

Curvature and Information
in the WWW
and Other Networks

Elisha Moses, Weizmann

PART I

A Short Review

I. General Setup and "Well-Known" Facts

WWW: The set of pages in the Web, and the links connecting them.

Currently about 10^9 pages.

An average of about 5 links/page.

View this as a (directed) graph.

Robots (up to 70% of traffic)

II. Random graphs

[Bollobás: Random Graphs]

Phase transitions: For large graphs, a property is usually true with probability going to 0 or 1 as n (the number of nodes) goes to ∞ .

Example I: Random graphs with n nodes and γn ($\gamma > 1/2$) links have a "giant connected" component of size $> n^{2/3}$ with probability going to 1 when $n \rightarrow \infty$.

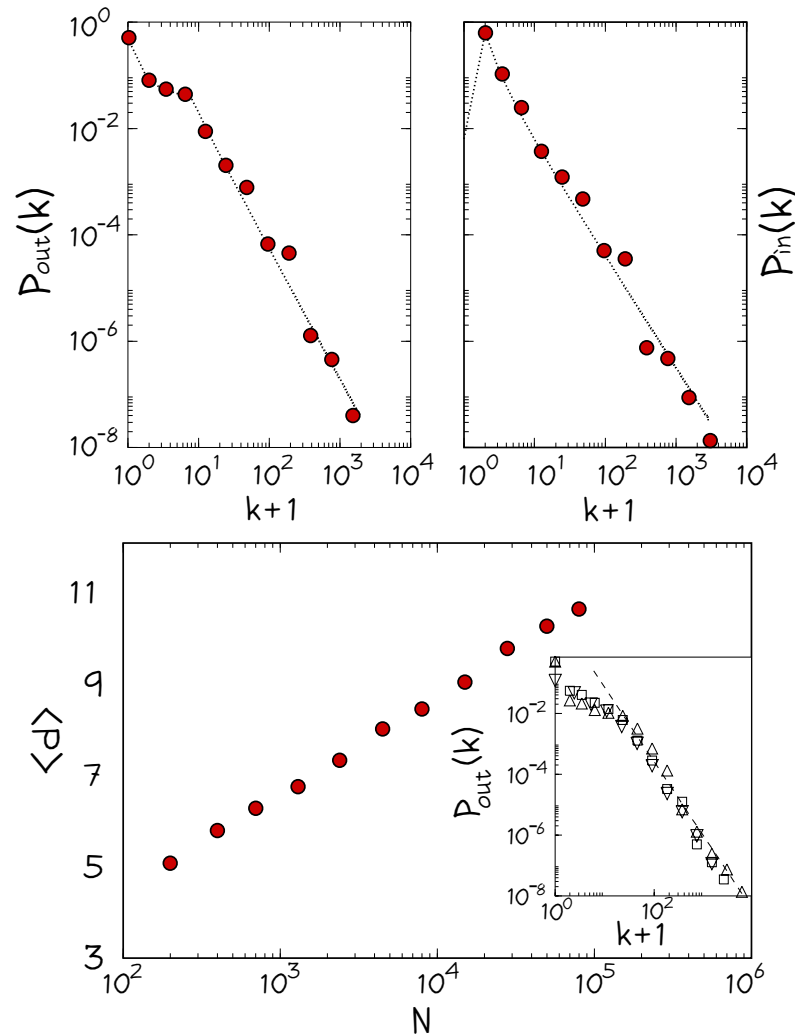
Example II: Random graphs with n nodes and kn links have an average of $O(k/n)$ triangles at each node when $n \rightarrow \infty$.

That is, the probability to find a triangle at a node $\rightarrow 0$

III. Power Laws

Most experimentally studied large graphs have a distribution of valencies which obey a power law:

$$\text{Prob}(v = k) \approx k^{-2.45}$$



a,b: Distribution of out- and in-links. (325,000 nodes, 1,500,000 links)

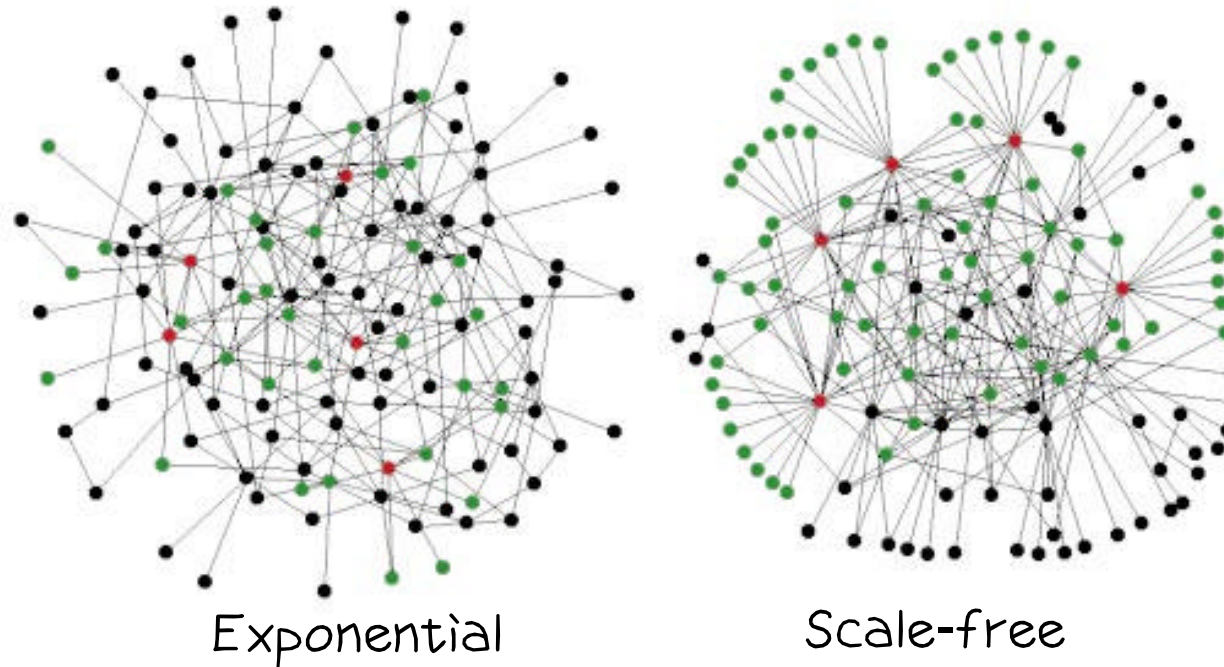
The law is $n^{-2.45}$

c: Predicted mean distance between nodes as function of network size:

$$d \approx 2 \log_{10} n$$

Inset: Data collapse for other networks [Albert, Jeong, Barabási, Nature, 1999]

