Fast Learning in Multi-Layer Feed-Forward Neural Nets with Small-World Architecture

Contents

- Part I: What is a Small-World network? Some examples.
- Part II: Small-World networks in neuroscience. Experiments and models.
- Part III: Learning in SWN.

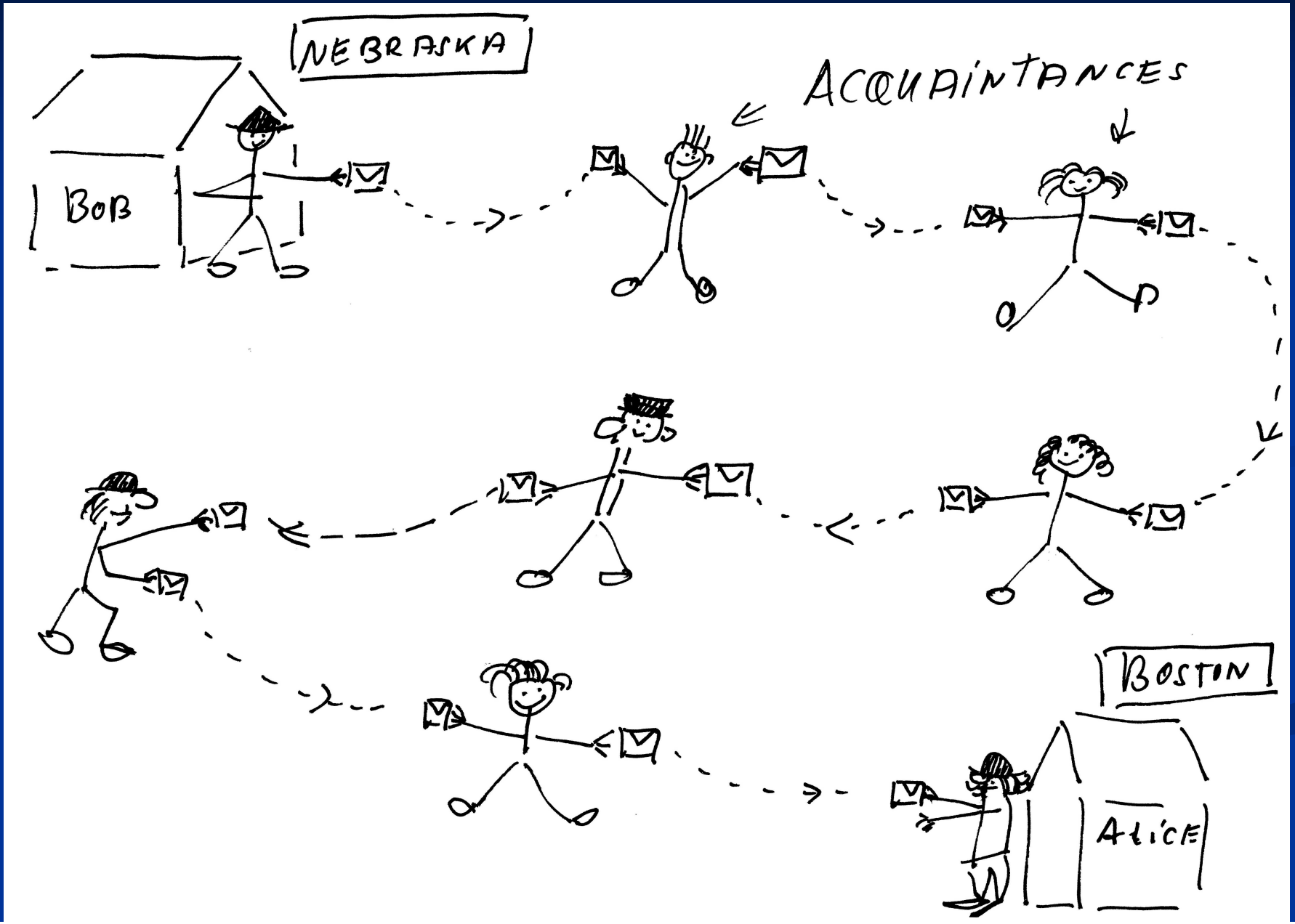
The first papers on SWN's

- D.J. Watts, S.H. Strogatz, Nature 393(1998) 442.
- R. Albert, H. Jeong, A.L. Barbasi, Nature 401 (1999) 130.
- B.A. Huberman, L.A. Adamic, Nature 401 (1999) 131.
- J.M. Kleinberg, Nature 406 (2000) 845.
- S.H. Strogatz, Nature 410 (2001) 268.
- D.J. Watts, P.S. Dodds, M.E.J. Newman, Science 296 (2002) 1302.

