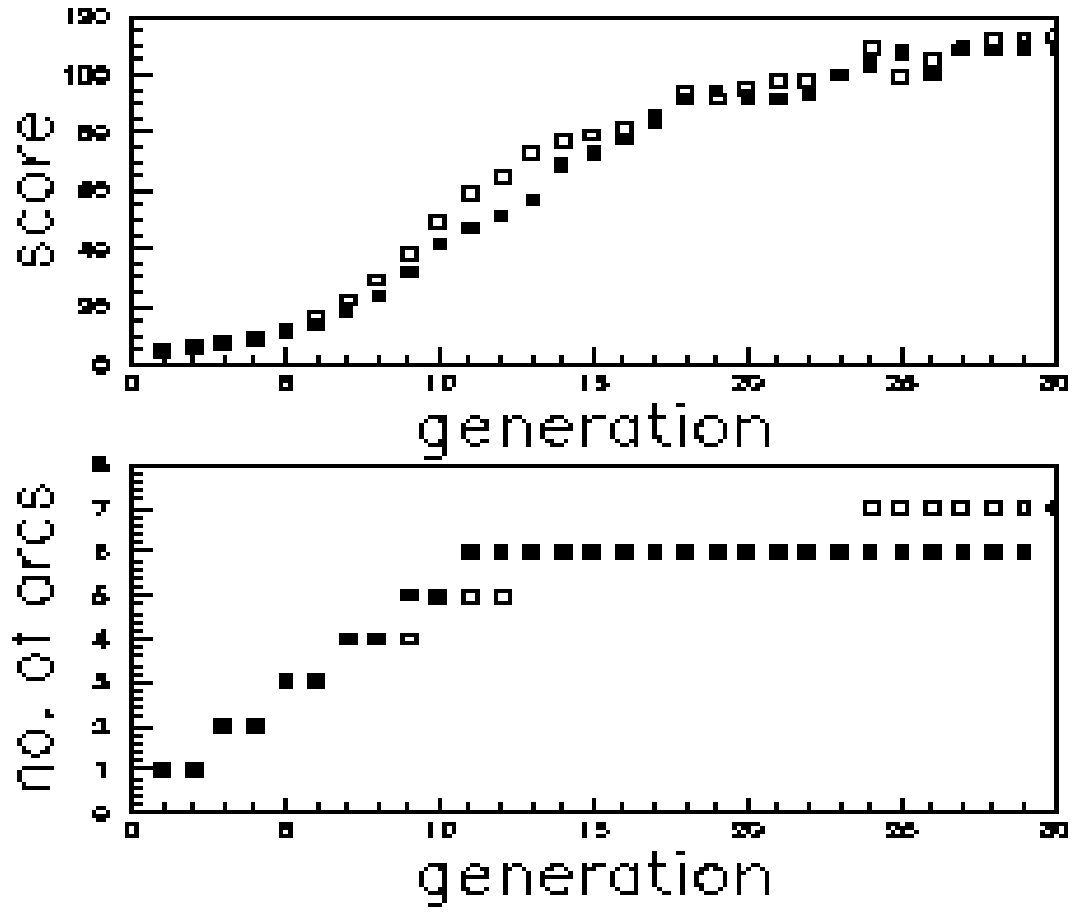
- Genetic algorithm for network structures:

  less prone to being trapped in local minima

  (encoding, recombination, mutation, selection, population size)

- Downhill simplex for parameter estimation:

  efficient for high-dimensional function maximization


score
generation
no. of arcs
generation

Analysis of Real Data

J.L. DeRisi, V.R. Iyer, P.O. Brown, "*Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale*", Science, vol. **278**, (1997) 680-686

DNA microarrays containing virtually every gene of Saccharomyces cerevisiae were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. …

Analysis of Real Expression Data

1. yeast glycolytic pathway

2. yeast cell cycle

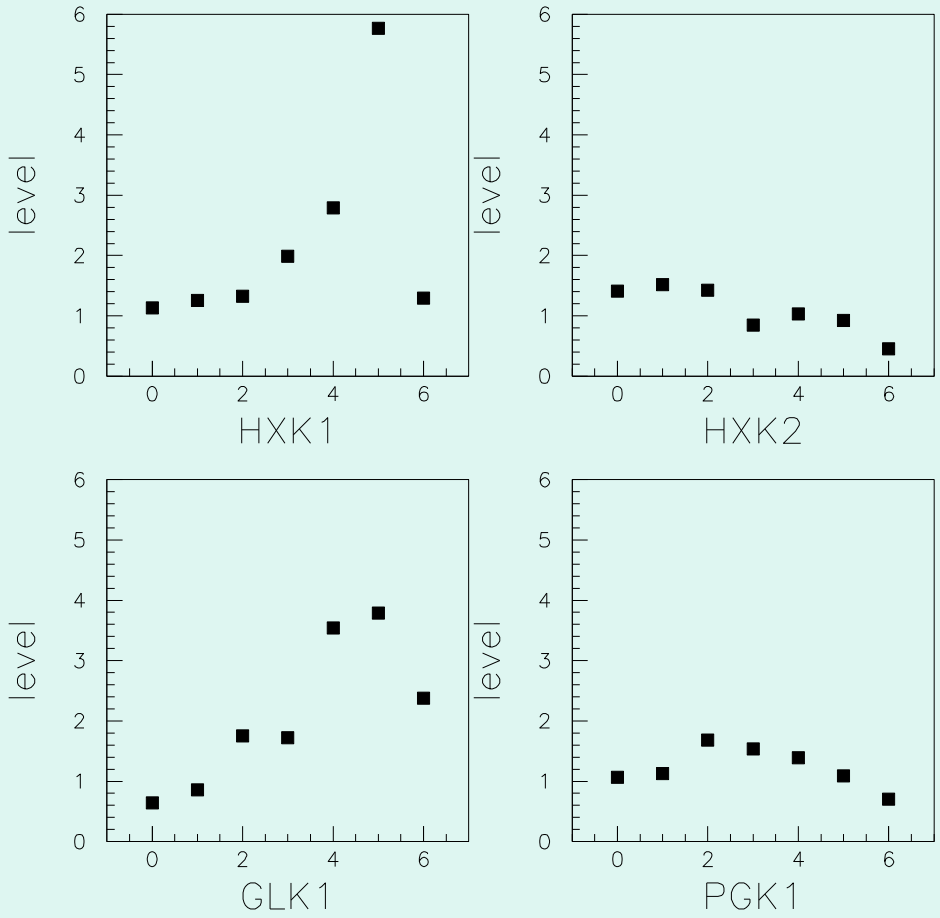
3. yeast stress response (heat shock)