Properties of Scale-free Graphs



130 nodes, 215 links, **red**: 5 nodes with highest valence

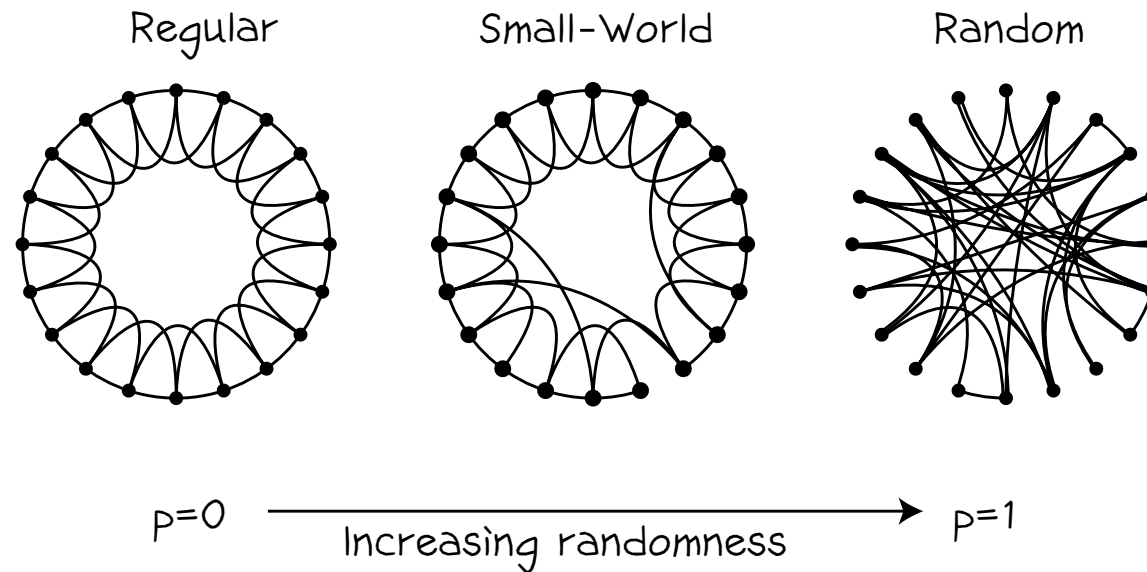
green: their neighbors.

[Albert, Jeong, Barabási, Nature, 1999]

“Authorities” (Central Organizers, Phonebooks)

Vulnerability to elimination of nodes only if authorities are hit

IV. The Small-World Effect



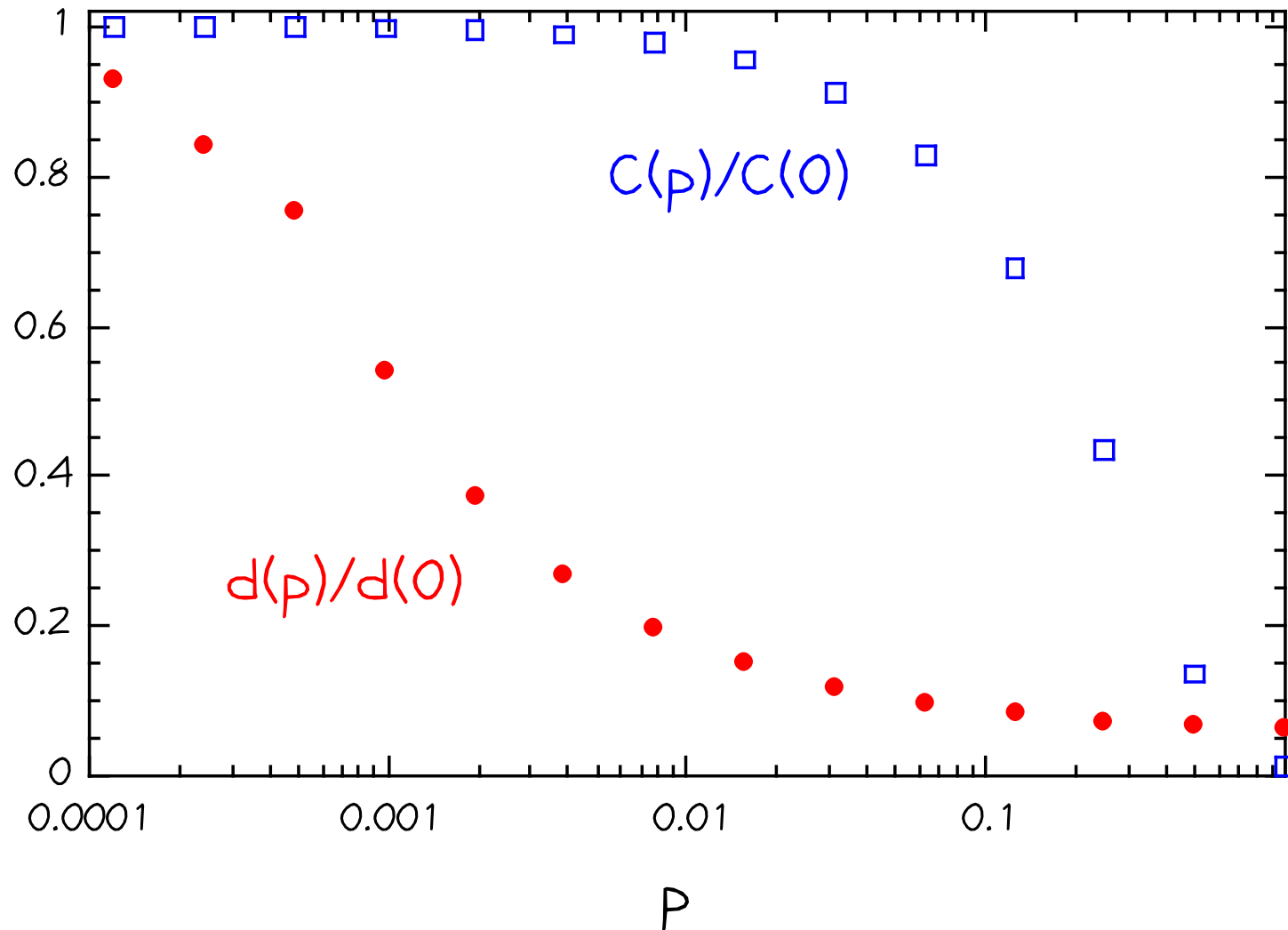
“rewiring” a graph [Watts and Strogatz, Nature 1998]

The diameter of the WWW is only about **19** (can reach any node with 19 clicks)

Well... provided they are connected
and one knows which to click. Perhaps through authorities

Like in chess: The mate is only about 40 moves away.

Remark: Connectivity and diameter are **not** related.



