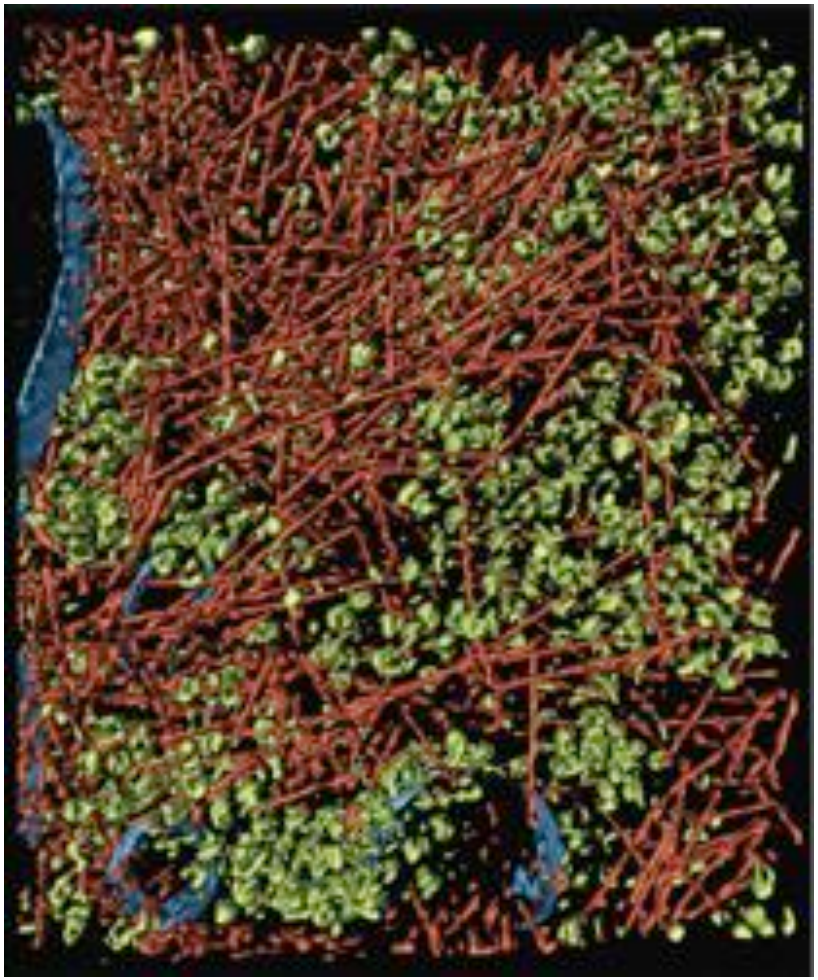
V 1 – Introduction

Mon, Oct 14, 2013

Bioinformatics 3 — Volkhard Helms

How Does a Cell Work?



Medalia et al, Science 298 (2002) 1209

A cell is a crowded environment
=> many different proteins,
metabolites, compartments, …

On a microscopic level
=> direct two-body interactions

At the macroscopic level
=> complex behavior

Can we understand the behavior from the interactions?

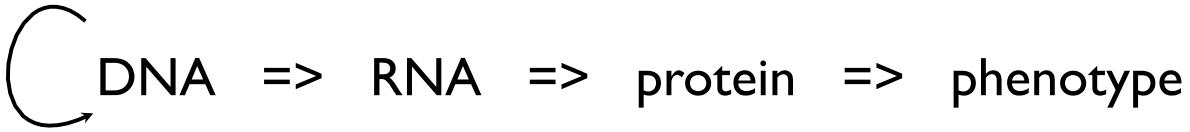
=> Connectivity

The view of traditional molecular biology

Molecular Biology: "One protein — one function"

mutation => phenotype

Linear one-way dependencies: regulation at the DNA level, proteins follow



Structural Biology: "Protein structure determines its function"

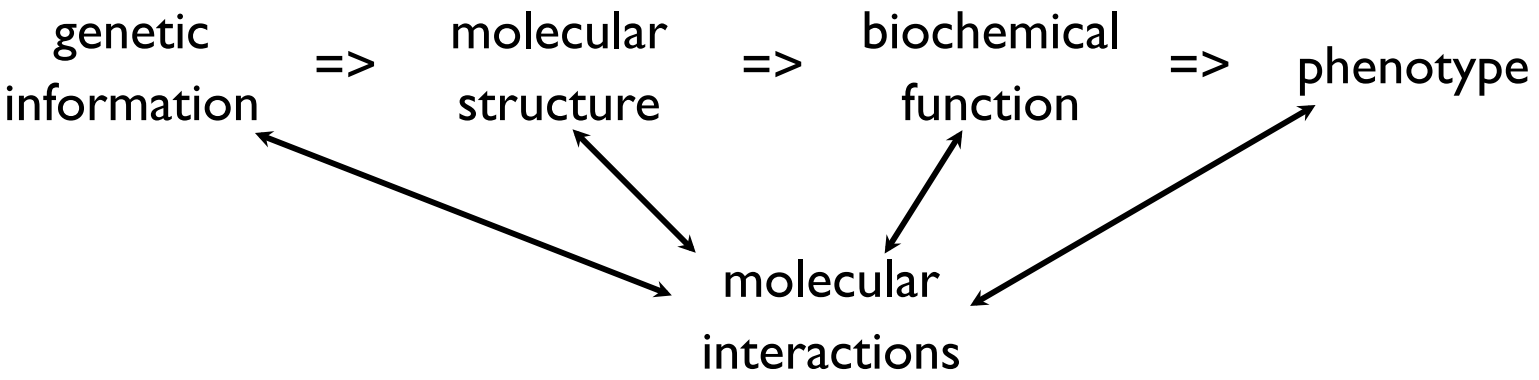
biochemical conditions => phenotype

No feedback, just re-action:

genetic information => molecular structure => biochemical function => phenotype

The Network View of Biology

Molecular Systems Biology: "It's both + molecular interactions"



➔ highly connected network of various interactions, dependencies

=> study networks

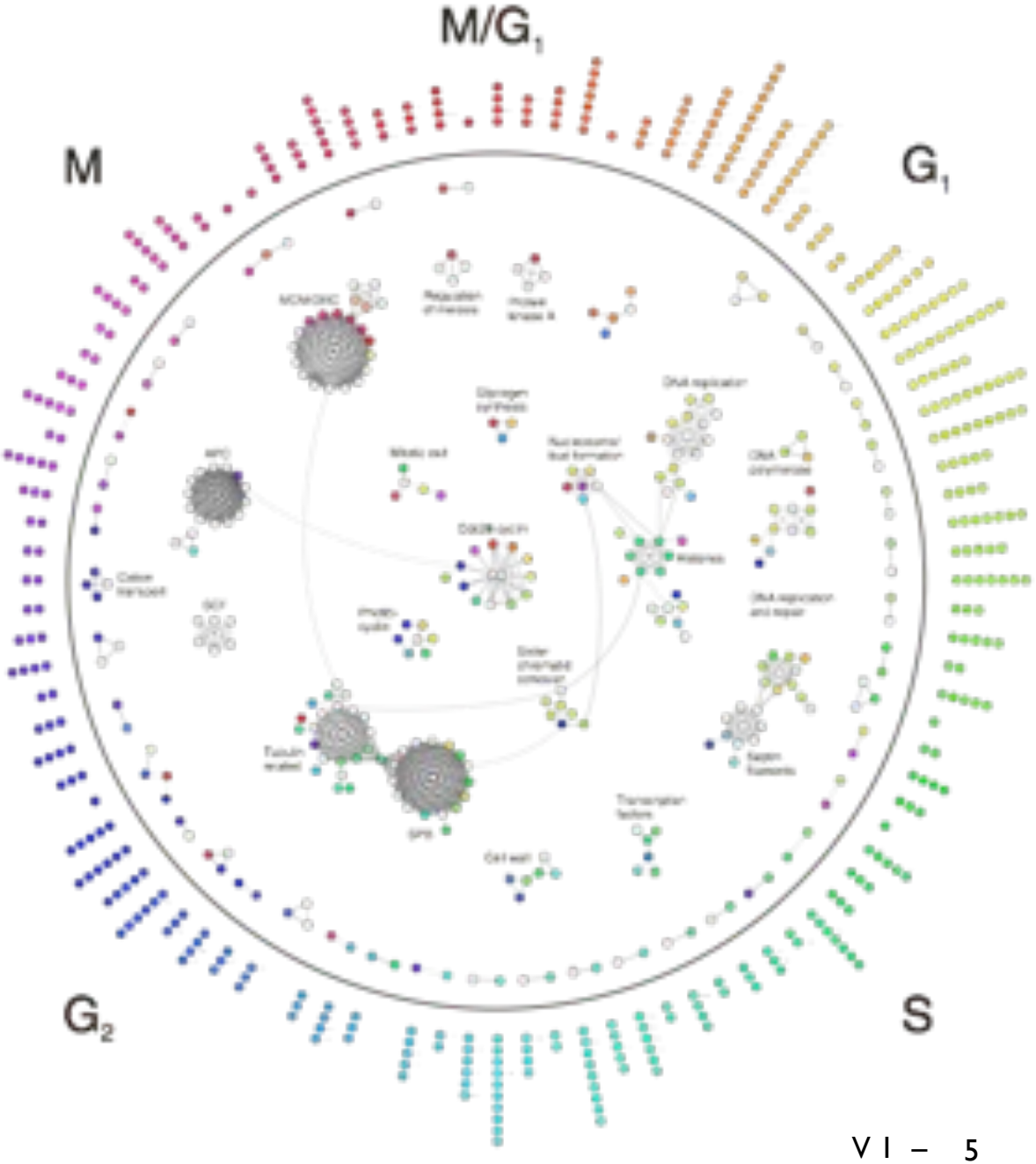
Example: Proteins in the Cell Cycle

From Lichtenberg et al, Science 307 (2005) 724:

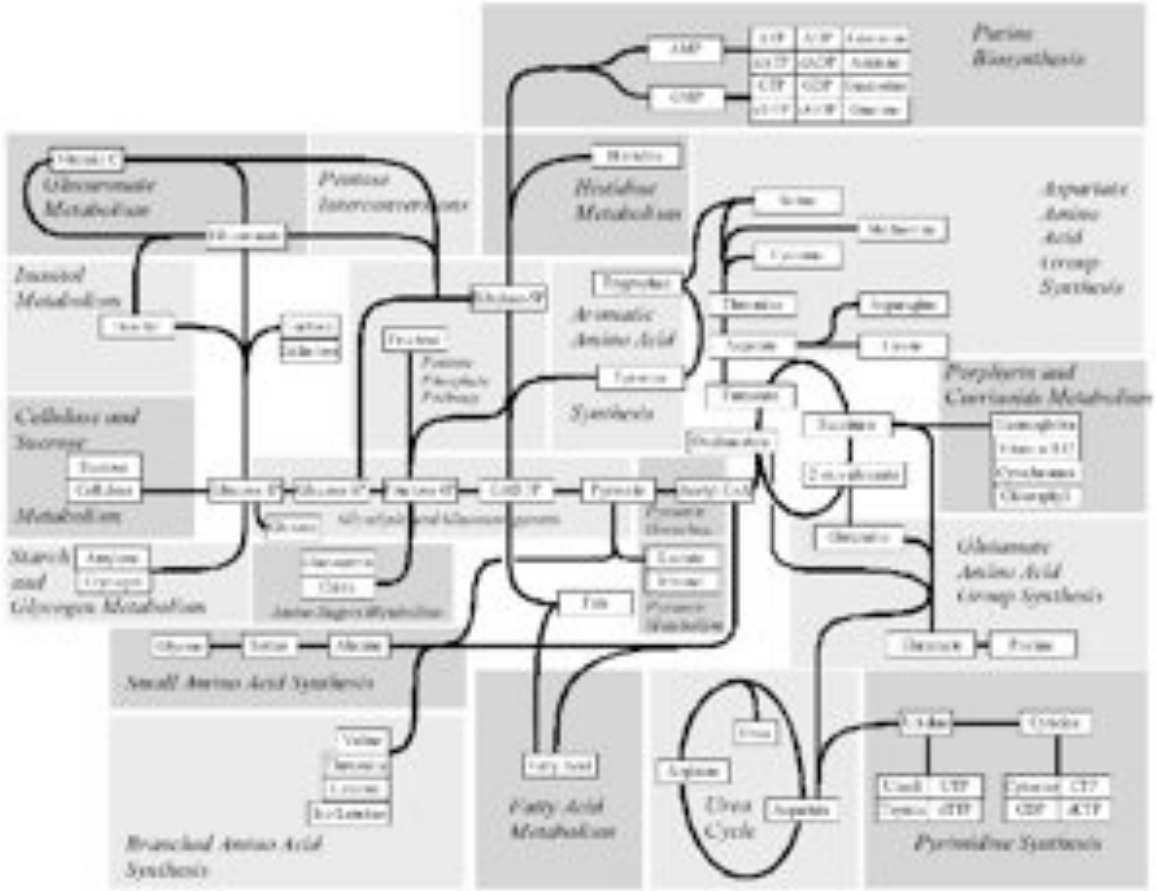
color coded assignment of proteins in time-dependent complexes during the cell cycle