available at http://genome-www.stanford.edu

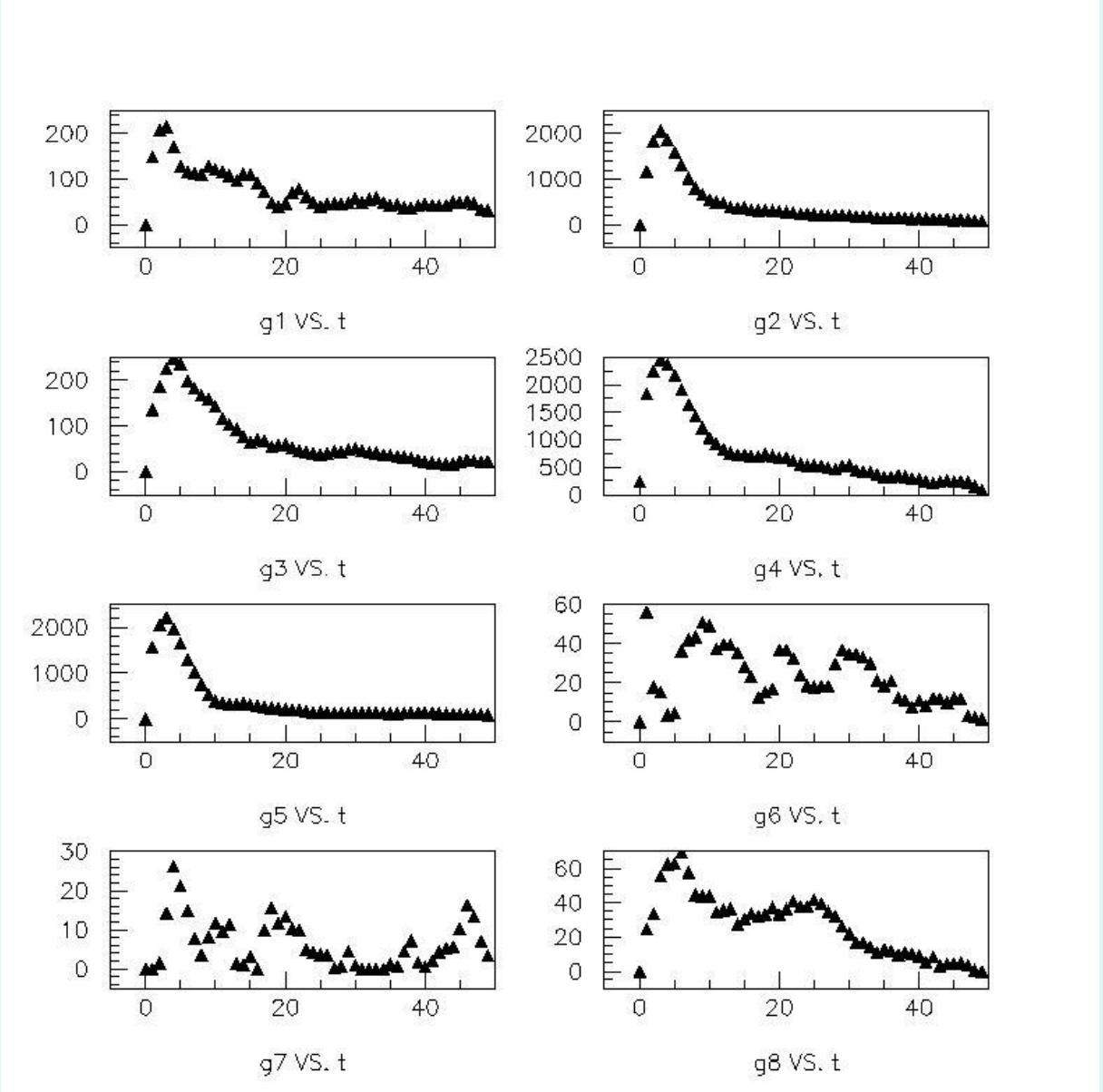
4. *E. coli* S.O.S. DNA damage repair

available at http://www.weizmann.ac.il/mcb/UriAlon

Real microarray data (yeast diauxic shift)



E. Coli S.O.S. system promoter activity data



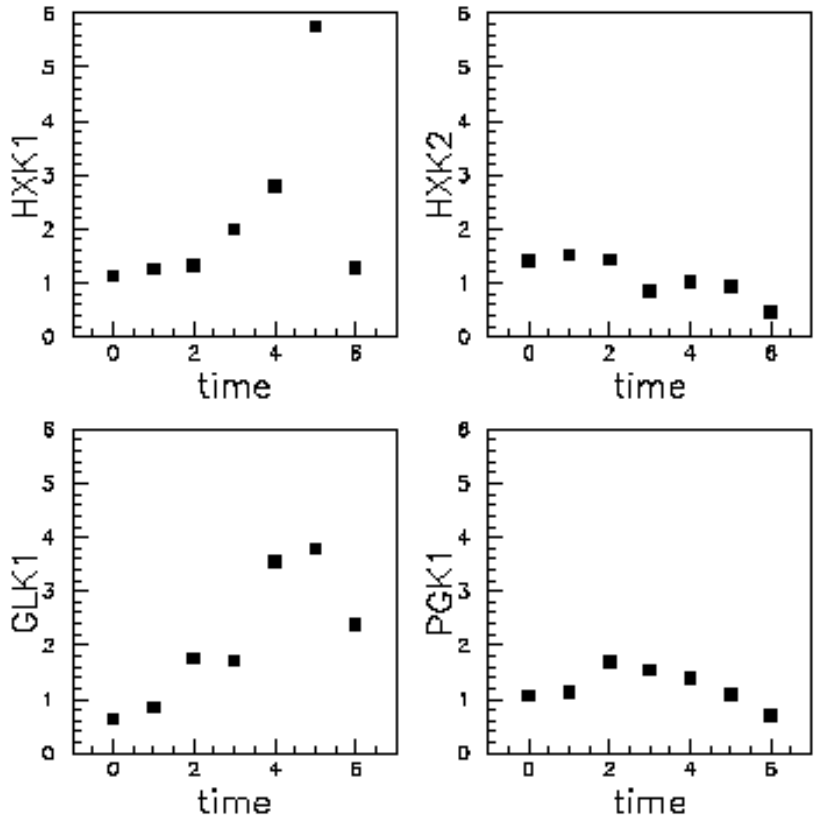


Fig. 7. Expression levels of genes *HXK1, HXK2, GLK1*, and *PGK1* across time. Data are from DeRisi *et al.* (1997).

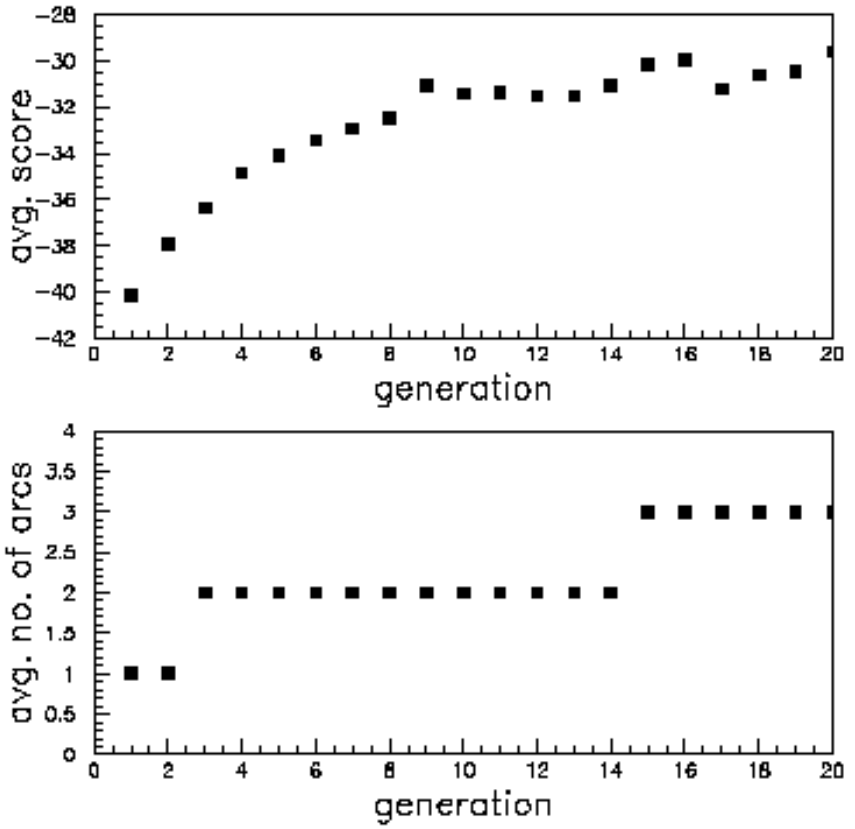
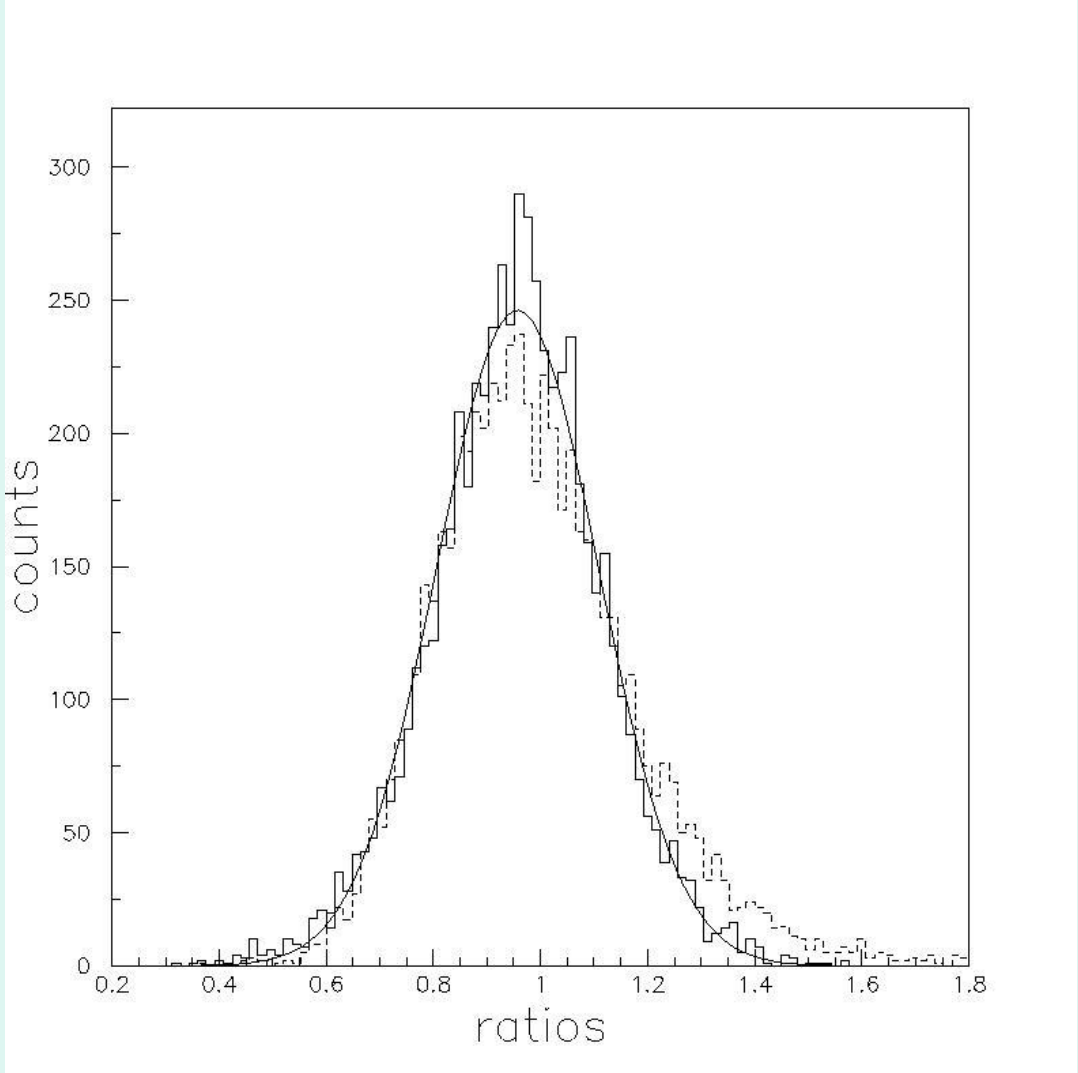