At the beginning: Milgram's Letter Experiment. Result: Six Degrees of Separation.

S. Milgram, "The Small-World Problem", Psychology Today 1, 60-67 (1967).

Short Average Path Length: L=7



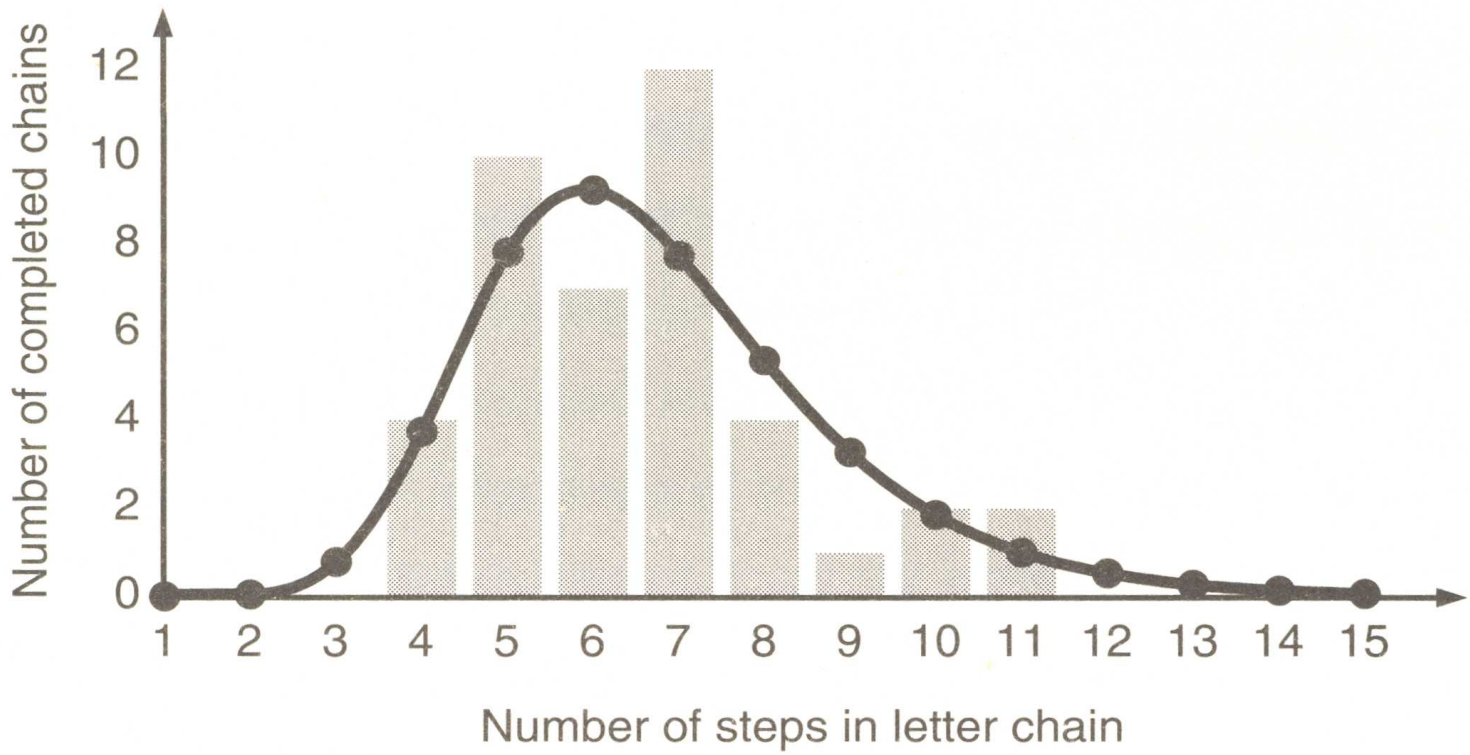
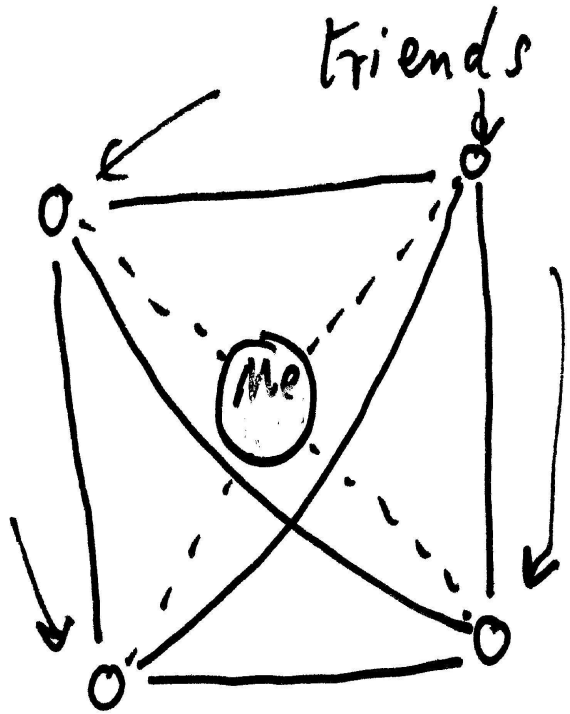
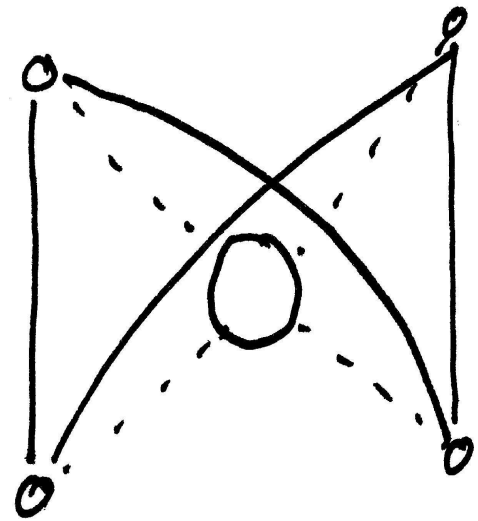