Curvature C and average distance d as a function of randomness.
[Watts and Strogatz, Nature 1998]

Remark: For the moment, think of curvature as “connectivity”

PART II

Curvature and Information

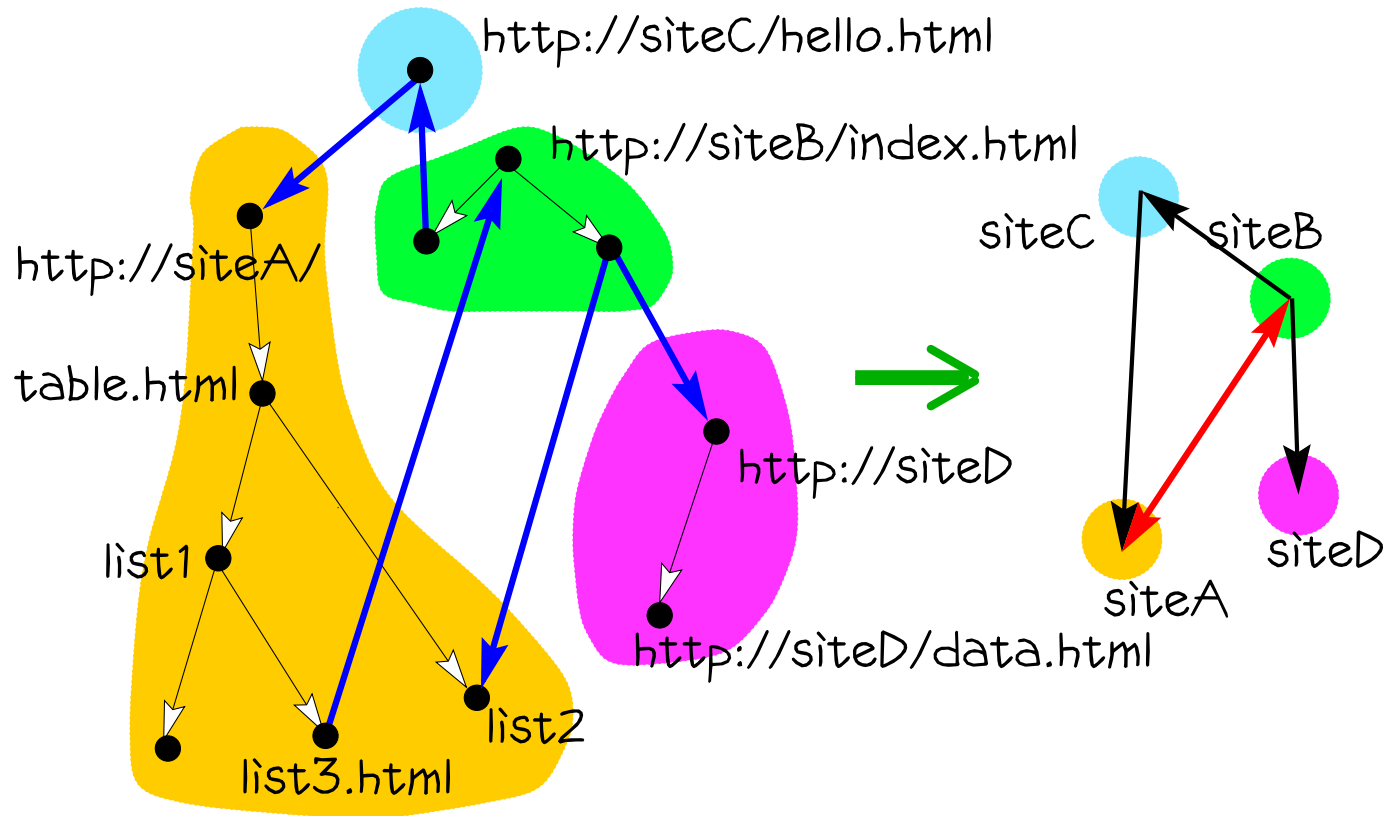
The general question: Is the whole more than the sum of its parts?

This work: Independent contributions add up to a layer of information which **no one in particular** has put there