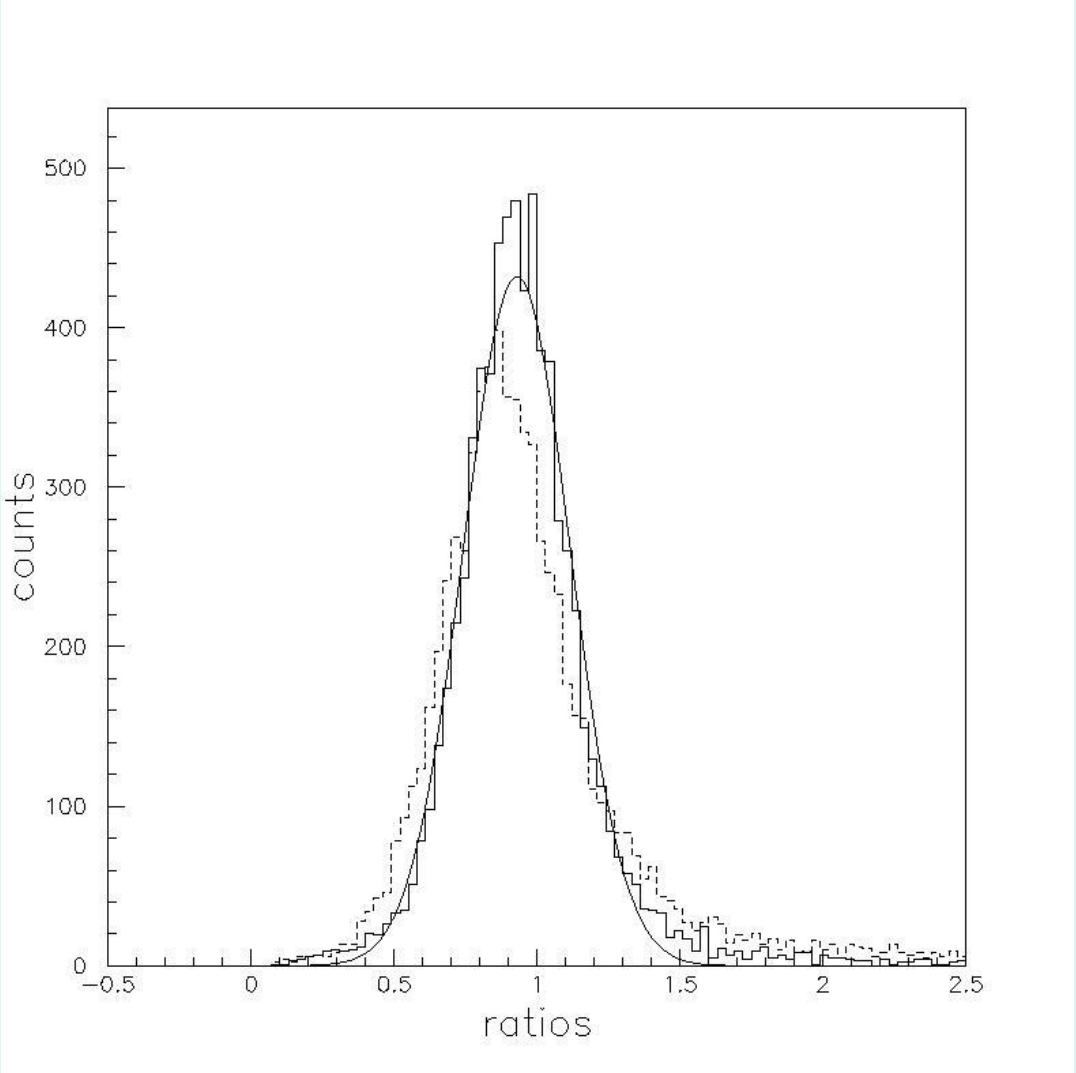
Fig. 8. Top (bottom): increase in the average score (number of arcs) over the generation number.

Genomic expression ratios of yeast glycolysis
σ=0.15



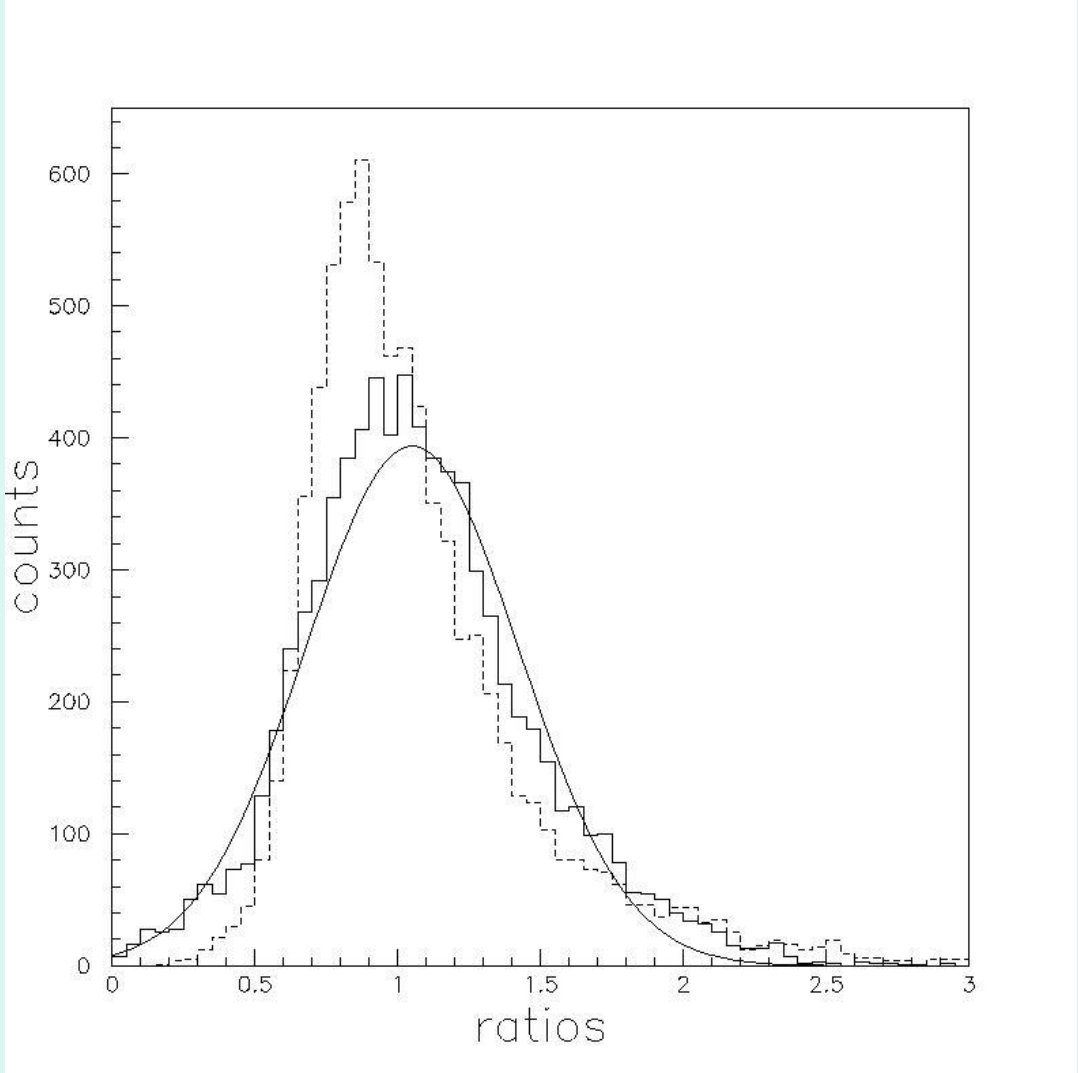
Genomic expression ratios of yeast cell cycle