=> protein complexes are transient

=> describe with a time dependent network



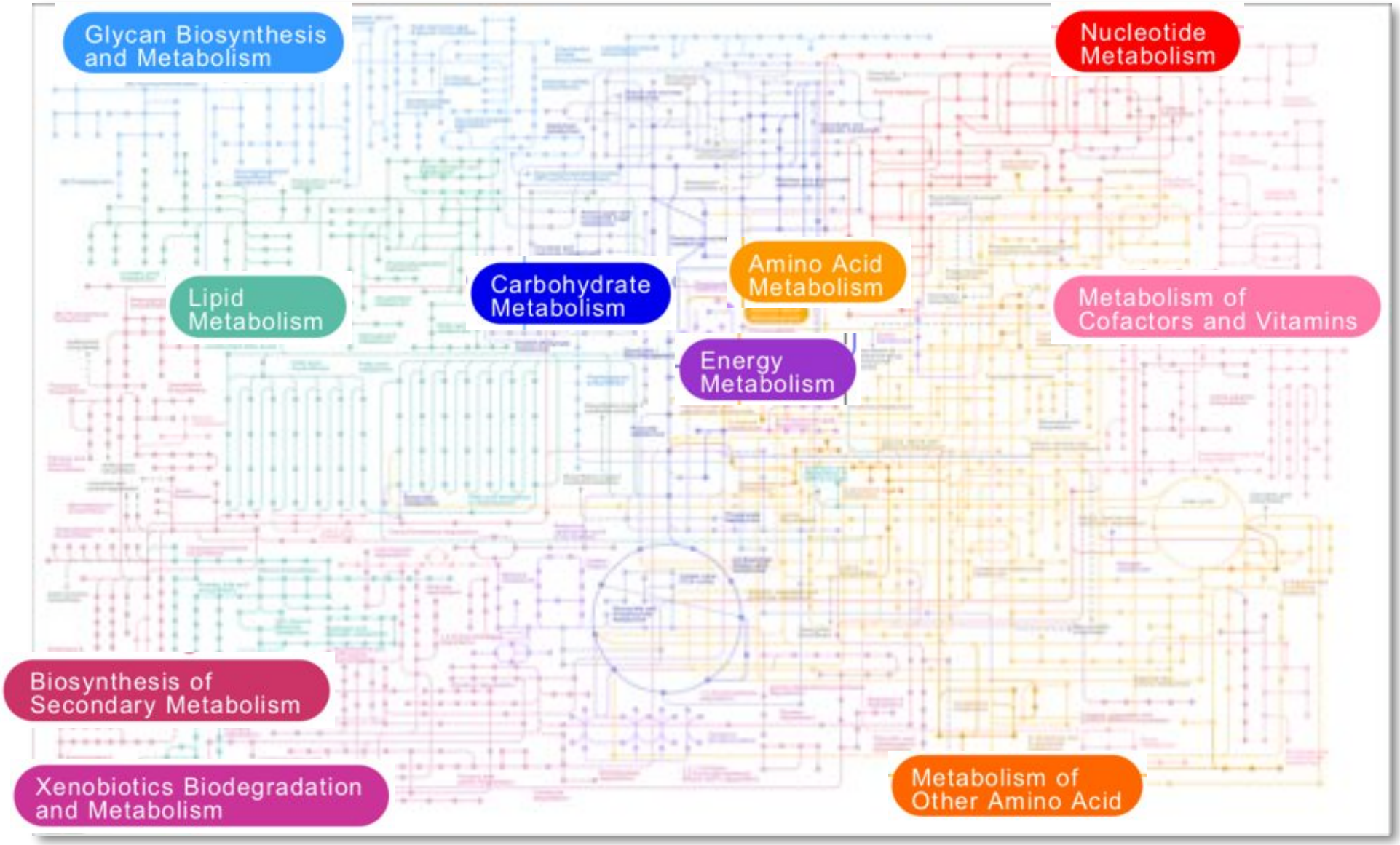
Major Metabolic Pathways



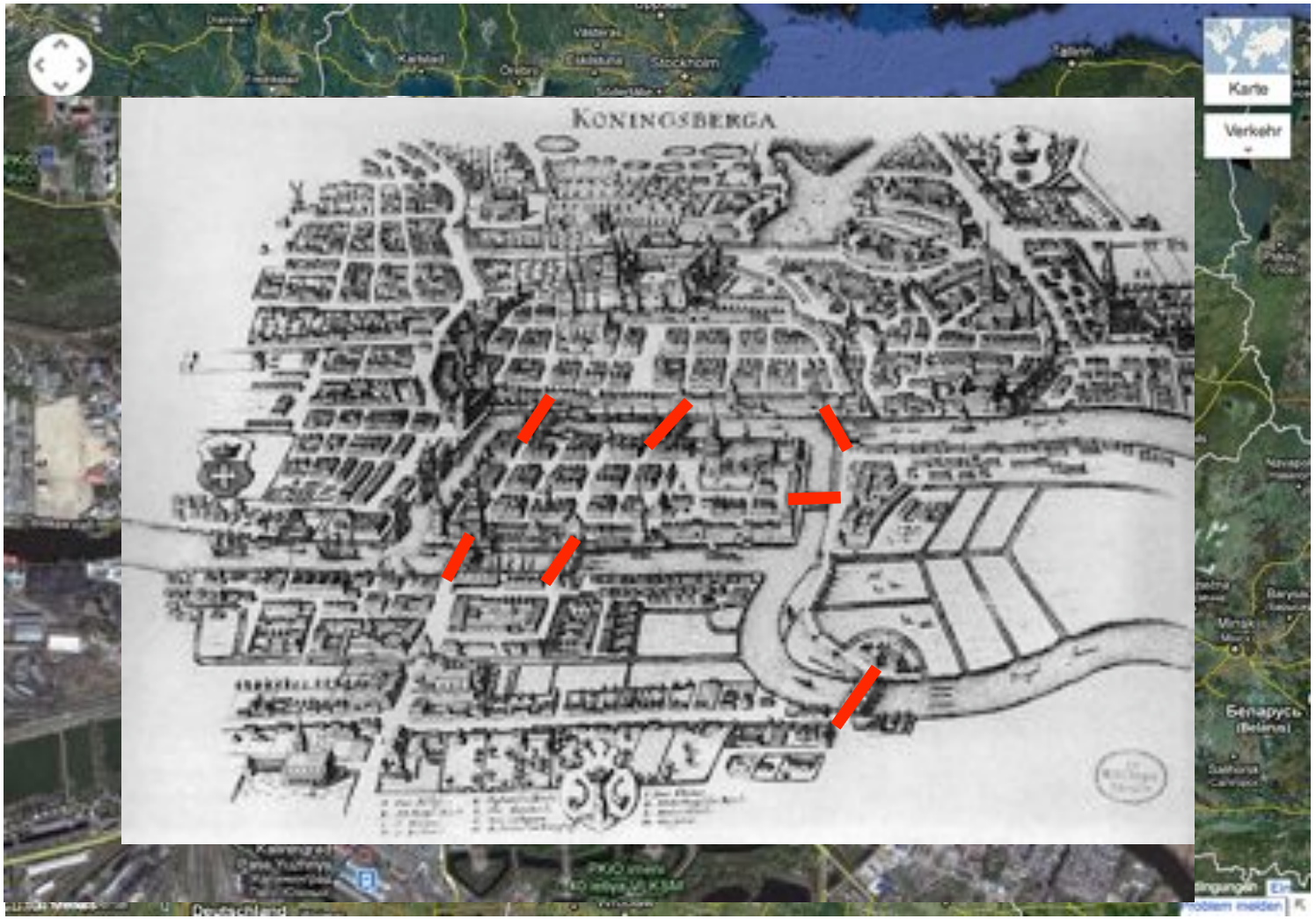
static connectivity <=> dynamic response to external conditions <=> different states during the cell cycle

http://www.mvv-muenchen.de/de/netz-bahnhoefe/netzplaene/index.html

Metabolism of *E. coli*

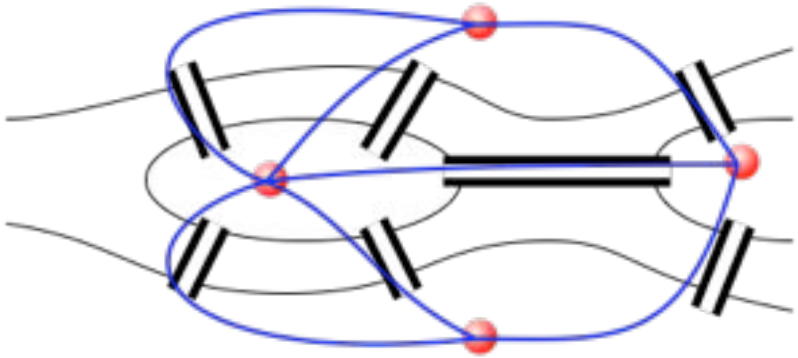


Euler @ Königsberg (1736)

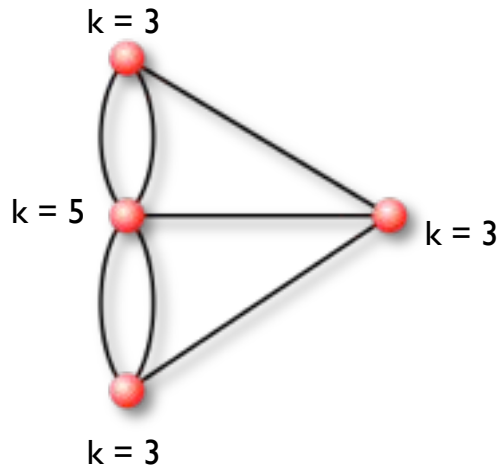


Can one cross all seven bridges once in one continuous (closed) path???

The Königsberg Connections



Turn problem into a graph:
i) each neighborhood is a node
ii) each bridge is a link
iii) straighten the layout



Continuous path
<=> ≤2 nodes with odd degree

Closed continuous path
<=> only nodes with even degree

see also: http://homepage.univie.ac.at/franz.embacher/Lehre/aussermathAnw/Spaziergaenge.html

Quantify the "Hairy Monsters"

Schwikowski, Uetz, Fields, *Nature Biotech.* **18**, 1257 (2001)

Network Measures:

- No. of edges, nodes
  => size of the network
- Average degree <*k*>
  => density of connections
- Degree distribution *P(k)*
  => structure of the network
- Cluster coefficient *C(k)*
  => local connectivity
- Connected components
  => subgraphs

Lecture – Overview

Protein complexes: **spatial** structure

=> experiments, spatial fitting, docking

Protein association:

=> interface properties, spatial simulations

Protein-Protein-Interaction Networks: **pairwise** connectivity

=> data from experiments, quality check

PPI: static network **structure**

=> network measures, clusters, modules, …

Gene regulation: cause and **response**

=> Boolean networks

Metabolic networks: steady state of **large networks**

=> FBA, extreme pathways

Metabolic networks / signaling networks: **dynamics**

=> ODEs, modules, stochastic effects

Systems Biology

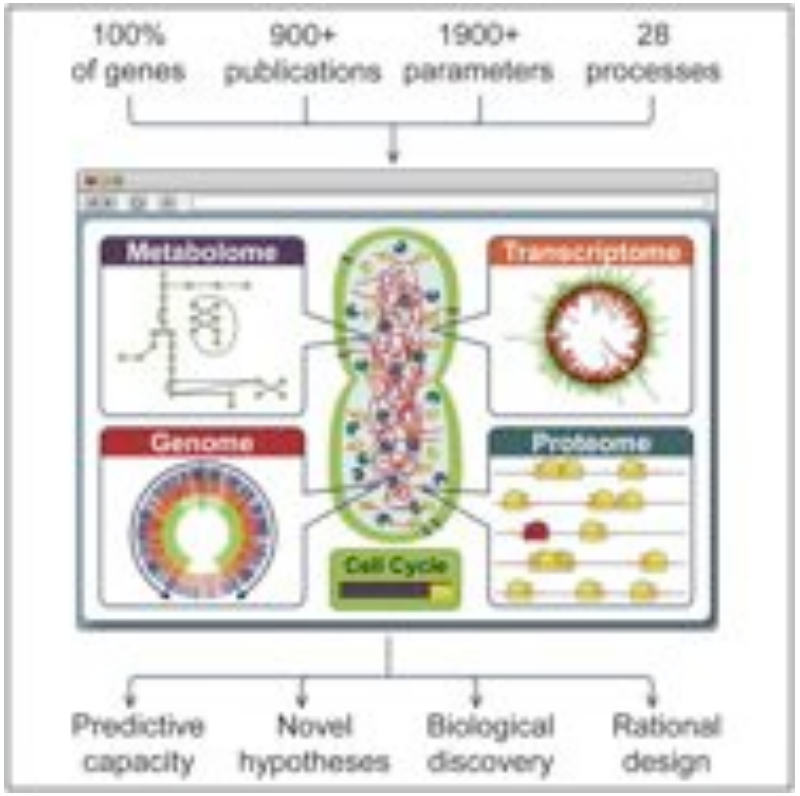
Appetizer: A whole-cell model for the life cycle of the human pathogen *Mycoplasma genitalium*