Avoid subjective criteria: Study Web by looking **only** at links

1. Reduction of Complexity

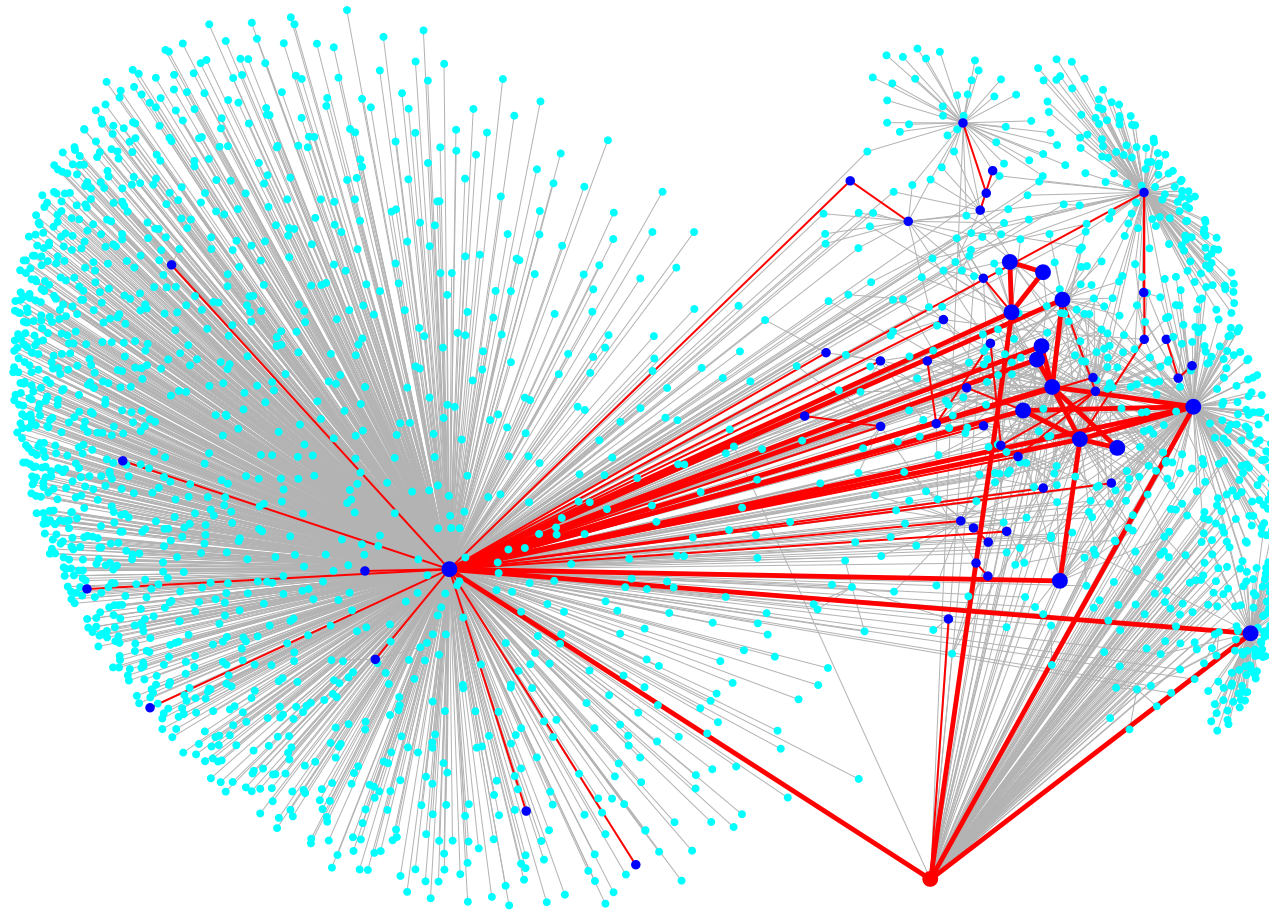
Clustering and Co-links



Note: The **Red** Co-link is a cooperative effect of 2 independent sites (people).

Nodes on the right are called **clusters**

Co-links are **special** because they signal mutual recognition. The clustering procedure defines a new graph of links between clusters, **and** a subgraph of co-links inside the clustered graph.



Piazzolla: The clustered graph of 48,000 nodes, with co-links enhanced.

II. Measuring Affinity

Mutual recognition (friendship) is **transitive**.

If $A \leftrightarrow B$, as well as $B \leftrightarrow C$ are co-linked, we expect $A \leftrightarrow C$ to be co-linked as well

If all co-linked neighbors of a site are mutually linked, they form a strong interest group

Quantitatively: **Triangles** [Watts & Strogatz]

$$\text{curvature}_n = c_n = \frac{t_n}{(v_n - 1)v_n/2}$$

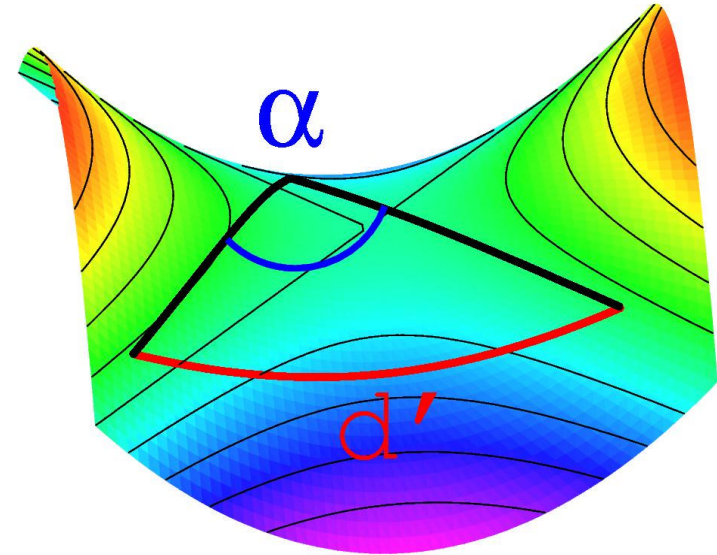
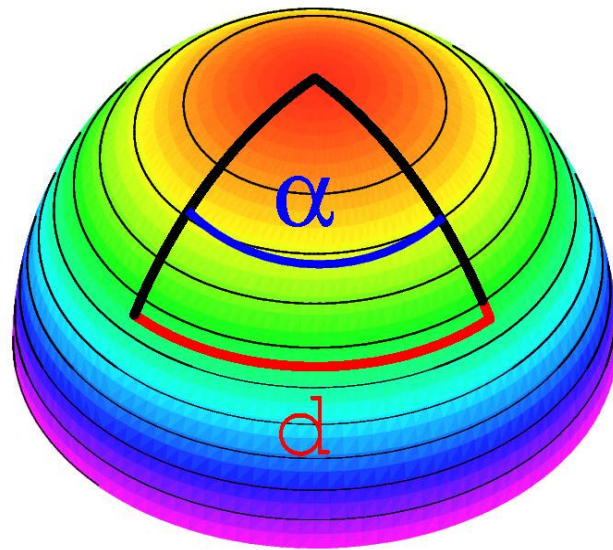
Number of triangles/number of possible triangles at node n