Figure 5.9. The results of the social network search model compared with Milgram's Nebraska results. The bars represent the forty-two completed chains that started in Nebraska, and the curve is the average over many simulated searches preformed according to our model.

High Local Clustering Coefficient C



$$C = \frac{6}{6} = 1$$



$$C = \frac{4}{6} = \frac{2}{3}$$

Clustering in social networks

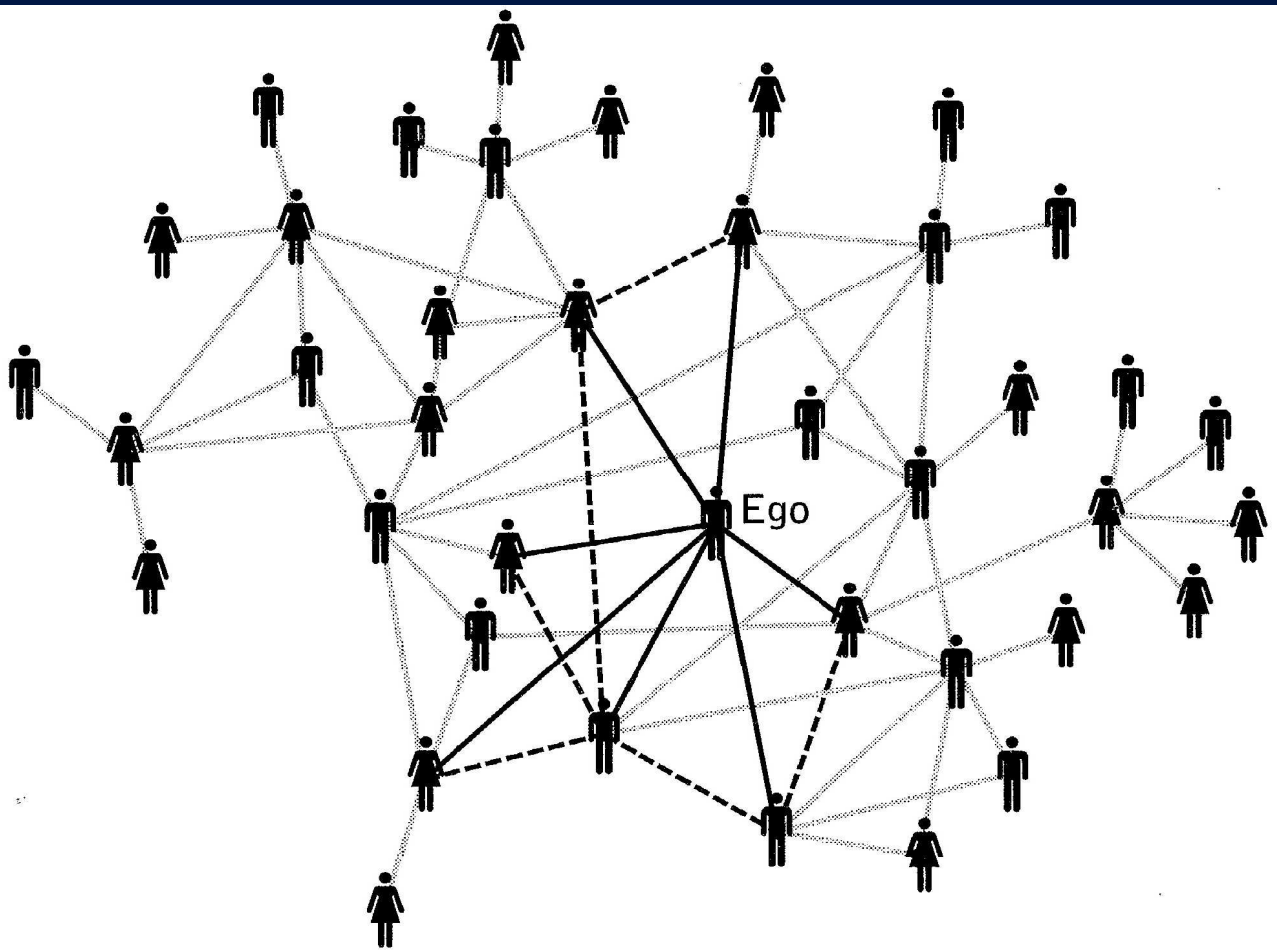


Figure 1.3. Real social networks exhibit clustering, the tendency of two individuals who share a mutual friend to be friends themselves. Here, Ego has six friends, each of whom is friends with at least one other.

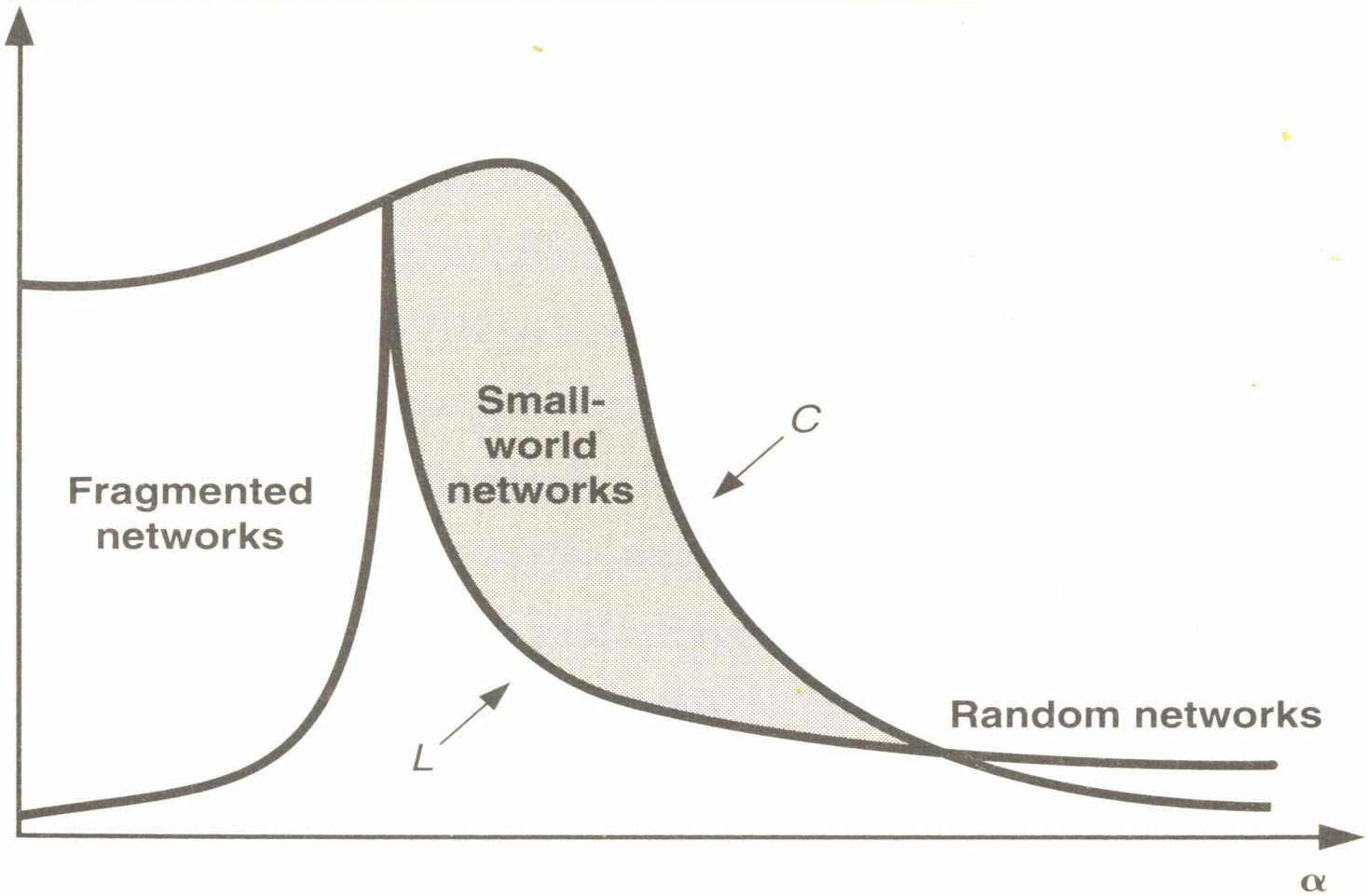
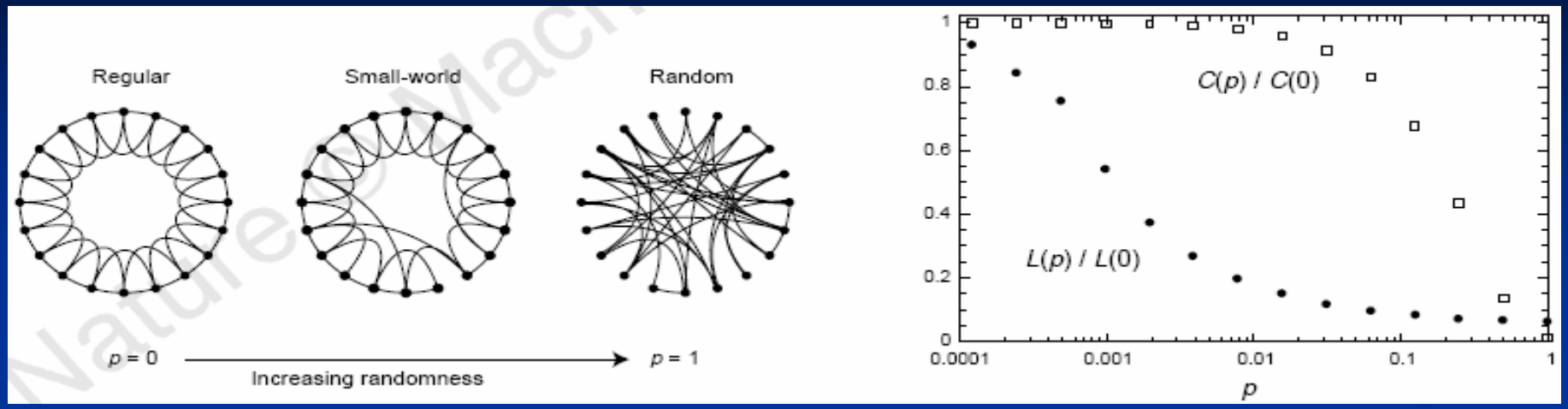