Theory

Cell

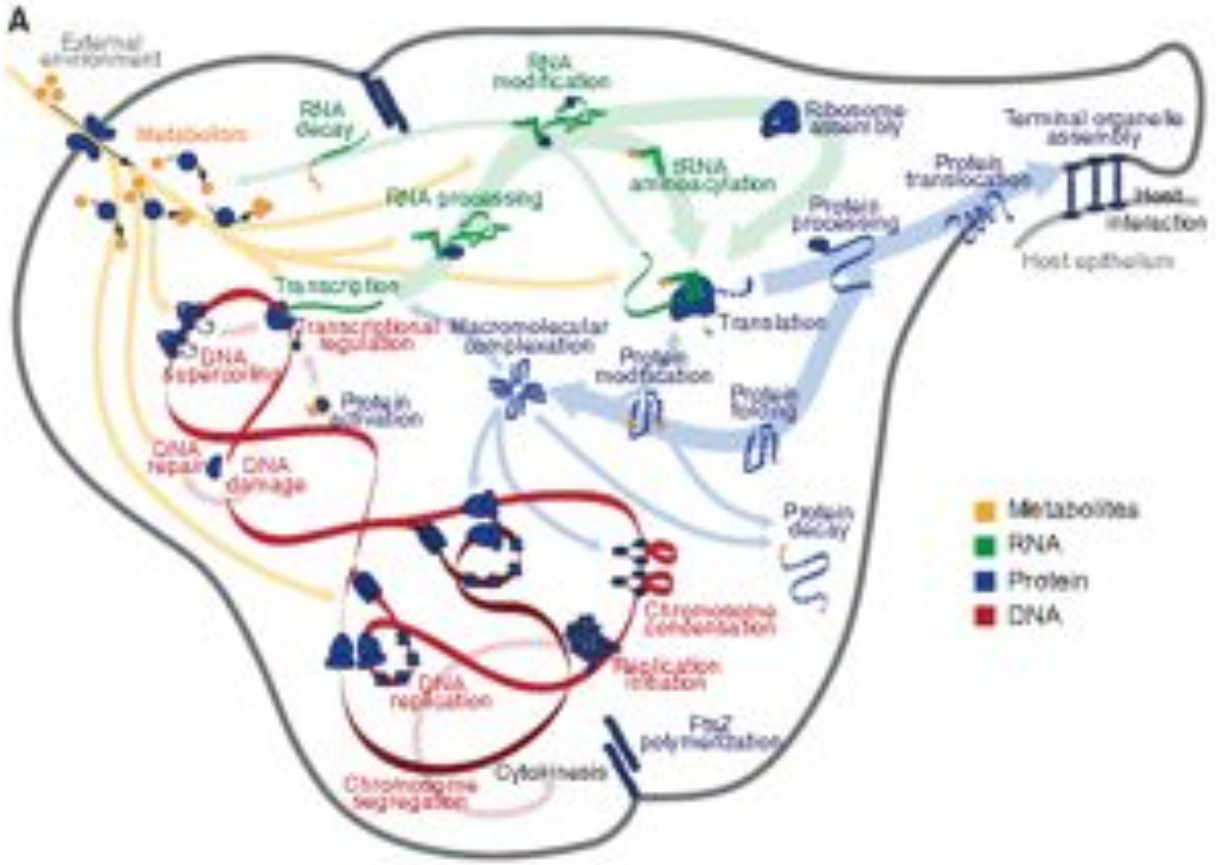
A Whole-Cell Computational Model Predicts Phenotype from Genotype

Jonathan R. Karr,[1,4] Jayodita C. Sanghvi,[2,4] Derek N. Macklin,[2] Miriam V. Gutschow,[2] Jared M. Jacobs,[2] Benjamin Bolival, Jr.,[2] Nacyra Assad-Garcia,[3] John I. Glass,[3] and Markus W. Covert[2,*]
[1]Graduate Program in Biophysics
[2]Department of Bioengineering
Stanford University, Stanford, CA 94305, USA
[3]J. Craig Venter Institute, Rockville, MD 20850, USA
[4]These authors contributed equally to this work
*Correspondence: mcovert@stanford.edu
http://dx.doi.org/10.1016/j.cell.2012.05.044



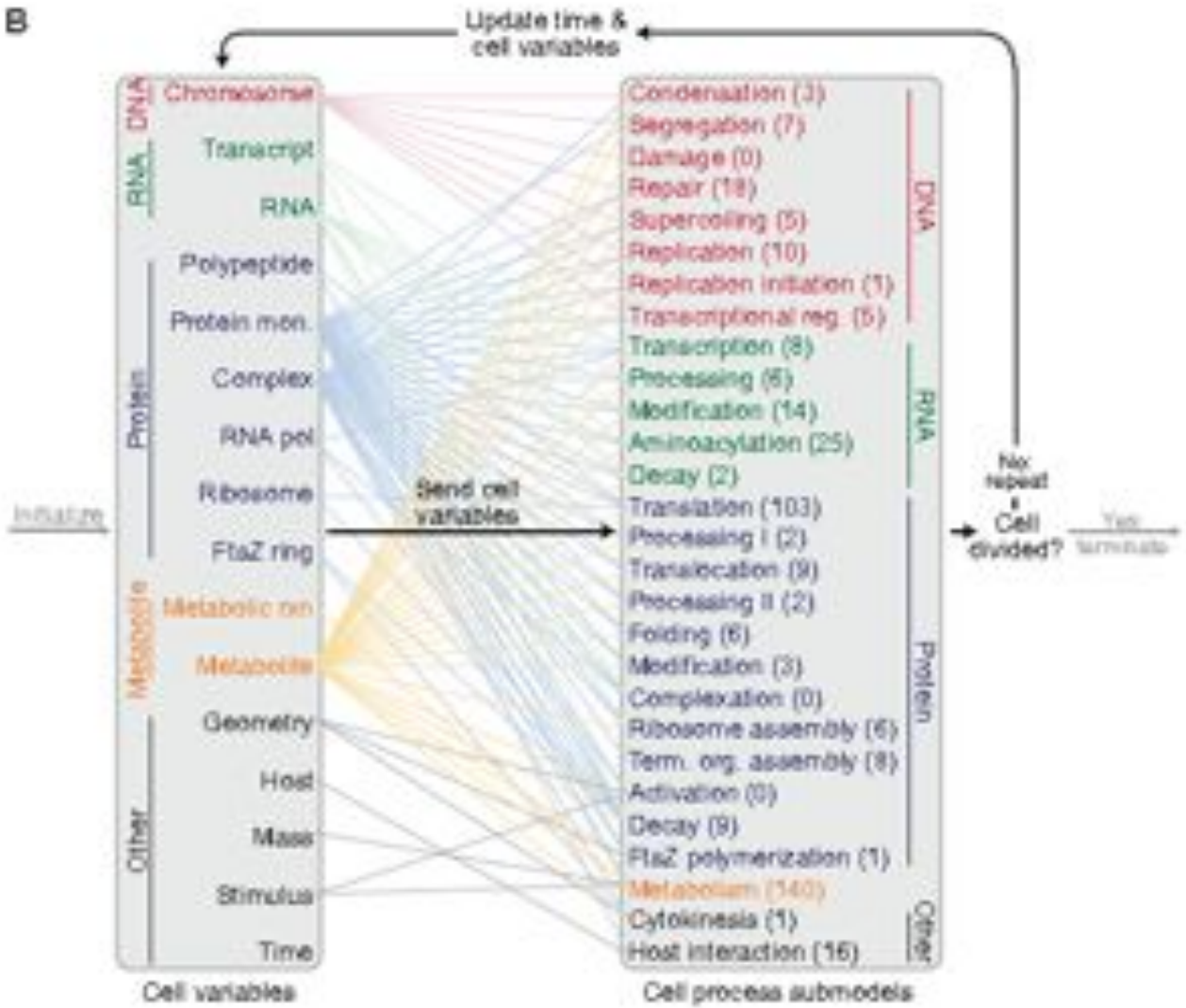
Cell 150, 389-401 (2012)

Divide and conquer approach (Caesar): split whole-cell model into 28 independent submodels