$$\text{curvature}_n = c_n = \frac{t_n}{(v_n - 1)v_n/2}$$

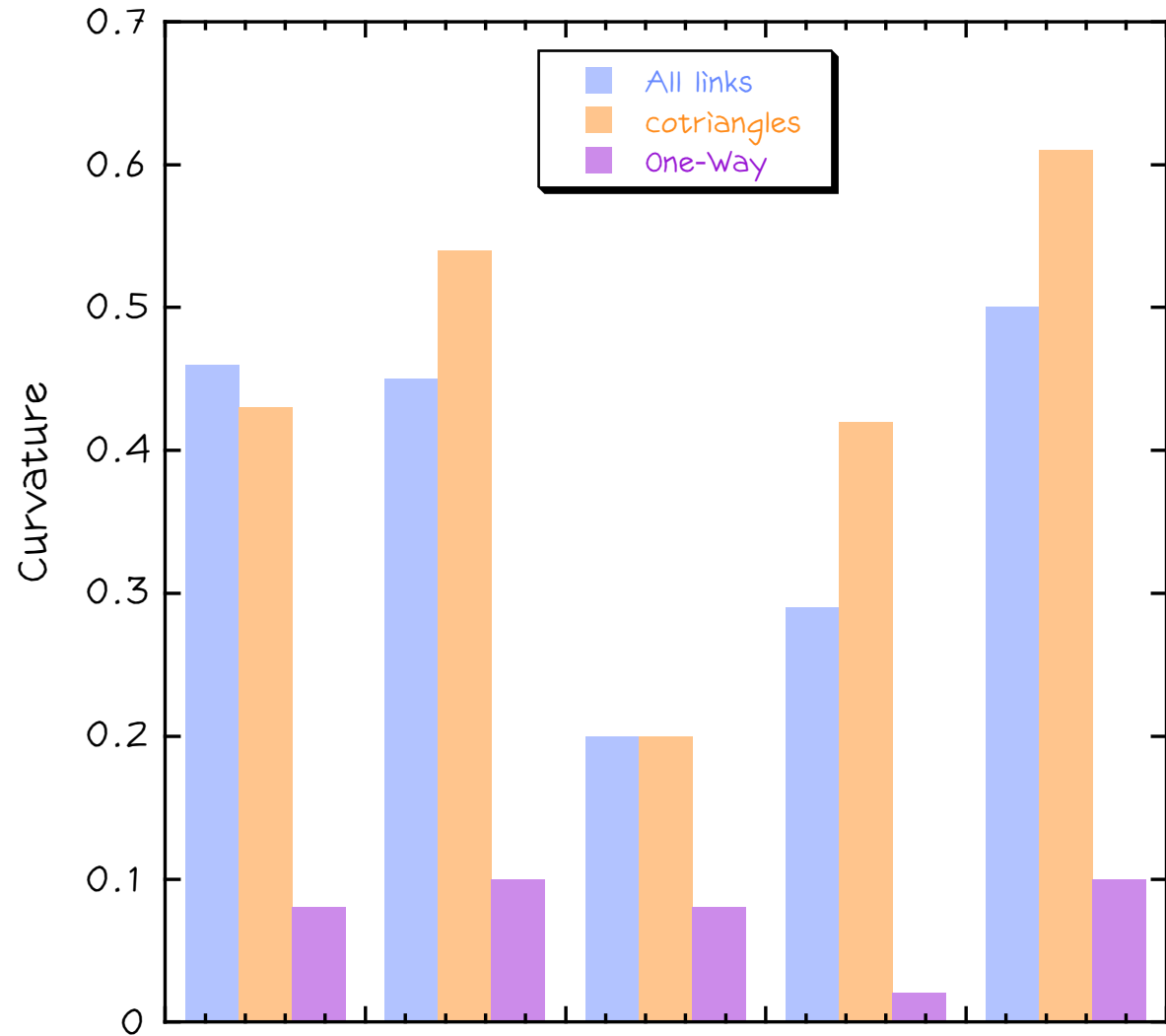
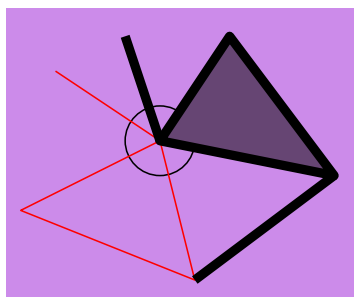
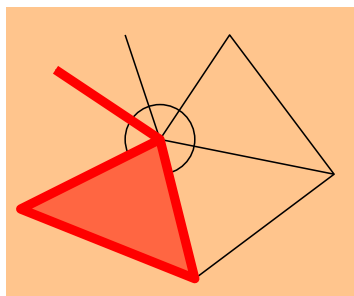
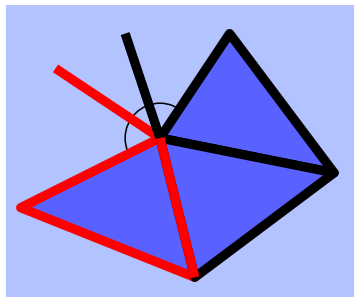
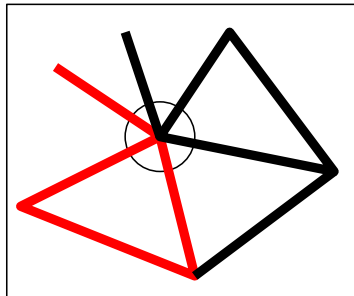
Why a **curvature**? Introduce distance on clustered graph (number of hops). $\langle c_n \rangle = 2$ - mean distance between nearest neighbors of n : $d = 1$ if triangle, $= 2$ if none

So triangles have average sides $1, 1, \langle c_n \rangle$. Law of cosines on surfaces. [Gromov, Bridson-Haefliger].