σ=0.19

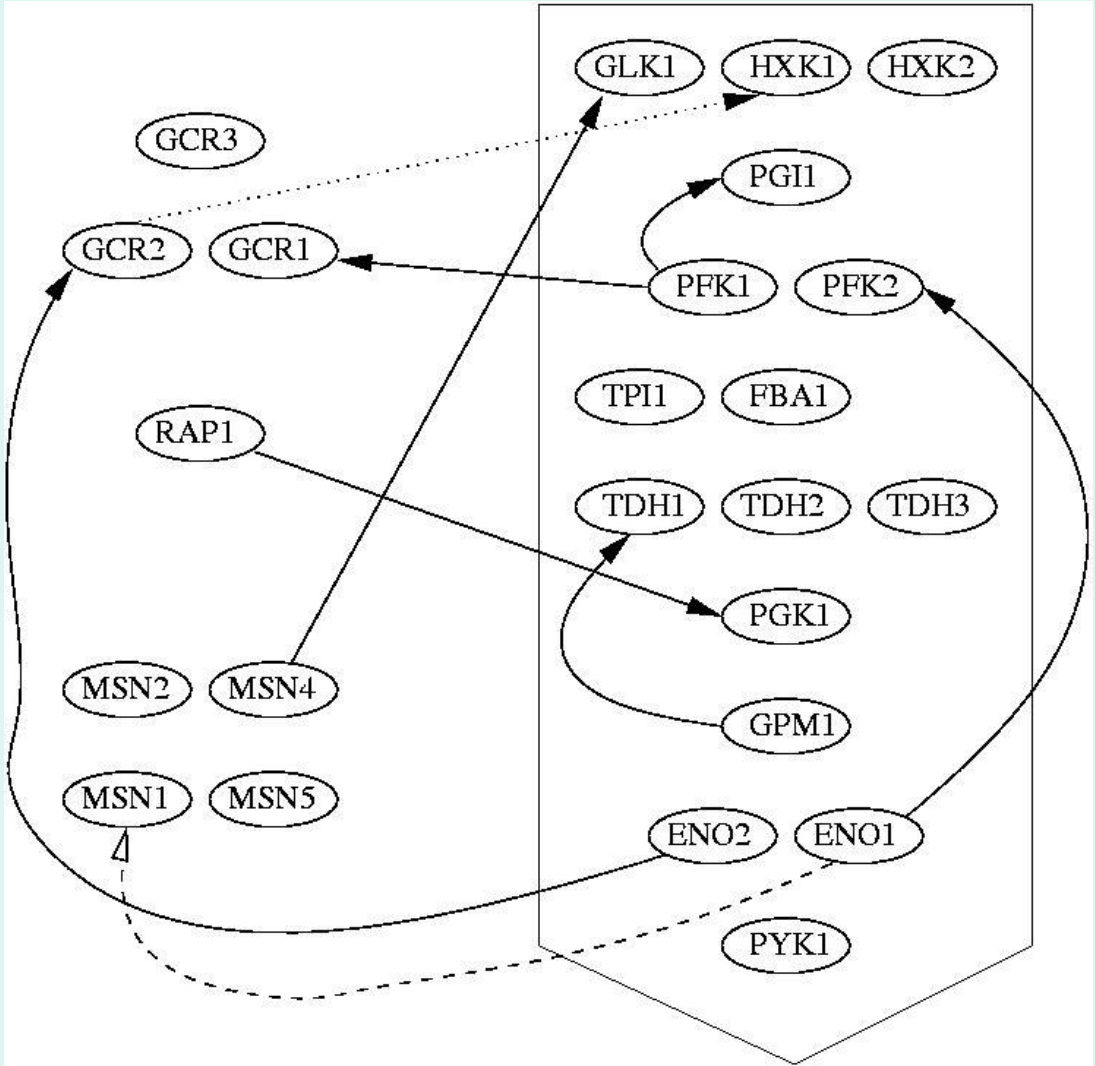


Genomic expression ratios of yeast heat shock response

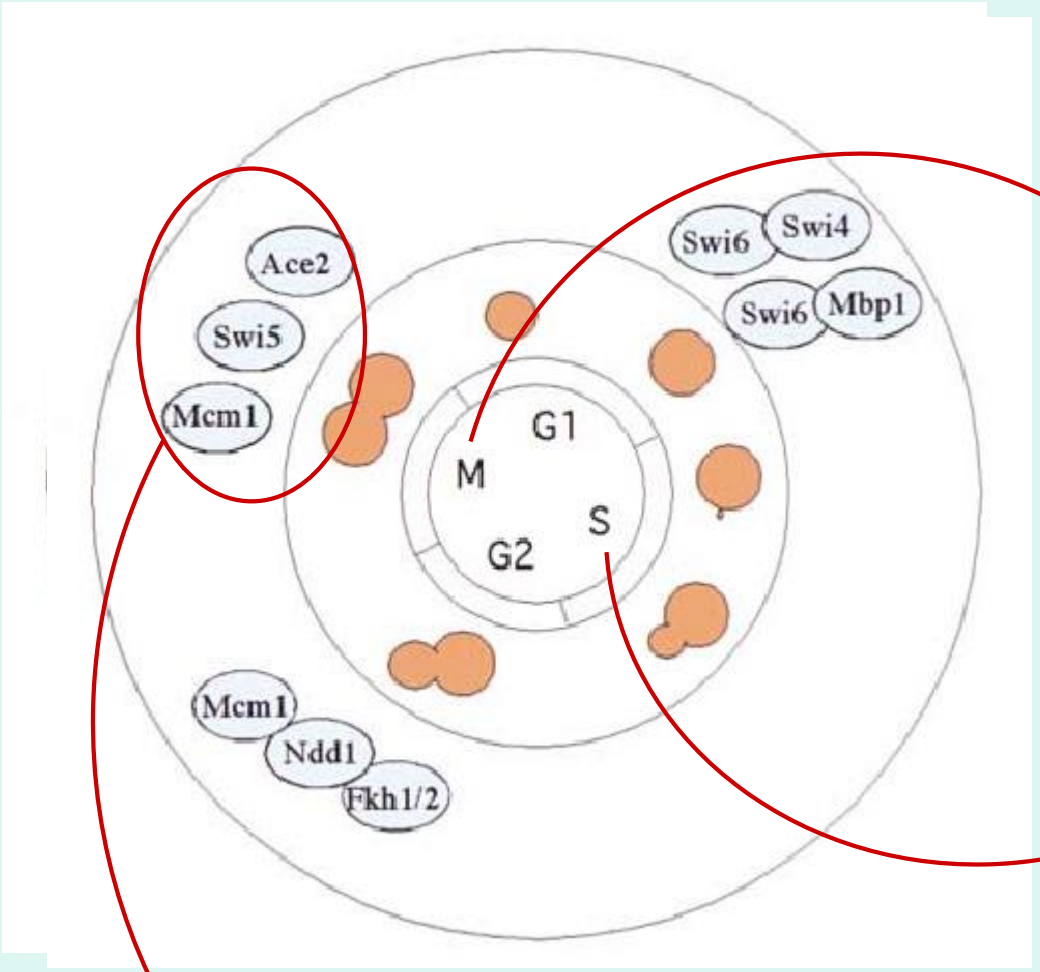
σ=0.39



Reconstructed yeast glycolytic pathway



Yeast Cell Cycle

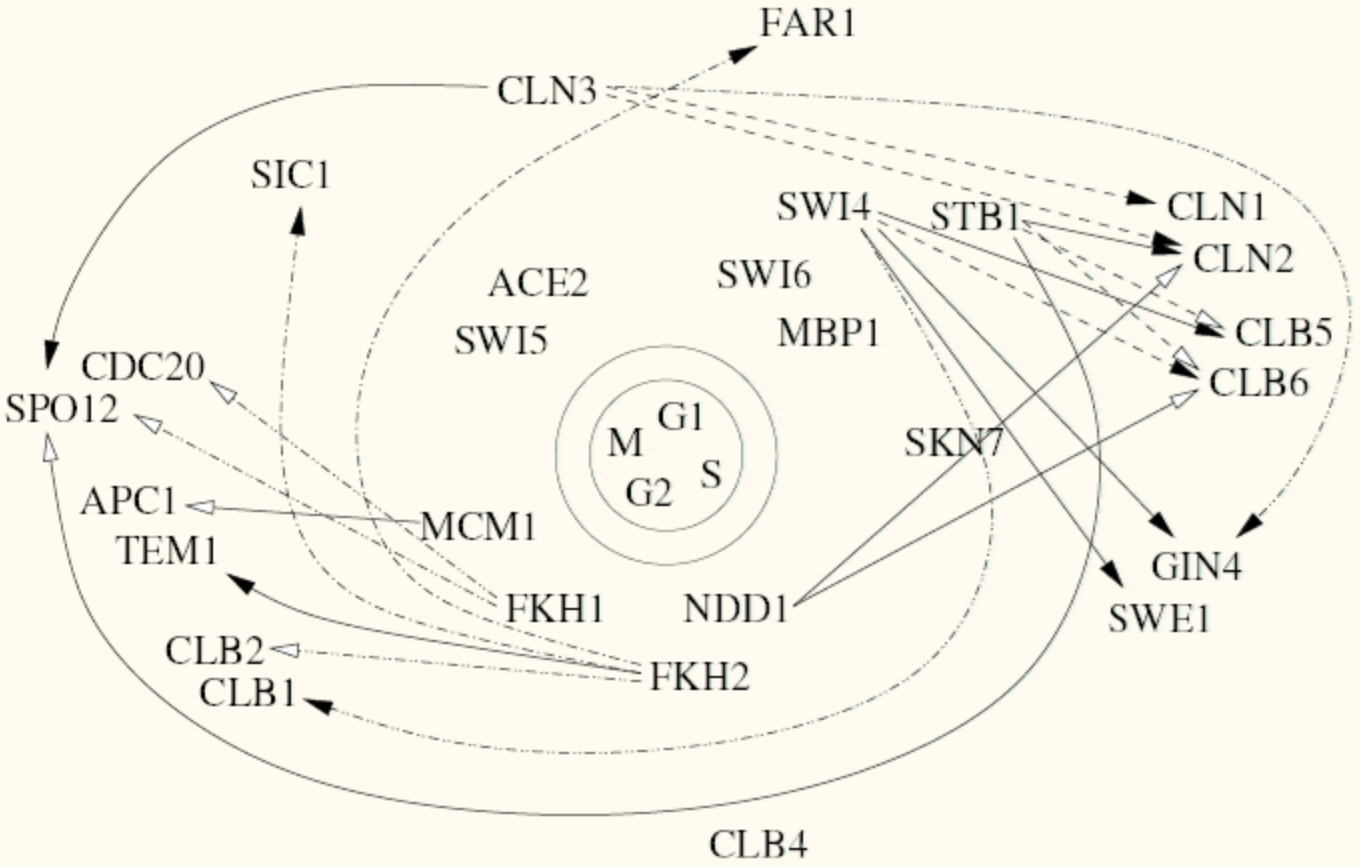


※ ~800 *genes* in the yeast cell cycle have been identified by Stanford group

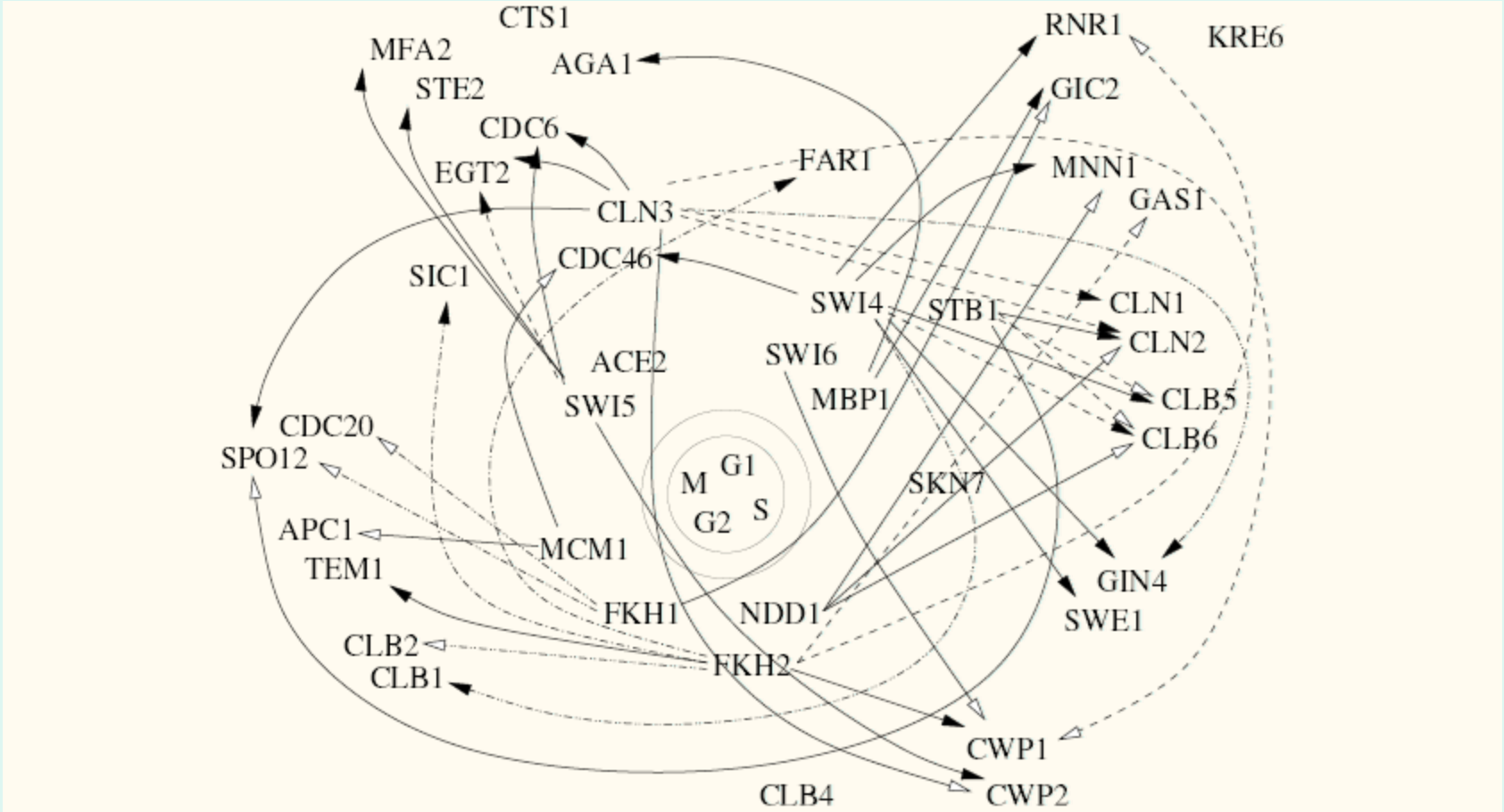
Mitosis phase
(cell division)

Synthesis phase
(*DNA* synthesis)

Transcription factor
(binding on the *gene*'s *promoter* to either activate the *gene*'s transcription, or repress it)



27: The reconstructed model of regulation of the yeast cell cycle dependent genes by the transcription factors. Solid, dashed, dotted, dotted dashed, and dotted dotted dashed line represents influence of regulation from time point $t, t-1, t-2, t-3$, and $t-4$, respectively. The sign of the w's are positive (negative) for solid (empty) arrow heads. See text for a discussion of the up or down regulation.