28 submodels are built / parametrized / iterated independently

Cell variables



System state is described by 16 cell variables

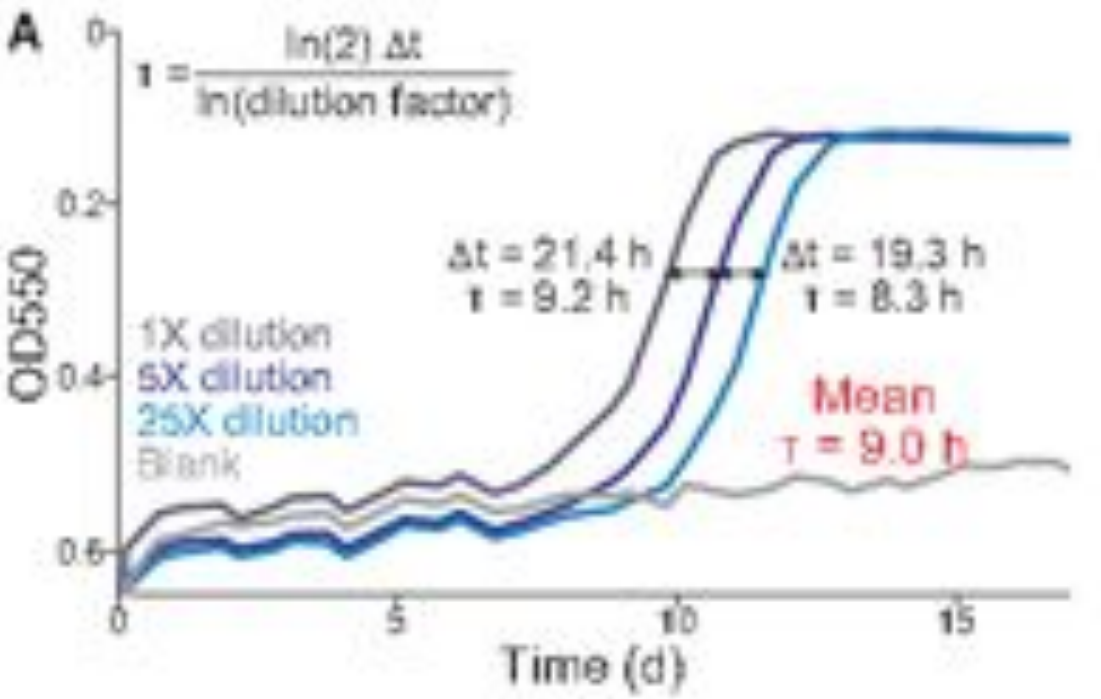
Colored lines: cell variables affected by individual submodels

Mathematical tools:
-Differential equations
-Stochastic simulations
-Flux balance analysis

List S1. Primary sources of the *M. genitalium* reconstruction.

Data source	Content
Bernstein *et al.*, 2002[602]	mRNA half-lives
BioCyc[6]	Genome annotation, metabolic reactions
BRENDA[570]	Reaction kinetics
CMR[668]	Genome annotation
Deuerling *et al.*, 2003[388]	Chaperone substrates
DrugBank[947]	Antibiotics
Eisen *et al.*, 1999[891]	DNA repair
Endo *et al.*, 2007[391]	Chaperone substrates
Feist *et al.*, 2007[556]	Metabolic reactions
Glass *et al.*, 2006[193]	Gene essentiality
Güell *et al.*, 2009[418]	Transcription unit structure
Gupta *et al.*, 2007[280]	N-terminal methionine cleavage
KEGG[113]	Genome annotation, orthology
Kerner *et al.*, 2005[389]	Chaperone substrates
Krause *et al.*, 2004[609]	Terminal organelle assembly
Lindahl *et al.*, 2000[462]	DNA damage
Morowitz *et al.*, 1962[870]	Cell chemical composition
NCBI Gene[64,777]	Genome annotation
Neidhardt *et al.*, 1990[393]	Cell chemical composition
Peil, 2009[105]	RNA modification
PubChem[587]	Metabolite structures
SABIO-RK[100]	Reaction kinetics
Solabia[754–759]	Media chemical composition
Suthers *et al.*, 2009[610]	Metabolic reactions
UniProt[96]	Genome annotation
Weiner *et al.*, 2000[431]	Promoters
Weiner *et al.*, 2003[569]	mRNA expression

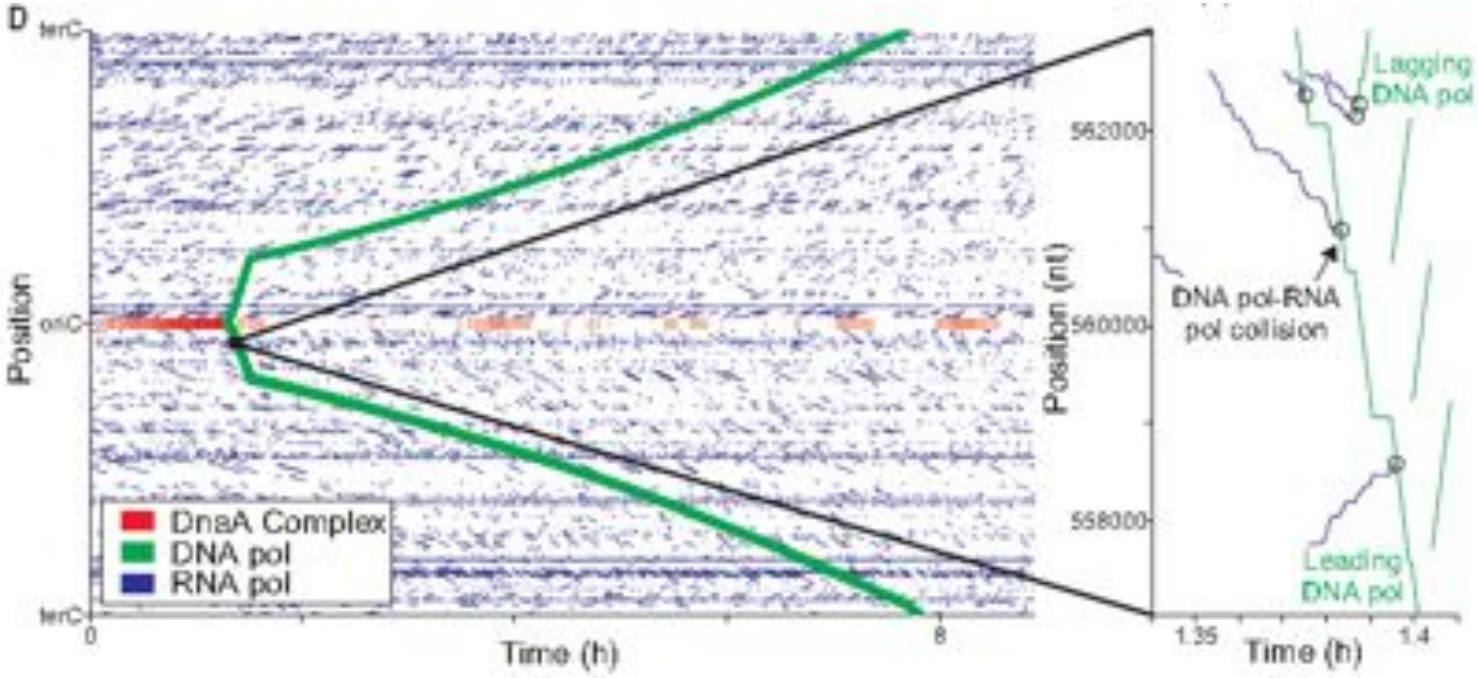
Growth of virtual cell culture



Growth of three cultures (dilutions indicated by shade of blue) and a blank control measured by OD550 of the pH indicator phenol red. The doubling time, t, was calculated using the equation at the top left from the additional time required by more dilute cultures to reach the same OD550 (black lines).

The model calculations were consistent with the observed doubling time!

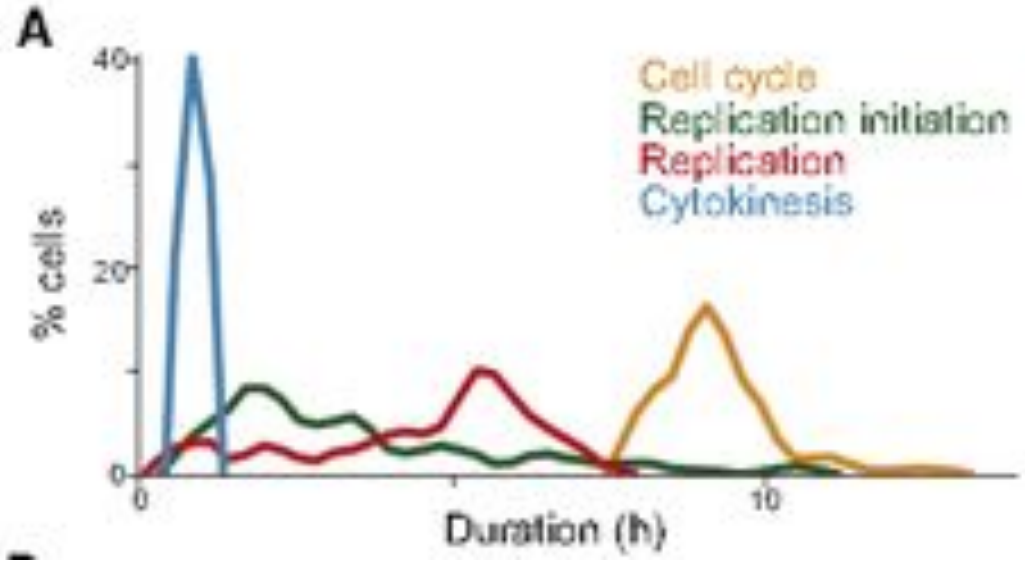
DNA-binding and dissociation dynamics



DNA-binding and dissociation dynamics of the oriC DnaA complex (red) and of RNA (blue) and DNA (green) polymerases for one in silico cell. The oriC DnaA complex recruits DNA polymerase to the oriC to initiate replication, which in turn dissolves the oriC DnaA complex. RNA polymerase traces (blue line segments) indicate individual transcription events. The height, length, and slope of each trace represent the transcript length, transcription duration, and transcript elongation rate, respectively.