Figure 3.4. Comparison between path length (L) and clustering coefficient (C). The region between the curves, where L is small and C is large (shaded), represents the presence of small-world networks.

What is a Small-World Network?



Watts, D. J. and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440-42

A "Small-World" network is a network architecture between a random network and a regular network. Starting from a regular network and at random rewiring some links to a far node yields SWN (Small-World network can be obtained also by adding links instead of rewiring old links)

There are other possible architectures, e.g. Scale-Free, Clustered, Modular

Alternative architecture: Branching network

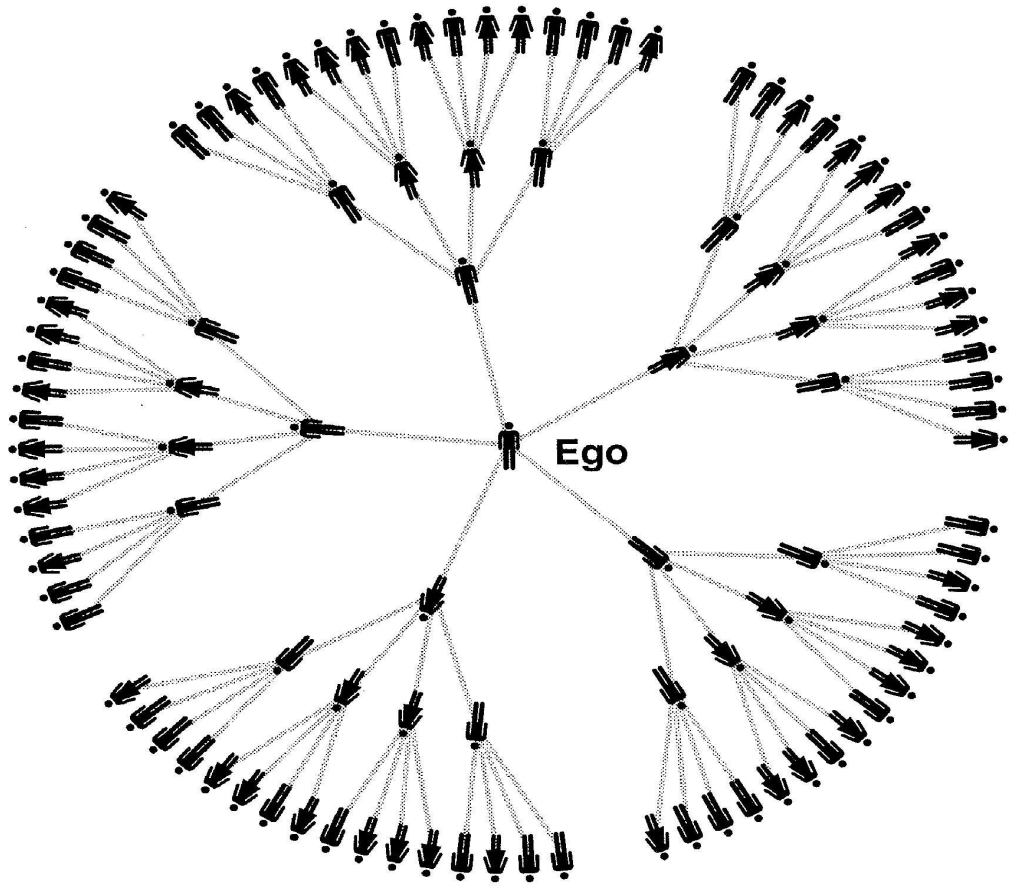


Figure 1.2. A pure branching network. Ego knows only 5 people, but within two degrees of separation, ego can reach 25; within three degrees, 105; and so on.

Possible architectures for network of internet (1964)