41: The reconstructed model of regulation of the yeast cell cycle dependent genes by the transcription factors. Solid, dashed, dotted, dotted dashed, and dotted dotted dashed line represents influence of regulation from time point $t, t-1, t-2, t-3$, and $t-4$, respectively. The sign of the w's are positive (negative) for solid (empty) arrow heads. See text for a discussion of the up or down regulation.

Features

- A **parsimonious** reconstruction:
  less noisy data → *more* reconstructed arcs

- Nonlinear model:
  power-law formalism is robust

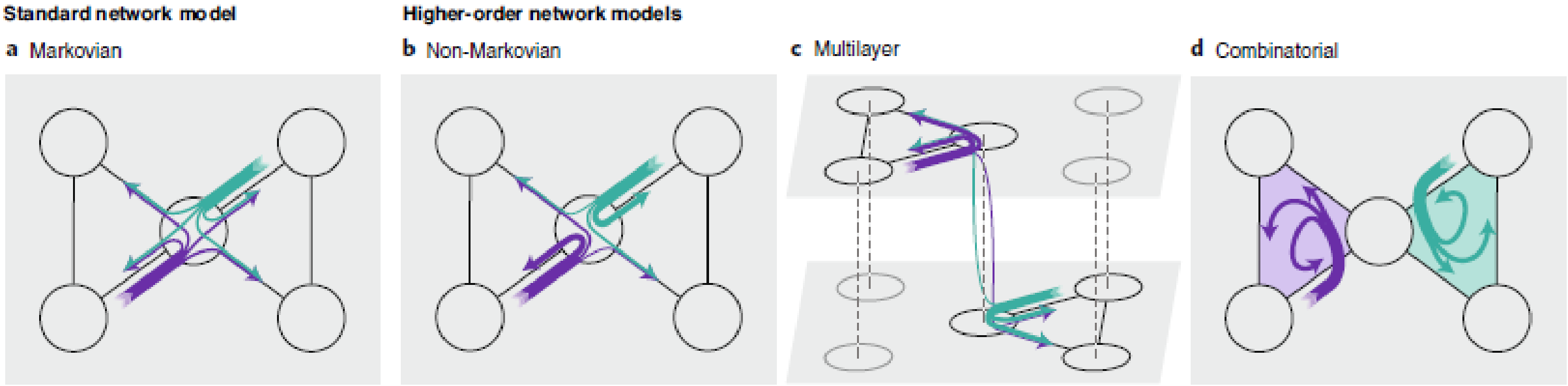
Higher Order Graphs

Some papers on higher order graphs:

- Scholtes, I. et al. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. Nature Communications 5, 5024 (2014).