Inset : several predicted collisions between DNA and RNA polymerases that lead to the displacement of RNA polymerases and incomplete transcripts.

Predictions for cell-cycle regulation

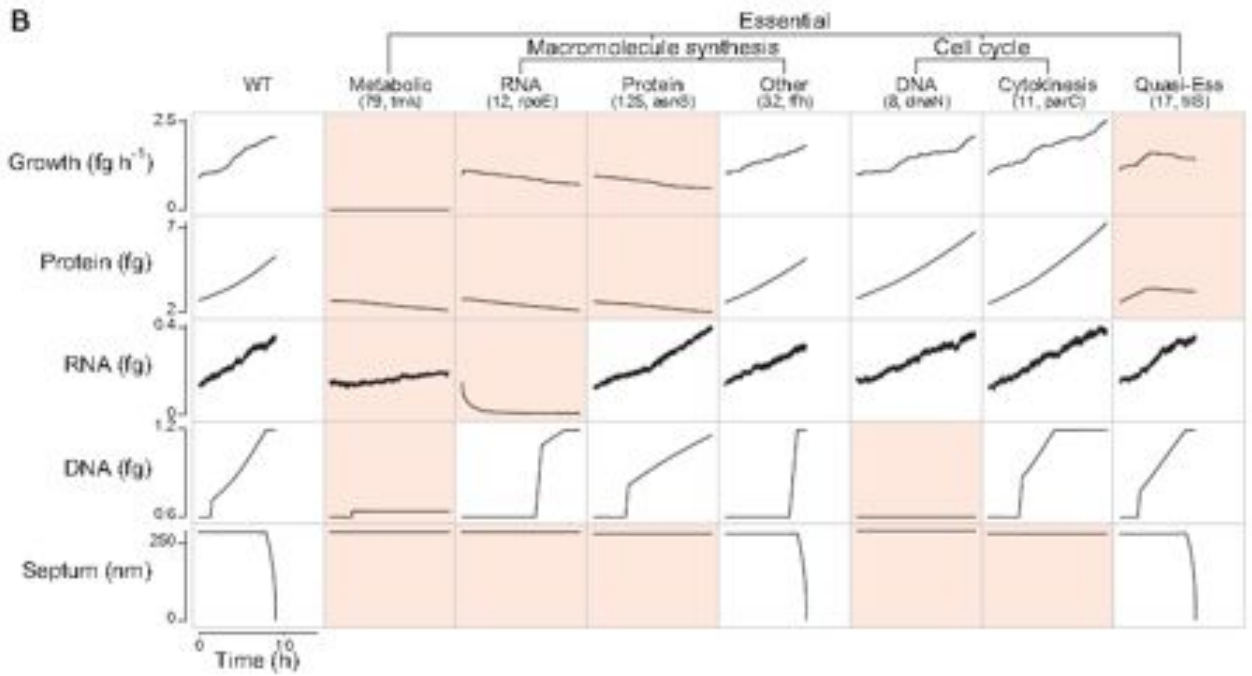


Distributions of the duration of three cell-cycle phases, as well as that of the total cell-cycle length, across 128 simulations.

There was relatively more cell-to-cell variation in the durations of the replication initiation (64.3%) and replication (38.5%) stages than in cytokinesis (4.4%) or the overall cell cycle (9.4%).

This data raised two questions:
(1) what is the source of duration variability in the initiation and replication phases; and
(2) why is the overall cell-cycle duration less varied than either of these phases?

Single-gene knockouts : essential vs. non-essential genes



Single-gene disruption strains grouped into phenotypic classes (columns) according to their capacity to grow, synthesize protein, RNA, and DNA, and divide (indicated by septum length).

Each column depicts the temporal dynamics of one representative in silico cell of each essential disruption strain class.

Dynamics significantly different from wild-type are highlighted in red.

The identity of the representative cell and the number of disruption strains in each category are indicated in parenthesis.

Literature

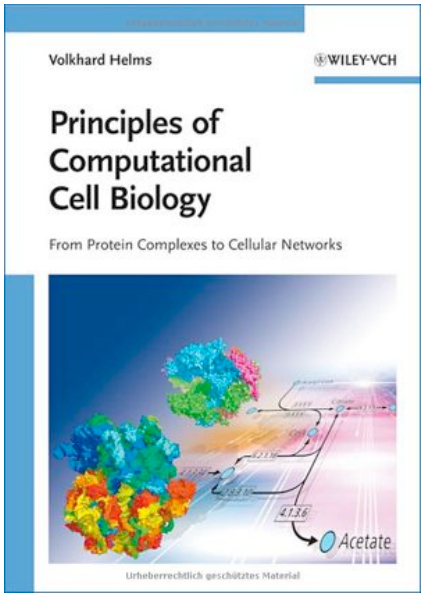
Lecture **slides** — available before the lecture

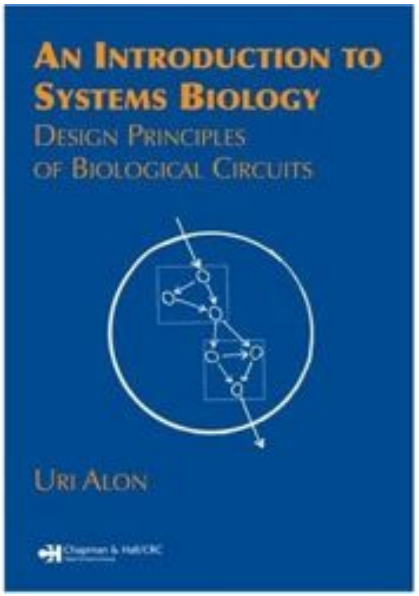
Suggested **reading**

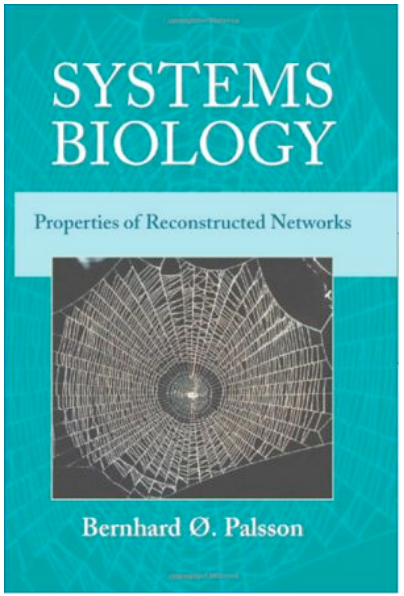
=> check our web page

http://gepard.bioinformatik.uni-saarland.de/teaching/…

Textbooks







=> check computer science library

How to pass this course

Schein = you need to pass 3 out of 4 short tests AND
you need to pass the final exam

Short tests: 4 tests of 45 min each

planned are: first half of lectures V6, V12, V18, V24

⇒ average grade is computed from 3 best tests

If you have passed 2 tests but failed 1-2 tests, you can select one failed test for an oral re-exam.

Final exam: written test of 120 min about assignments

requirements for participation:

- 50% of the points from the assignments
- one assignment task presented @ blackboard
- need to pass 2 short tests

Will take place at the end of the semester

In case you are sick (short test or final exam) you should bring a medical certificate to get a re-exam.

Assignments

Tutors: Christian Spaniol, Ruslan Akulenko,
Mohamed Hamed, Duy Nguyen

Tutorial: ?? Wed, 12:00–14:00, E2 1, room 007

10 assignments with 100 points each

Assignments are part of the course material (not everything is covered in lecture)

=> **one** solution for **two** students (or one)

=> hand-written or one **printable** PDF/PS file per email

=> content: data analysis + interpretation — **think!**

=> no 100% solutions required!!!

=> attach the source code of the programs for checking (no suppl. data)

=> present one task at the **blackboard**

Hand in at the following Fri electronically until 13:00 or
printed at the start of the lecture.