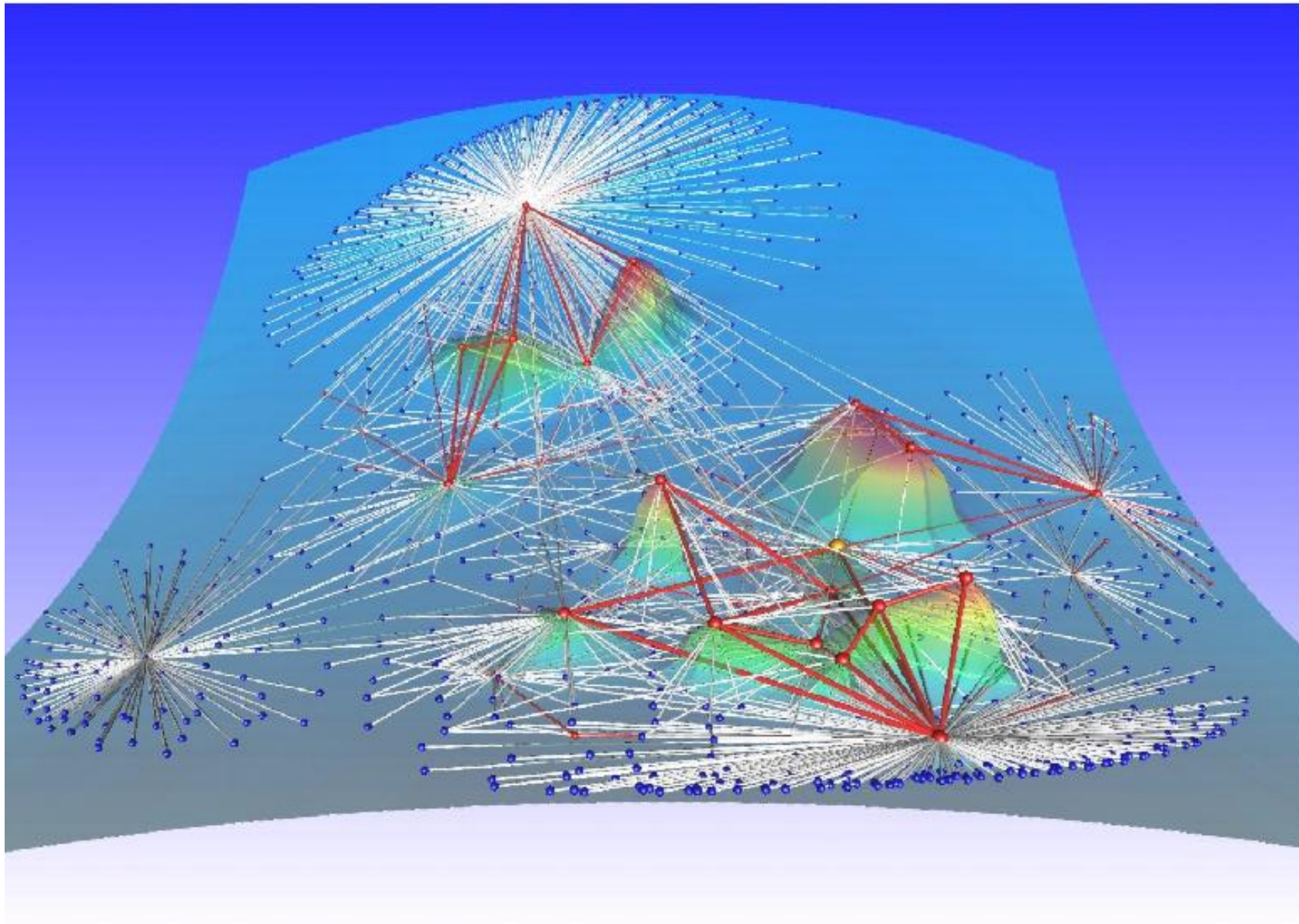
$d < d'$



III. Experiments

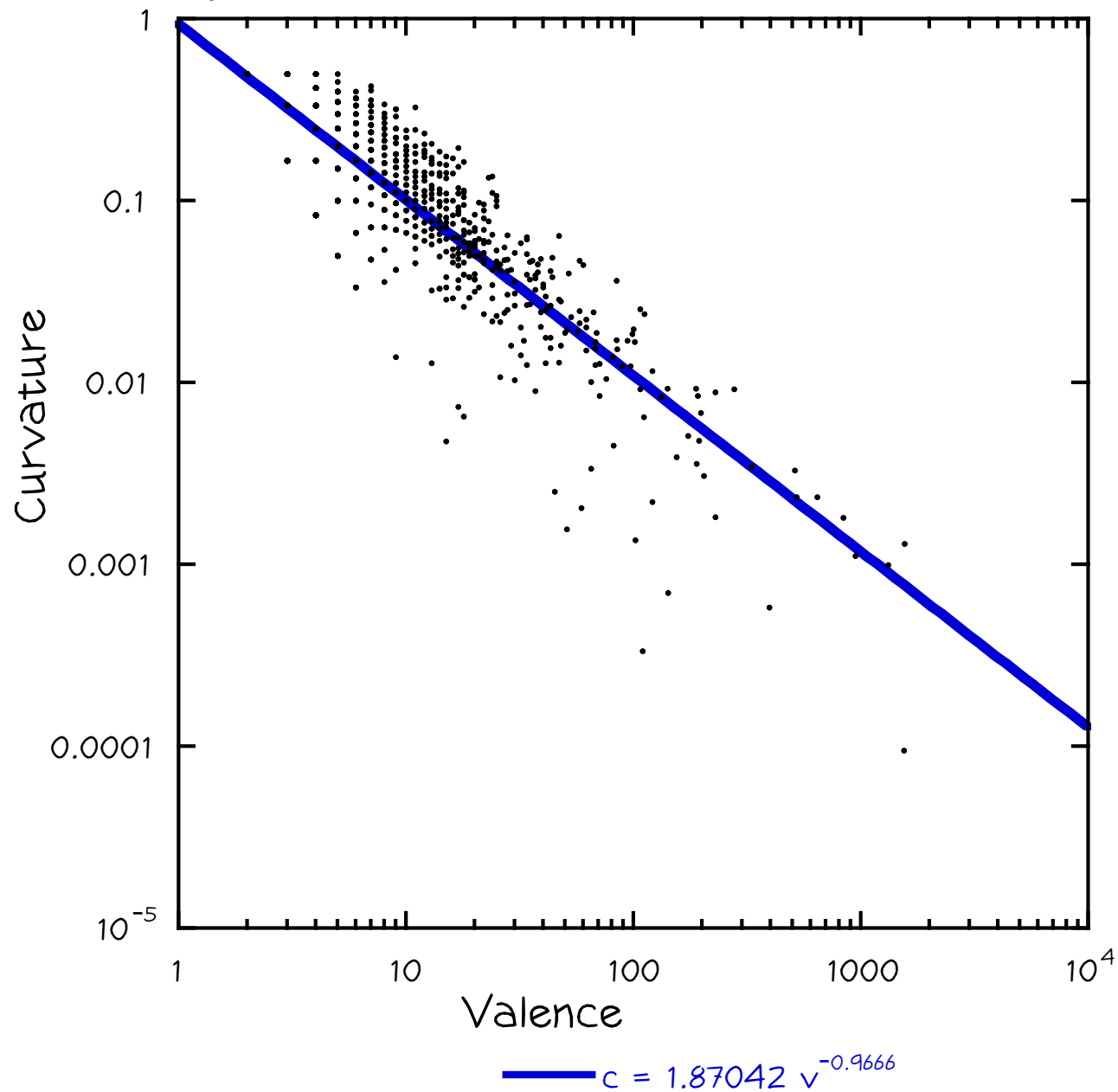


	Shakespeare	Needlework	Revisionist	Piazzolla	Mol. Biol.
URL	277,114	341,398	66,771	47,978	318,705
Clusters	1,560	1,498	947	2,125	1,518
Links	3,730	4,440	1,307	2,577	6,351
Co-Links	321	727	67	70	868



Swiss revisionist: The landscape of curvature

It seems that $c_n \approx 1/v_n$ (on average, but of course not individually)



IV. Generalization and Extension

Consider other graphs and see whether curvature captures "context":

To apply these ideas one needs data, that is, some network (**nodes and links**) and a set of "**subjects**" onto which at least a part of the nodes map.

The question is then whether the landscape of curvature reveals hidden contextual connections.

We have so far studied 4 more cases.