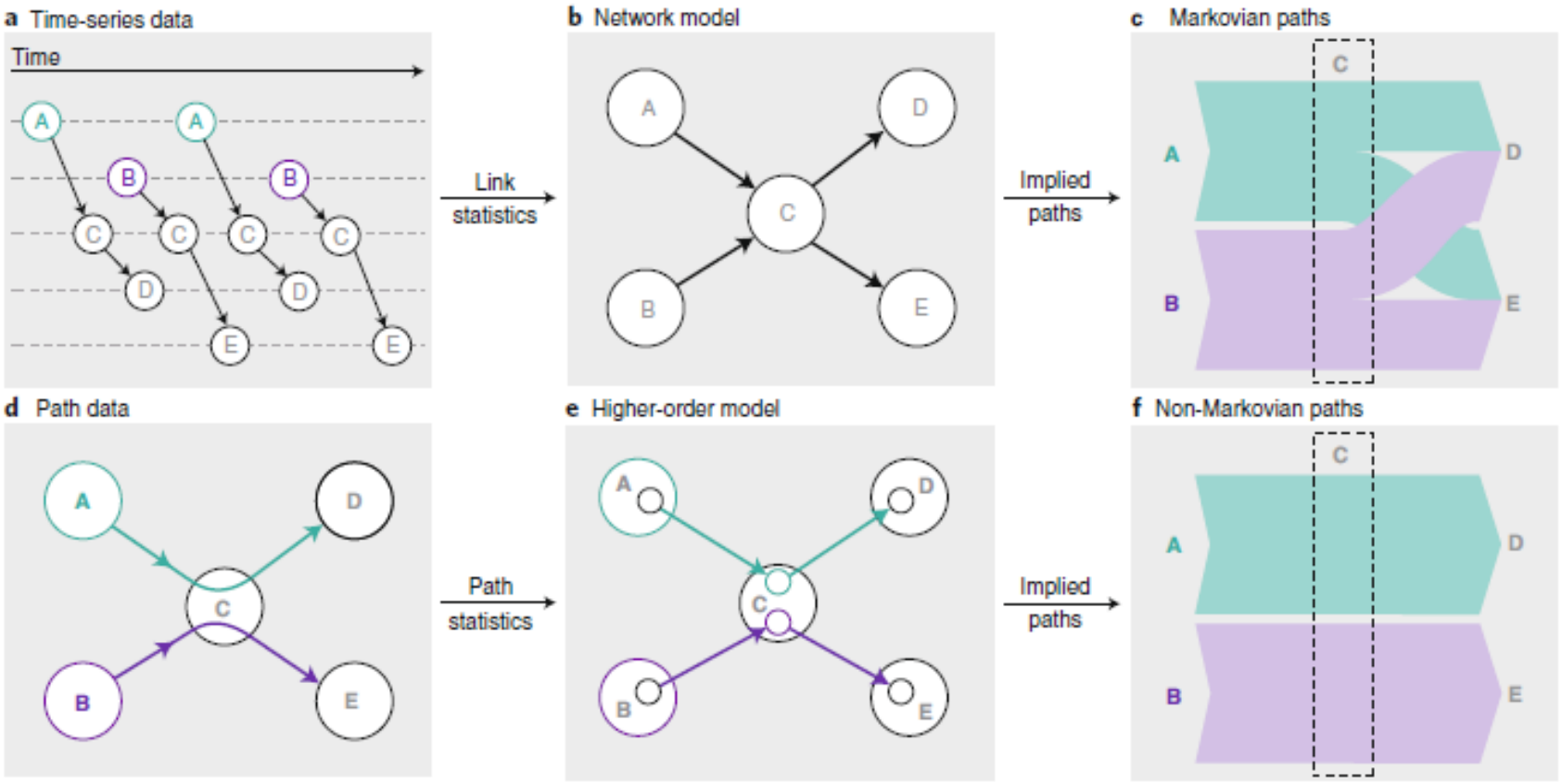
- Lambiotte, R. et al. From networks to optimal higher-order models of complex systems, Nature Physics 15 (2019) 313–320.

Warm-Up Examples:



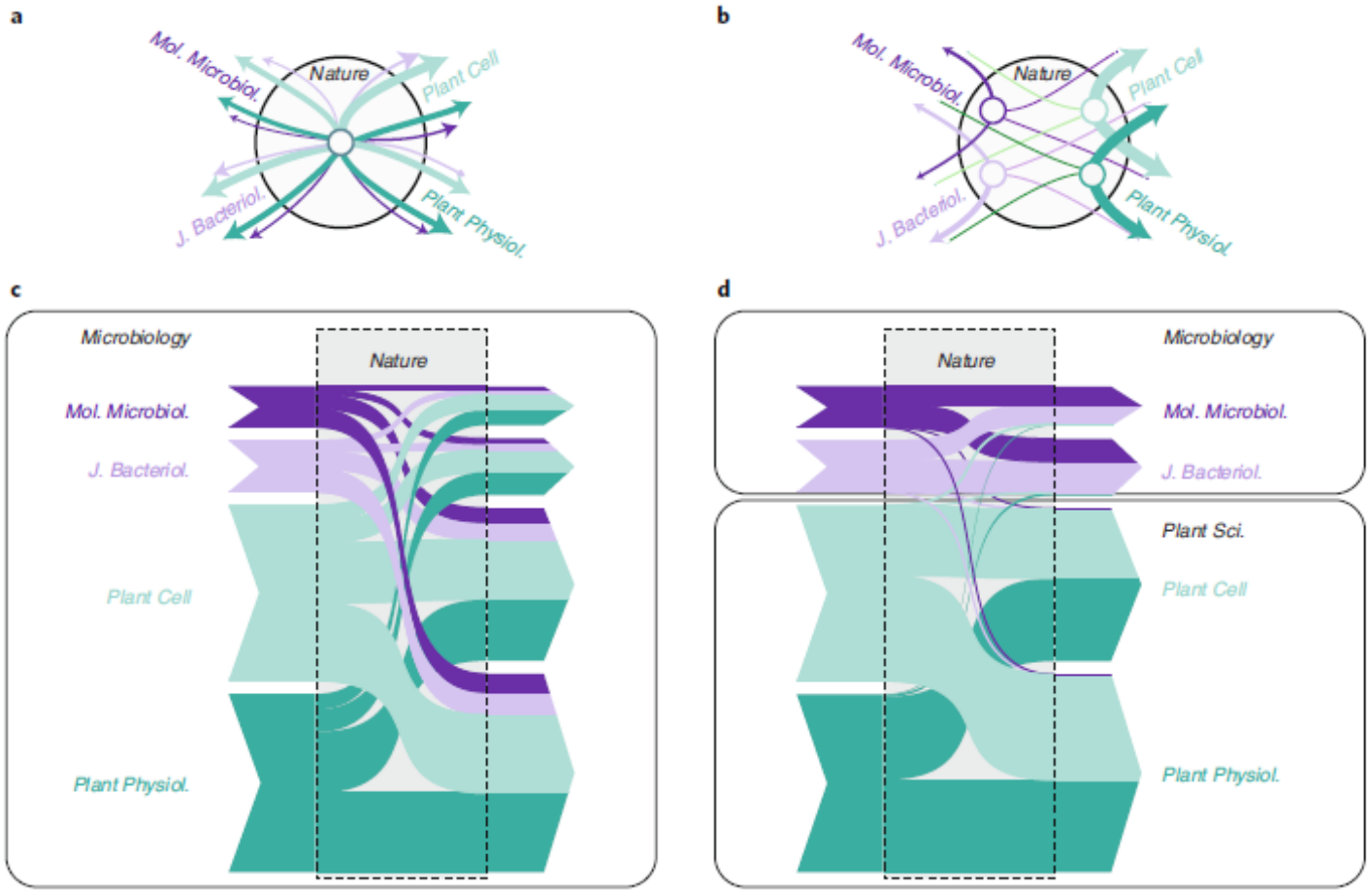
Different approaches to model an ego network with higher-order dependencies between nodes. **a**–**d**, Ego (central node) communicates by different means with two friends (left nodes) and two colleagues (right nodes). Green and purple arrows highlight paths from one friend (purple) and one colleague (green) through ego. To which nodes these paths can continue depends on the constraints set by a standard network model with Markovian dynamics (**a**), a non-Markovian network model (**b**), a multilayer network with Markovian dynamics within layers (**c**) and a simplicial complex where the paths move between links that share a triangle (**d**). The thickness of the arrows indicates the volume of flows between nodes.

Warm-Up Examples:



Non-Markovian higher-order models can better capture the topology of paths in complex systems.
a, A rich source of path information is time-series data that capture interaction sequences between the components of a system. **b**,**c**, Focusing on pairwise interactions, network models abstract a system's topology with nodes and links (**b**) while assuming that paths are transitive and Markovian (**c**). **d**, Due to the chronological ordering of interactions, the actual paths of indirect influence in time-series data can deviate from this assumption. **e**,**f**, Focusing on paths rather than pairwise interactions, higher-order network models with, for example, state nodes (**e**) can capture the actual topology of indirect influence (**f**).