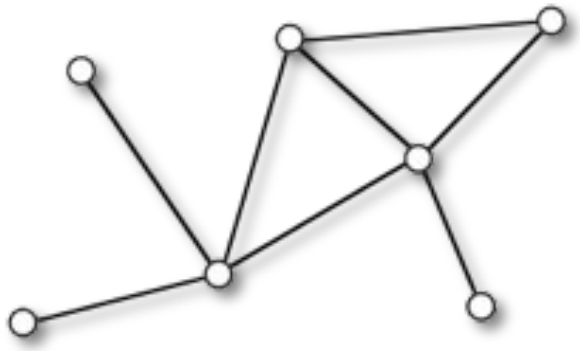
Some Graph Basics

Network <=> Graph

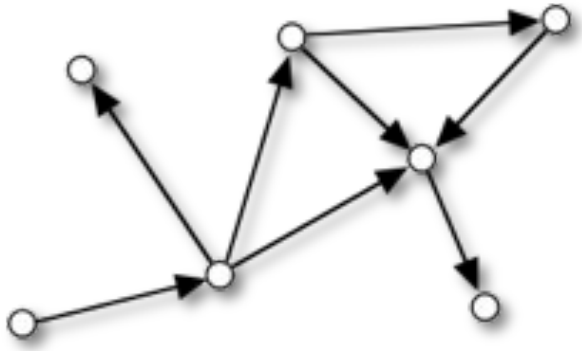
Formal **definition**:

A **graph** *G* is an ordered pair (*V*, *E*) of a set *V* of **vertices** and a set *E* of **edges**.

$$G = (V, E)$$



undirected graph



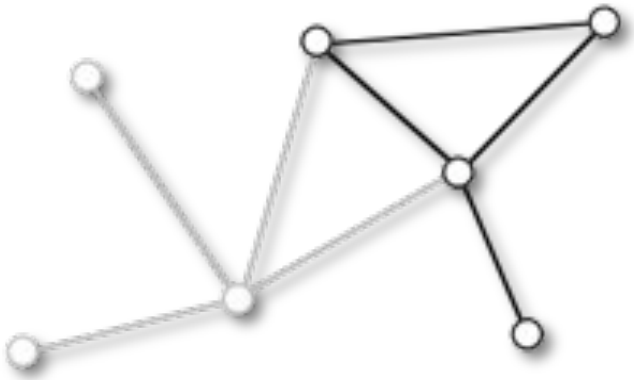
directed graph

If $E = V^{(2)}$ => fully connected graph

Graph Basics II

Subgraph:

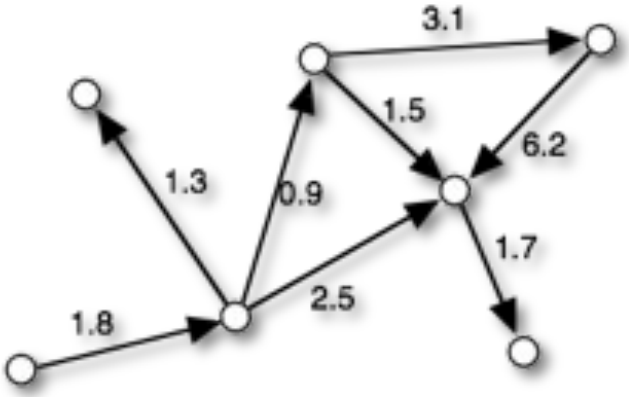
$G' = (V', E')$ is a subset of $G = (V, E)$



Practical question: how to define useful subgraphs?

Weighted graph:

Weights assigned to the edges



Note: no weights for vertices
=> why???

Walk the Graph

Path = sequence of connected vertices

start vertex => internal vertices => end vertex

Two paths are **independent** (internally vertex-disjoint),

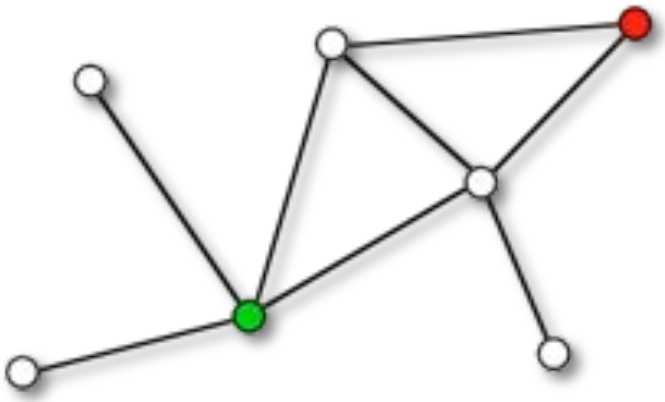
if they have no internal vertices in common.

Vertices *u* and *v* are **connected**, if there exists a path from *u* to *v*.

otherwise: disconnected

Trail = path, in which all edges are distinct

Length of a path = number of vertices || sum of the edge weights



How many paths connect the green to the red vertex?

How long are the shortest paths?

Find the four trails from the green to the red vertex.

How many of them are independent?

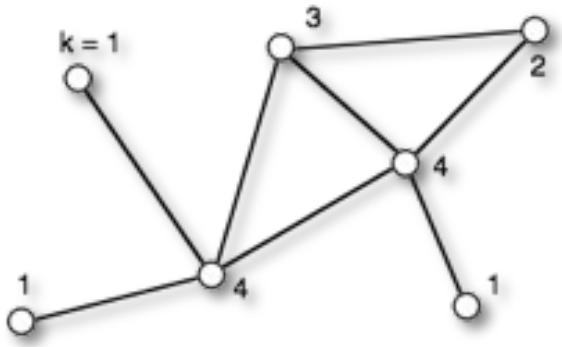
Local Connectivity: Degree

Degree *k* of a vertex = number of edges at this vertex

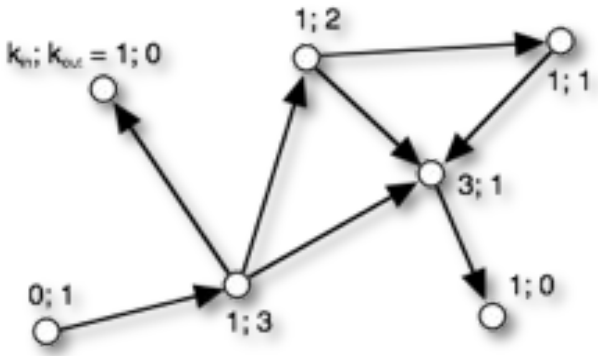
Directed graph => distinguish k_{in} and k_{out}

Degree distribution *P(k)* = fraction of nodes with *k* connections

$$P(k) = \frac{n_k}{N}$$



k	0	1	2	3	4
P(k)	0	3/7	1/7	1/7	2/7



k	0	1	2	3
$P(k_{in})$	1/7	5/7	0	1/7
$P(k_{out})$	2/7	3/7	1/7	1/7

Basic Types: Random Network

Generally: *N* vertices connected by *L* edges

More specific: **distribute** the edges **randomly** between the vertices

Maximal number of links between *N* vertices:

$$L_{max} = \frac{N(N-1)}{2}$$

=> **propability** *p* for an edge between two randomly selected nodes:

$$p = \frac{L}{L_{max}} = \frac{2L}{N(N-1)}$$

=> **average degree** λ

$$\lambda = \frac{2L}{N} = p\,(N-1)$$

path lengths in a random network grow with log(*N*) => small world

Random Network: *P(k)*

Network with *N* vertices, *L* edges
=> Probability for a random link:

$$p = \frac{2L}{N(N-1)}$$

Probability that random node has links to *k* other particular nodes:

$$W_k = p^k (1-p)^{N-k-1}$$

Probability that random node has links to any *k* other nodes:

$$P(k) = \binom{N-1}{k} W_k = \frac{(N-1)!}{(N-k-1)!\,k!} W_k$$

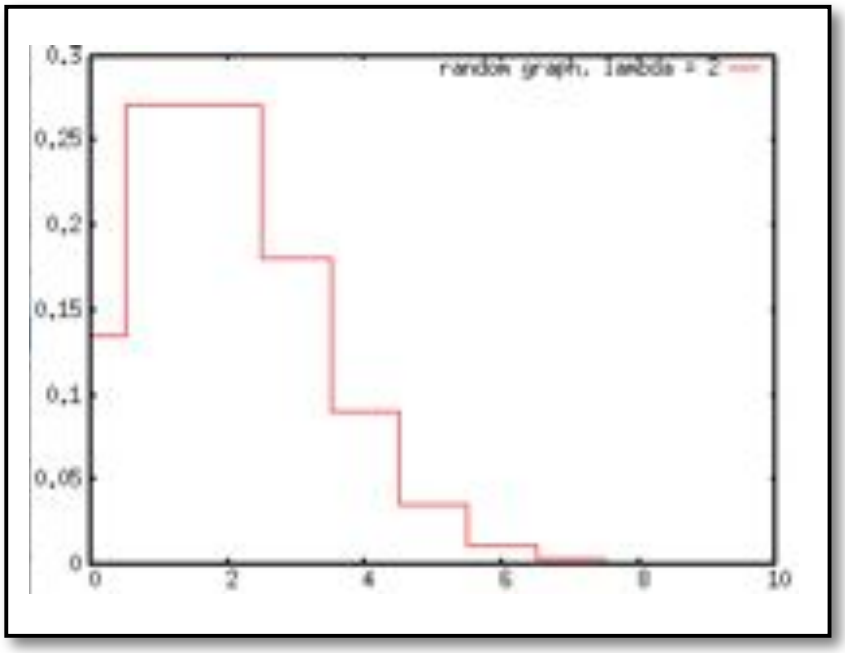
Limit of large graph: $N \to \infty$, $p = \lambda / N$

$$\begin{aligned}
\lim_{N\to\infty} P(k) &= \lim_{N\to\infty} \frac{N!}{(N-k)!\,k!} p^k (1-p)^{N-k} \\
&= \lim_{N\to\infty} \left(\frac{N(N-1)\dots(N-k+1)}{N^k}\right) \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{N}\right)^N \left(1-\frac{\lambda}{N}\right)^{-k} \\
&= \qquad 1 \qquad\qquad \frac{\lambda^k}{k!} \qquad e^{-\lambda} \qquad\quad 1 \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