IVa. A Worm Brain

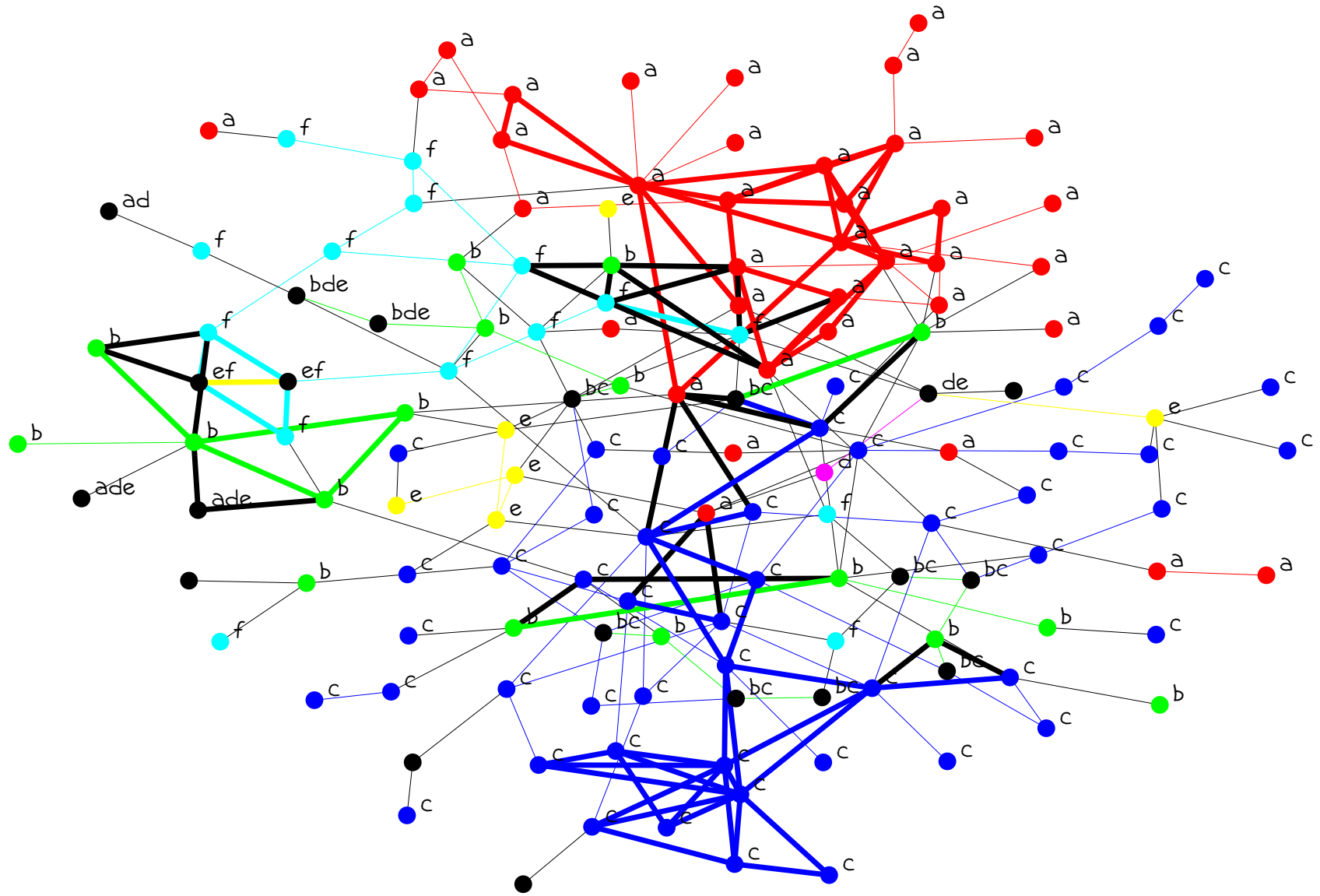
Caenorhabditis elegans (1.5 mm),
959 cells, 302 neurons,
all connections known

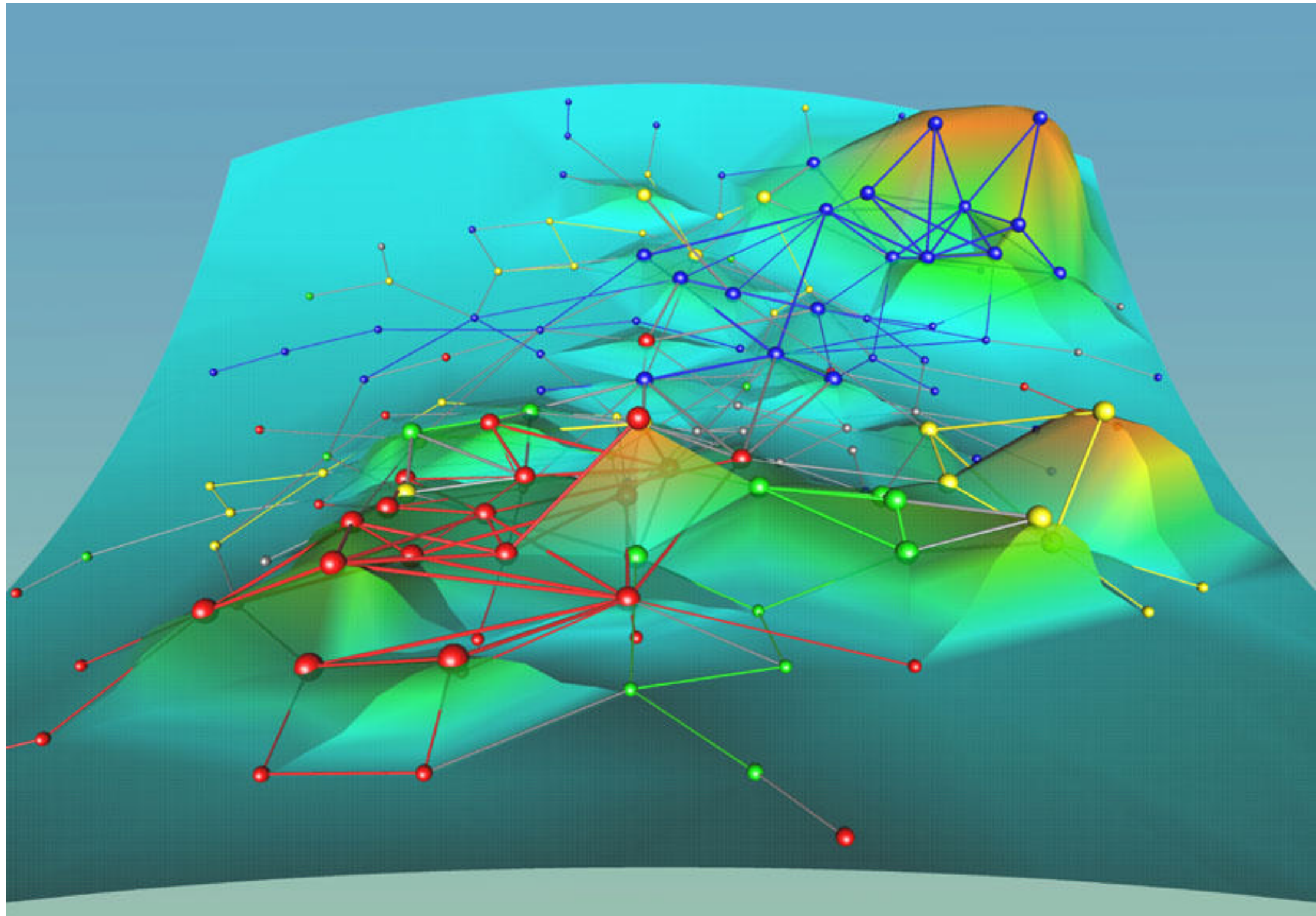
Nodes: The Neurons

Links: The Neural Connections

Subjects: Functionality

a: amphids, b: other head sensors
c: motor neurons, d: ventral motor
e: tail ganglia, f: egg-laying