Warm-Up Examples:



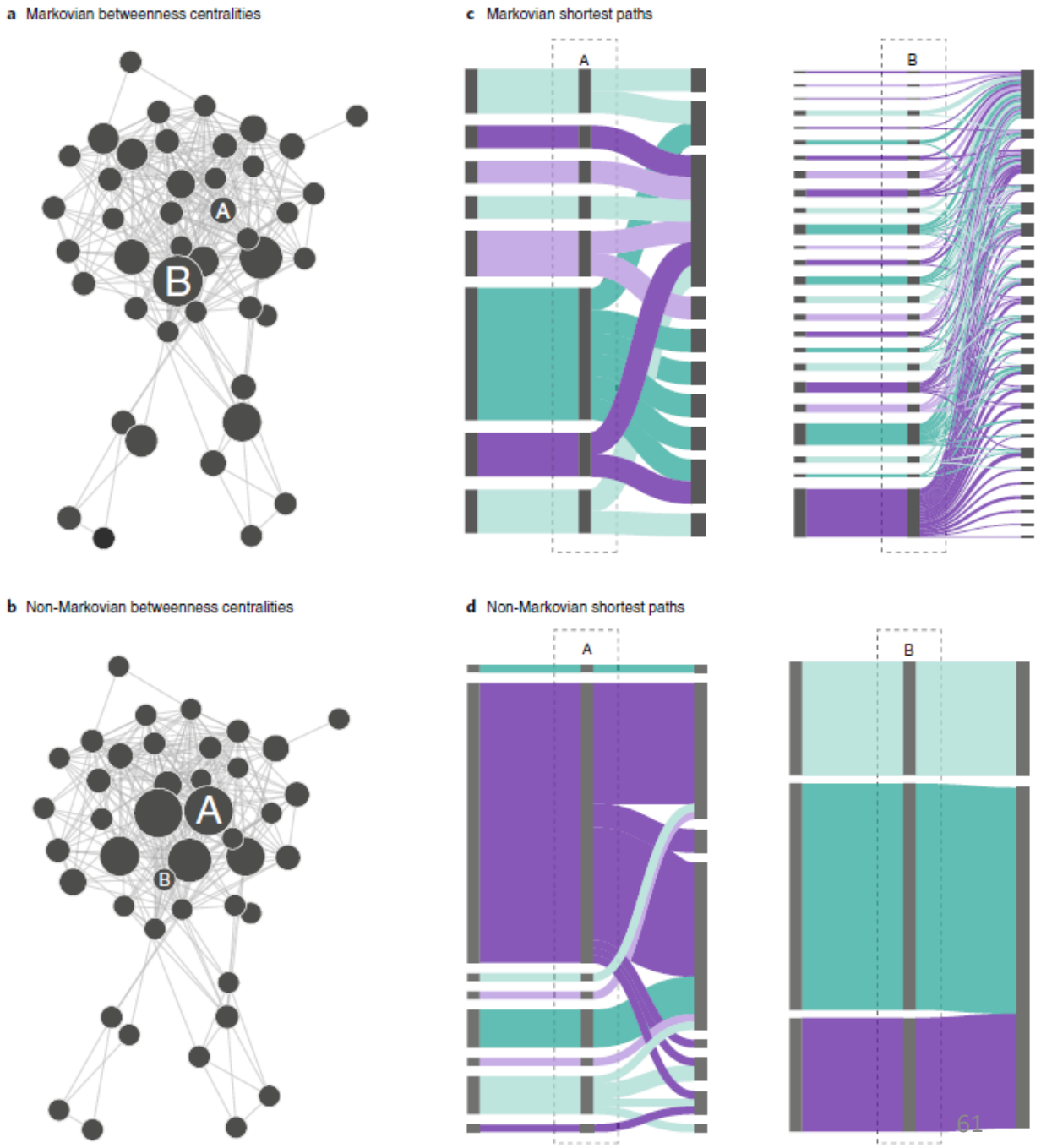
Community detection of paths can capture overlapping communities.

The underlying data from Thomson Reuters Web of Science are chains of citing articles aggregated in journals, like in the previous figure with nodes interpreted as articles in journals A–E. **a**, A standard first-order Markov representation of citation flows from four specialized journals through multidisciplinary *Nature*. **b**, A second-order representation with one state node for each citing journal. **c**, The standard network representation mixes flows and washes out the boundary between fields. **d**, A second-order Markov model captures the fact that citation flows through a multidisciplinary journal depends on where they come from and highlights overlapping fields in *Nature*.

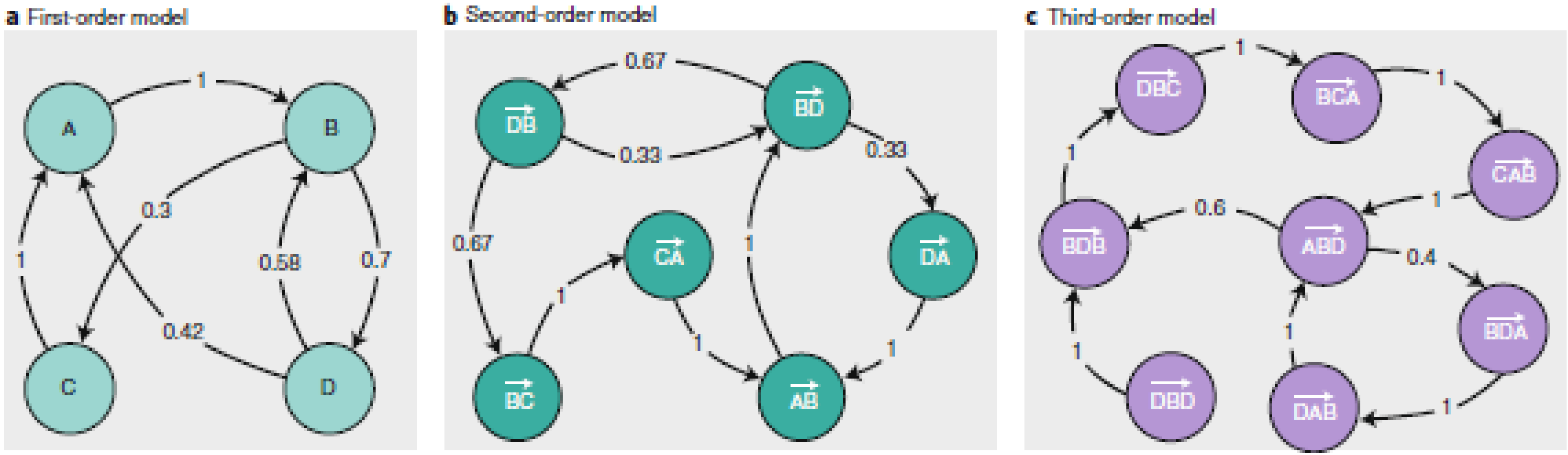
Warm-Up Examples:

Non-Markovian paths change the centrality of nodes in time-stamped social network data.

a,**b**, Betweenness centralities calculated based on shortest paths in a network model of time-stamped interactions (**a**) do not capture the true importance of nodes calculated based on causal paths that respect causality in the underlying time-series data (**b**). **c**,**d**, The alluvial diagrams highlight the fact that the chronological order of interactions alters the shortest causal paths passing through nodes A and B (**d**), compared with what we would expect based on the topology of direct interactions (**c**), thus considerably changing the betweenness centrality of nodes.