Random Network: *P(k)*

Many independently placed edges => **Poisson statistics**

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$



k	*P(k | λ = 2)*
0	0.135335283237
1	0.270670566473
2	0.270670566473
3	0.180447044315
4	0.0902235221577
5	0.0360894088631
6	0.0120298029544
7	0.00343708655839
8	0.000859271639598
9	0.000190949253244
10	3.81898506488e-05

=> Small probability for $k >> \lambda$

Basic Types: Scale-Free

Growing network a la Barabasi and Albert (1999):

- start from a small "nucleus"
- add new node with *n* links
- connect new links to existing nodes with probability $\alpha\ k$
  (preferential attachment; β (BA) = 1)

$$p_i = \left(\frac{k_i}{\sum k_i} \right)^{\beta}$$

=> "the rich get richer"

Properties:

- power-law degree distribution:

$$P(k) \propto k^{-\gamma}$$

with γ = 3 for the BA model

- self-similar structure with highly connected hubs (no intrinsic length scale)
  => path lengths grow with log(log(*N*))
  => very small world

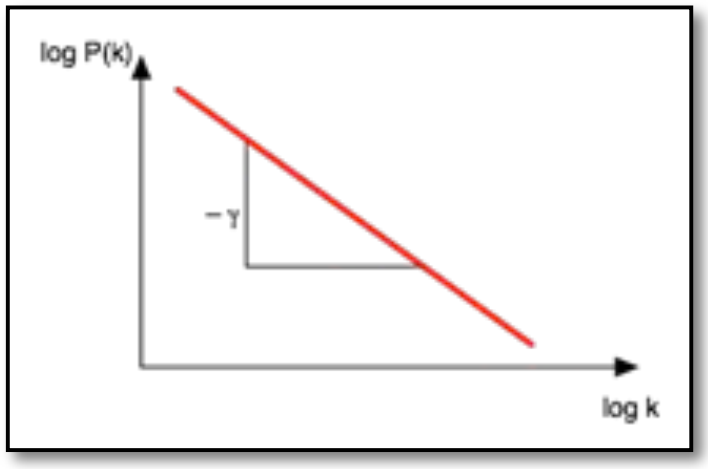
The Power-Law Signature

Power law (Potenzgesetz) $P(k) \propto k^{-\gamma}$

Take log on both sides:

$$\log(P(k)) = -\gamma \log(k)$$

Plot log(*P*) vs. log(*k*) => straight line



Note: for fitting γ against experimental data it is often better to use the integrated *P(k)*
=> integral smoothes the data

$$\int_{k_0}^{k} P(k)dk = \left[-\frac{k^{-(\gamma-1)}}{\gamma}\right]_{k_0}^{k}$$

Scale-Free: Examples

The World-Wide-Web:

=> growth via links to portal sites

Flight connections between airports

=> large international hubs, small local airports

Protein interaction networks

=> some central,

ubiquitous proteins

http://a.parsons.edu/~limam240/blogimages/16_full.jpg

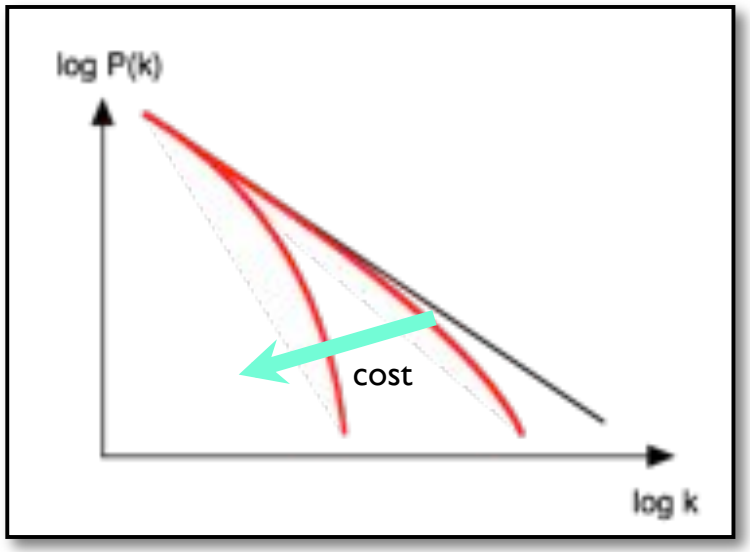
Saturation: Ageing + Costs

Example: network of movie actors (with how many other actors did an actor appear in a joint movie?)

Each actor makes new acquaintances for ~40 years before retirement
=> limits maximum number of links

Example: building up a physical computer network

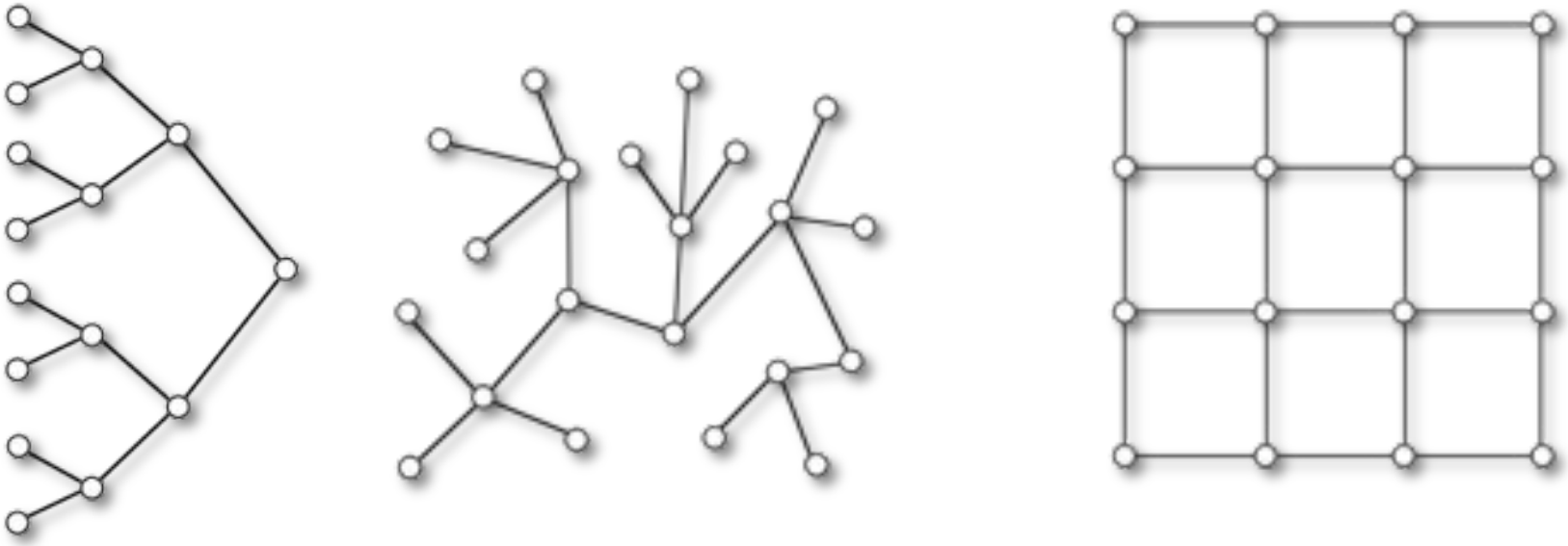
It gets more and more expensive for a network hub to grow further
=> number of links saturates



Hierarchical, Regular, Clustered…

Tree-like network with similar degrees
=> like an organigram
=> hierarchic network

All nodes have the same degree
and the same local neighborhood
=> regular network



P(k) for these example networks? (finite size!)

Note: most real-world networks are somewhere in between the basic types

Summary

What you learned **today**:

=> **networks** are everywhere

=> how to get the "**Schein**" for BI3

=> basic network **types** and **definitions**:
random, scale-free, degree distribution, Poisson distribution, ageing, …

Next lecture:

=> clusters, percolation

=> algorithm on a graph: Dijkstra's shortest path algorithm

=> looking at graphs: graph layout

Further Reading:

R. Albert and A–L Barabási, „Statistical mechanics of complex networks“
Rev. Mod. Phys. **74** (2002) 47-97