IVb. Yeast Protein Interactions

Nodes: Yeast proteins

Links: A documented interaction in the DIP database

Subjects: Functionality according to YPD

p2: Pol II transcription

vt: Vesicular transport

ch: Chromatin/chromosome structure

ps: Protein synthesis

cs: Cell stress

cc: Cell cycle control

rp: RNA processing/modification

mr: Mating response

mi: Mitosis

IVc. E. Coli Transcriptional Control

[Uri Alon, Shai Shen-Orr]

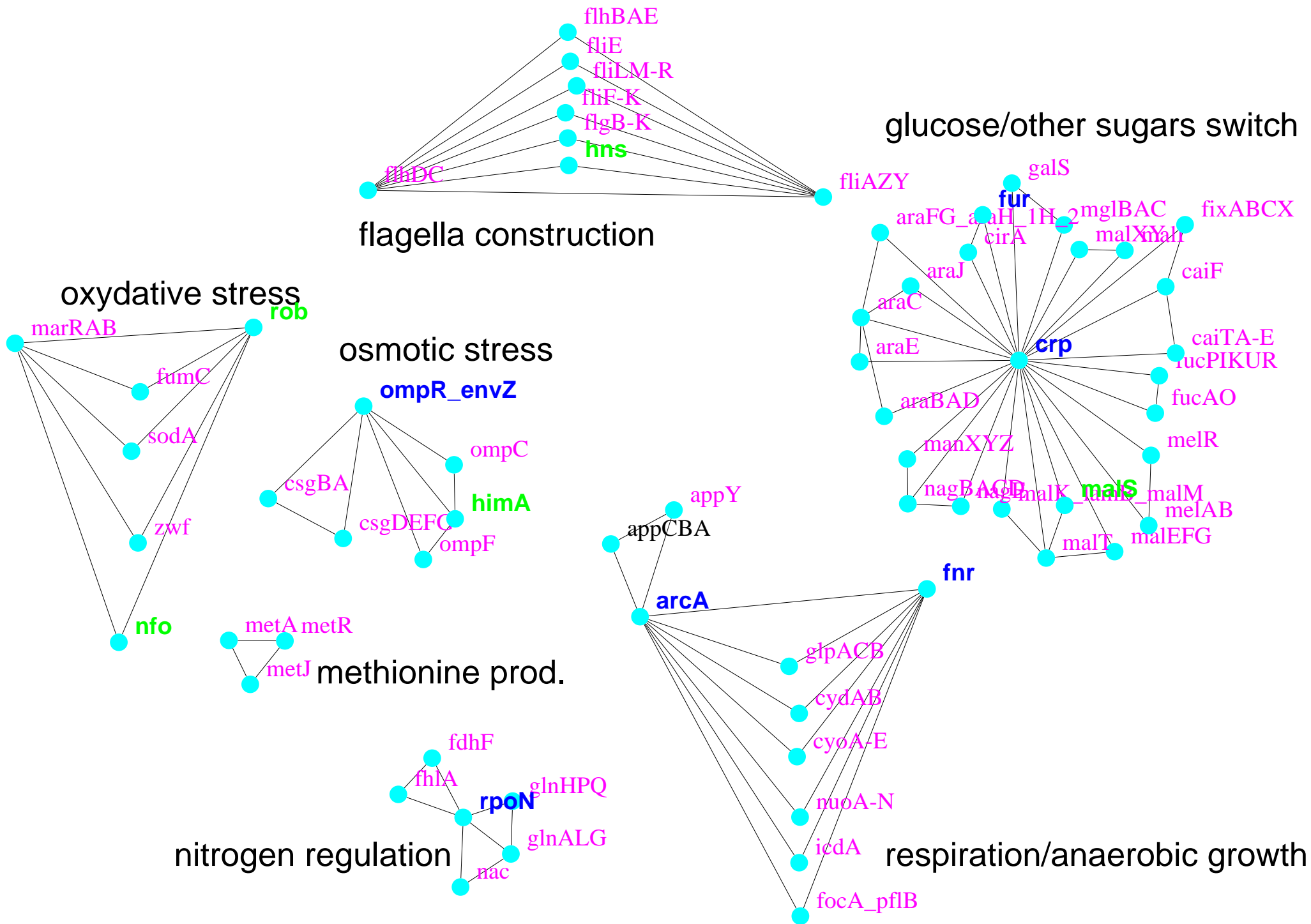
Nodes: genes

Links: regulation of activity

Codes: global regulator (sensor)

global adaptor

others



IVd. Mathematical Physics Papers

(mp_arc)

Nodes: Authors

Links: Two people co-authoring a paper = co-link
reference = directed link

Subjects: Section headings of talks and poster sessions in
math-physics and other conferences

42: Quantum Mechanics

36: Statistical Mechanics

27: Equilibrium Statistical Mechanics

25: Quantum Field Theory

18: Operator Algebras and Noncommutative Geometry

17: PDE

16: Quantum Chaos and the Semiclassical Limit

15: Nonequilibrium Statistical Mechanics

13: Condensed Matter Physics

Outlook

Spectral properties:

[N. Biggs: Algebraic Graph Theory]

Laplacian on Graph without weights:

Determinant counts number of paths.

Multiplicities \Leftrightarrow sites have same neighbors

[F. Chung: Spectral Graph Theory, AMS]

With weights $1/\sqrt{v_n}$:

Closer to Riemannian Geometry

(**global** properties)

Entropy:

Have a finite set of subjects S . Each node maps to an element (or subset) of S . Consider connected components of the graph defined by points with curvature $\geq c_*$.

The map induces a partition of the connected component with probability p_i to find subject i .

$$\text{Entropy} = - \sum_{i \in S} p_i \log p_i$$

Information Theory: [Eckmann & Collet]

Triangles provide information

Having a density of α triangles and k links per node in random graph with n nodes reduces the number of possible graphs from

$$\exp(kn \log n) \Rightarrow \exp\left(\left(k - \frac{\alpha}{11k^2}\right)n \log n\right)$$

Complexity decreases to $(1 - \alpha/11k^3)/\text{link}$.

Information incr. to $(1 + O(\alpha)) \log n/\text{node}$.

Describing a link among n nodes costs $O(\log n)$ bytes \Rightarrow a byte of link-address gives advantage of $1 + O(\alpha)$ over a byte of link in a random graph

Speculations: Is the Web alive?

Ill defined question at this stage, no precise definition of living object:

Criteria:

? **Reproduction**

- **growth** (inert material is organized and patterned)
- **defenses and immune system**
- **perpetuation** Web is here to stay ("genetic information" collected and perpetuated by crawlers)

Is the Web alive? Question does not make much sense, but we treat it as though it were.

A new form of organism