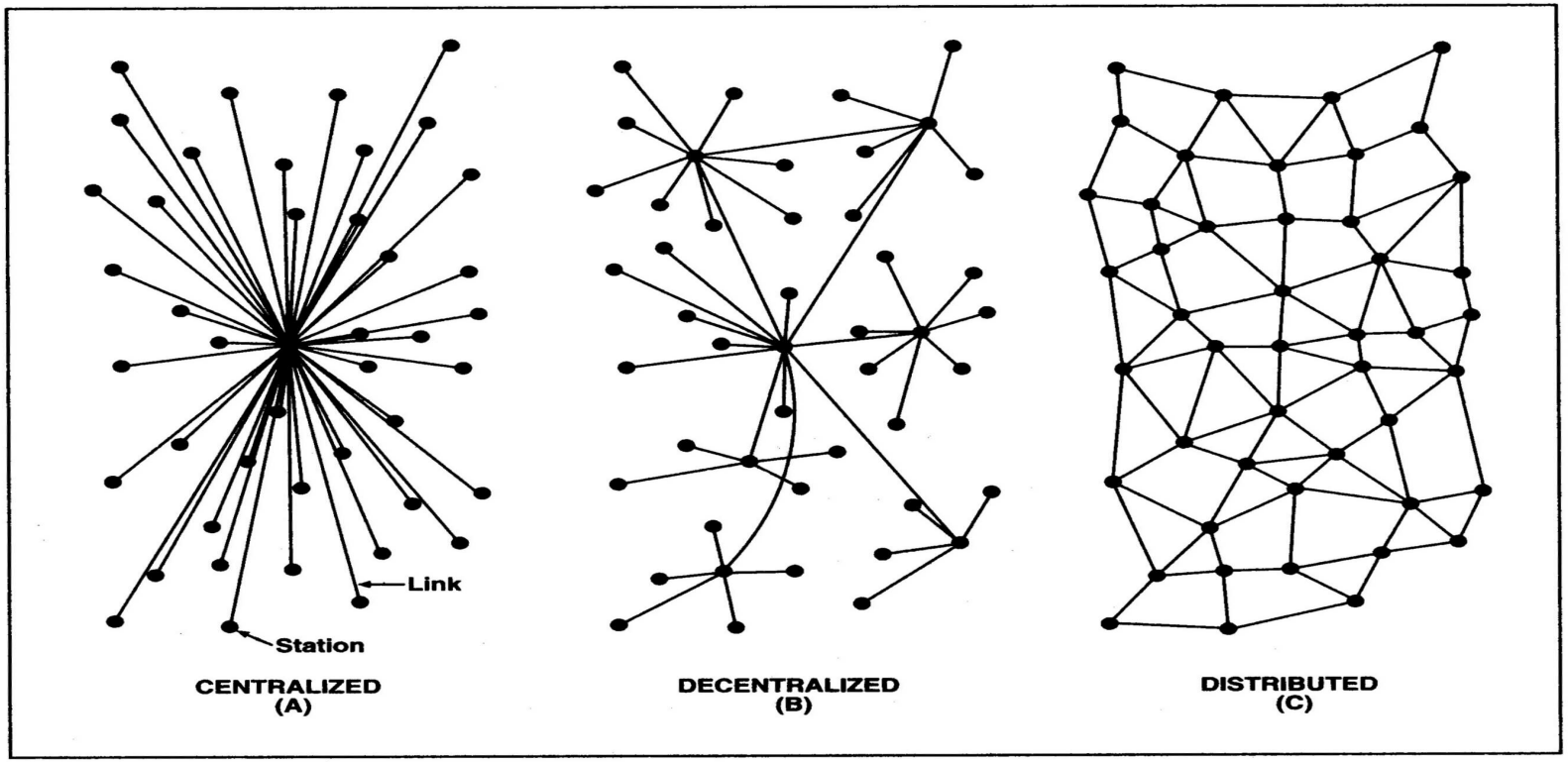


Figure 11.1 Paul Baran's Networks. *In 1964, Paul Baran began thinking about the optimal structure of the Internet. He suggested that there were three possible architectures for such a network—centralized, decentralized, and distributed—and warned that both the centralized and decentralized structures that dominated communications systems of the time were too vulnerable to attack. Instead, he proposed that the Internet should be designed to have a distributed, mesh-like architecture. (Reproduced with permission of Paul Baran.)*

- Why is high local clustering beneficial?

  Redundancy in case of breakdown of node

  Small risk of error. Stability of system.

  Why is short path length beneficial?

  Rapid communication (Society: Milgram exp.,

  WWW: fast search.

  Brain: fast response, coherent activation in motor cortex.)

- Ideal network:

  Fully connected – each node is linked to each other node.

  Then:

  C=1(maximal), L=1(minimal)

  Best possible network ?

  Not possible in brain! Brain connections are diluted.

Examples of small-world networks…

(1) Human society:

Milgram's letter experiment.

Organization of board members of big corporations.

Films actors network.

Paul Erdös' (mathematician) publication network.

(2) Engineering:

Grid of electrical power lines in western US.

(3) Communication:

WWW,

Internet.

(4) Biology:

Metabolic network of bacterium E.coli,

Neural network of nematode worm C. elegans.