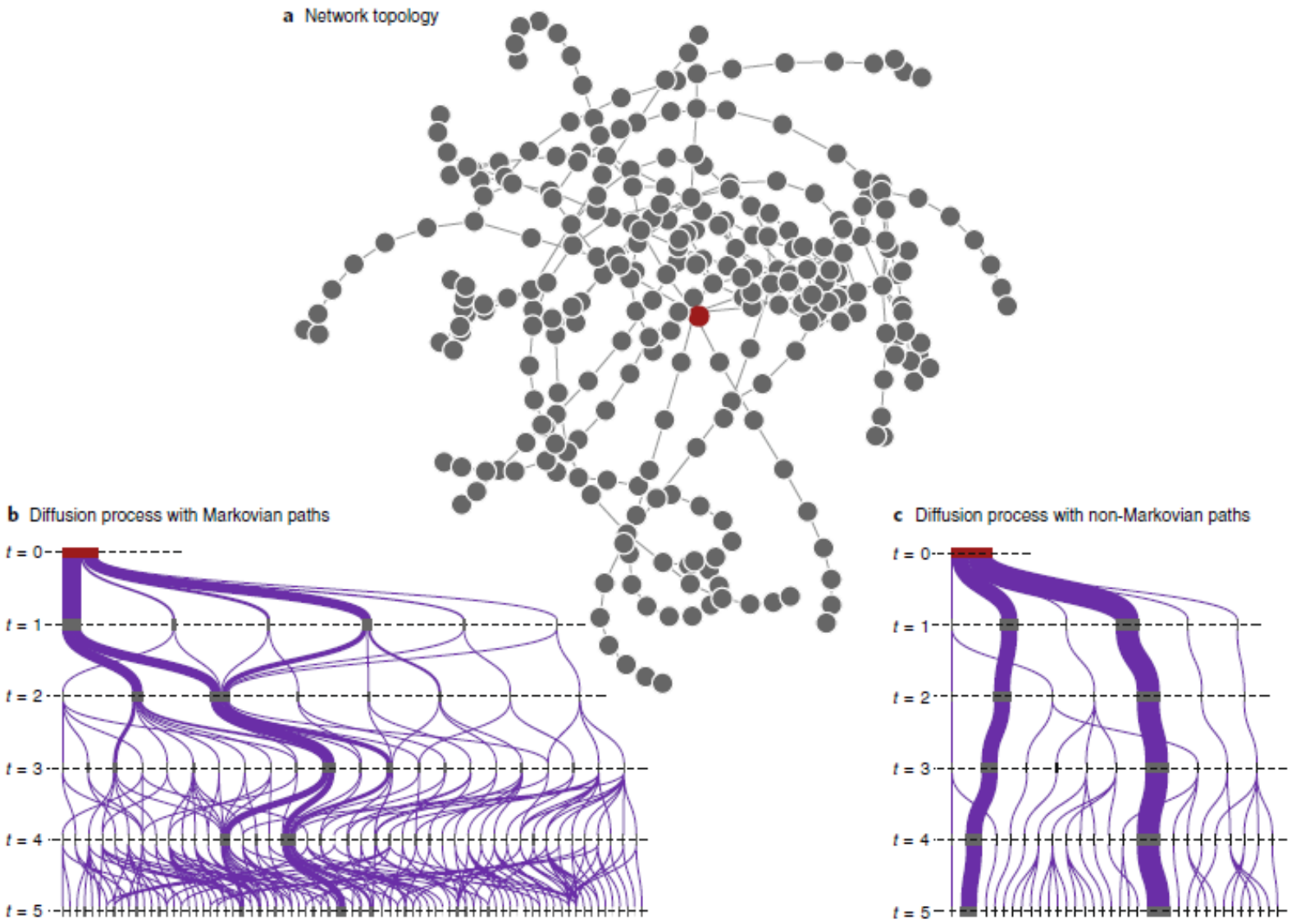


Warm-Up Examples:



De Bruijn graphs with m dimensions help generalize network analytic methods to higher-order models.
a, First-order model with dimensionality $m = 1$ for a set of observed causal paths between four nodes A, B, C and D. **b**, Second-order model with $m = 2$. **c**, Third-order model with $m = 3$. Starting from a first-order network model, higher-order models can be generated by an iterative line graph construction. The absence of transitions that correspond to a possible transitive path in the underlying first-order network, such as $\overrightarrow{BDB} \rightarrow \overrightarrow{DBD}$, indicates constraints in the observed paths that change the causal topology of the system.

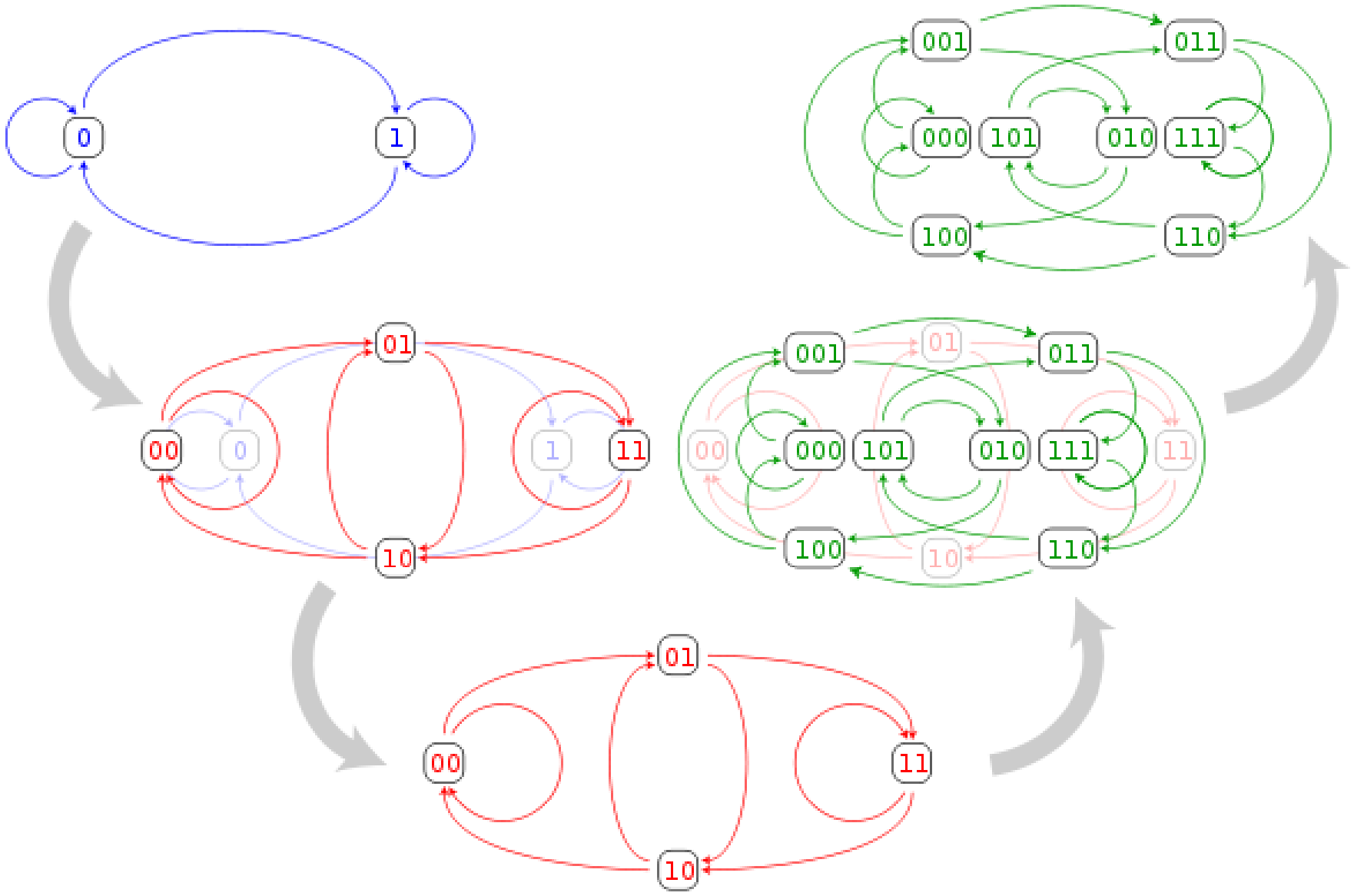
Warm-Up Examples:



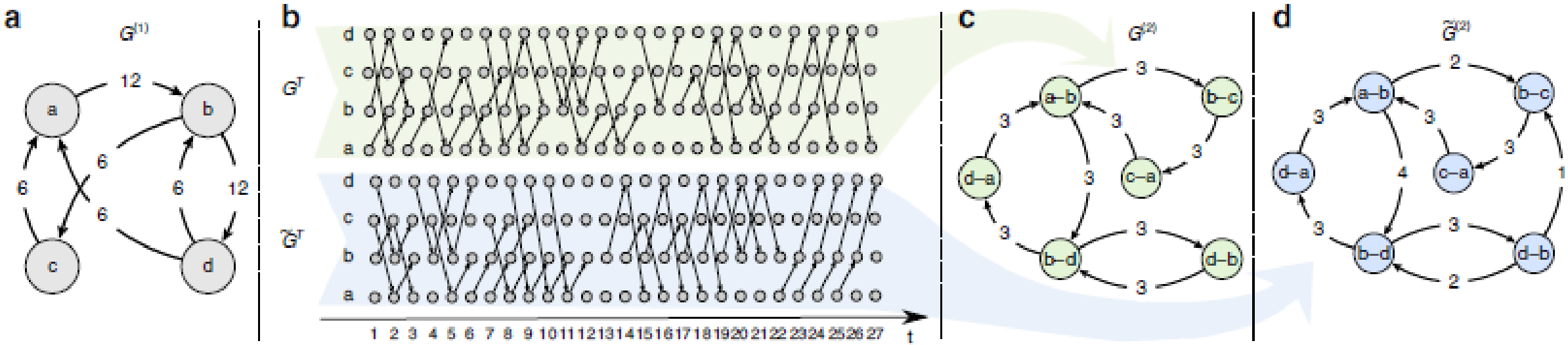
Non-Markovian paths in networked systems influence the evolution of diffusion processes.

a, Network model of the London Tube system, where links capture direct train connections between stations. **b**,**c**, The flow diagrams show the first five steps of a discrete-time diffusion process starting in node highlighted in red in **a**. The widths of flows capture the number of passengers moving on paths between particular nodes in the process. While **b** shows the dynamics of the process using transitive and Markovian paths in the network, **c** shows the evolution of diffusion across the non-Markovian paths created by the specific ordering of train connections in the London Tube system. The causal topology created by such non-Markovian paths influences dynamical processes and challenges our understanding of real complex systems.

De Buijin Graph explained



Causality in Higher-Order Graphs: An Example



Two temporal networks with the same first-order, but different second-order time-aggregated networks. (a) Time-aggregated network G(1), whose edge weights capture the number of times each edge occurred in a temporal network. The time-aggregated network is consistent with both temporal networks shown in b. (b) Time-unfolded representations of two temporal networks, each consisting of 4 nodes and 27 time steps, both consistent with $G^{(1)}$. Differences in their causality structures are highlighted by the corresponding second-order aggregate networks shown in c and d. Both second order aggregate networks are consistent with $G^{(1)}$.