Statistics of Small-World networks

Network	N	C	L	$\langle k \rangle$	λ	C_{rand}	L_{rand}
C. Elegans	282	0.28	2.65	7.68	NA	0.025	2.1
Macaque VC	32	0.55	1.77	9.85	NA	0.318	1.5
Cat Cortex	65	0.54	1.87	17.48	NA	0.273	1.4

Table II from:

Scale-free brain functional networks

Victor M. Eguiluz,Dante R. Chialvo,Guillermo Cecchi,Marwan Baliki,and A. Vania Apkarian

Network	Size	Clustering coefficient	Average path length	Degree exponent
Internet, domain level [13]	32711	0.24	3.56	2.1
Internet, router level [13]	228298	0.03	9.51	2.1
WWW [14]	153127	0.11	3.1	$\gamma_{\text{in}} = 2.1 \quad \gamma_{\text{out}} = 2.45$
E-mail [15]	56969	0.03	4.95	1.81
Software [16]	1376	0.06	6.39	2.5
Electronic circuits [17]	329	0.34	3.17	2.5
Language [18]	460902	0.437	2.67	2.7
Movie actors [5, 7]	225226	0.79	3.65	2.3
Math. co-authorship [19]	70975	0.59	9.50	2.5
Food web [20, 21]	154	0.15	3.40	1.13
Metabolic system [22]	778	–	3.2	$\gamma_{\text{in}} = \gamma_{\text{out}} = 2.2$

Comparison of SWN model with experiment:

TABLE 3.2 STATISTICS OF SMALL WORLD NETWORKS

	L_{ACTUAL}	L_{RANDOM}	C_{ACTUAL}	C_{RANDOM}
MOVIE ACTORS	3.65	2.99	0.79	0.00027
POWER GRID	18.7	12.4	0.080	0.005
C. ELEGANS	2.65	2.25	0.28	0.05

L=Path Length; C=Clustering Coefficient.

Example of Social Network:

- Film actors network. Kevin Bacon game:
- K. Bacon: Bacon number = 0.
- Movie star having acted in a film together with K.B.: Bacon number = 1.
- Move star who has acted together with someone who has acted with K.B.: Bacon number = 2.
- Etc.

Statistics of Kevin Bacon Network:

TABLE 3.1 DISTRUBUTION OF ACTORS ACCORDING TO BACON NUMBER

BACON NUMBER	NUMBER OF ACTORS	CUMULATIVE TOTAL NUMBER OF ACTORS
0	1	1
1	1,550	1,551
2	121,661	123,212
3	310,365	433,577
4	71,516	504,733
5	5,314	510,047
6	652	510,699
7	90	510,789
8	38	510,827
9	1	510,828
10	1	510,829

Example of Network in Biology: Food Web of Fish in Ocean

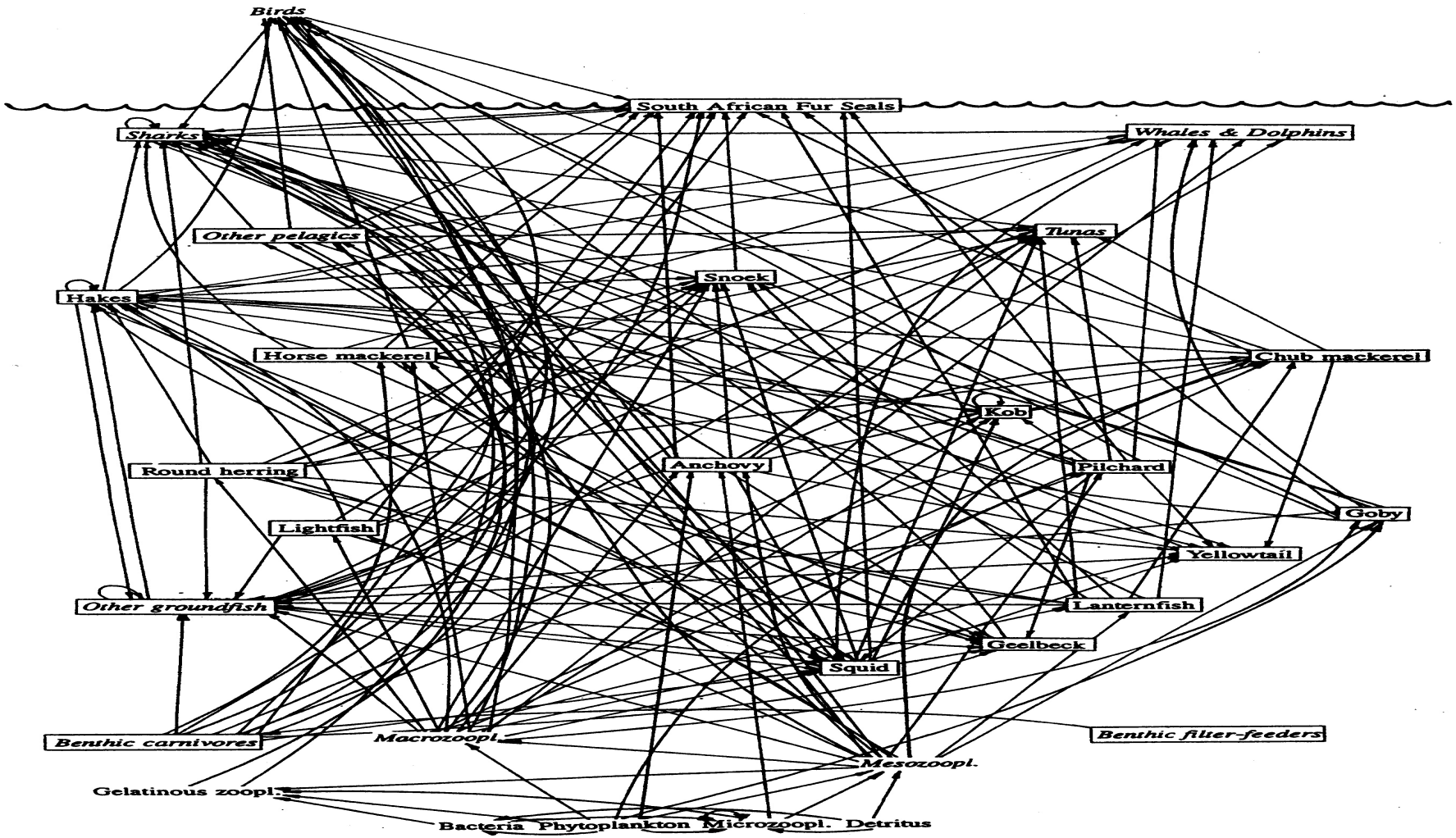


Figure 1. A portion of the food web for the Benguela ecosystem, which is located off the western coast of South Africa. (Reprinted by permission of Peter Yodzis.)

Network of Interactions Between Proteins in Yeast

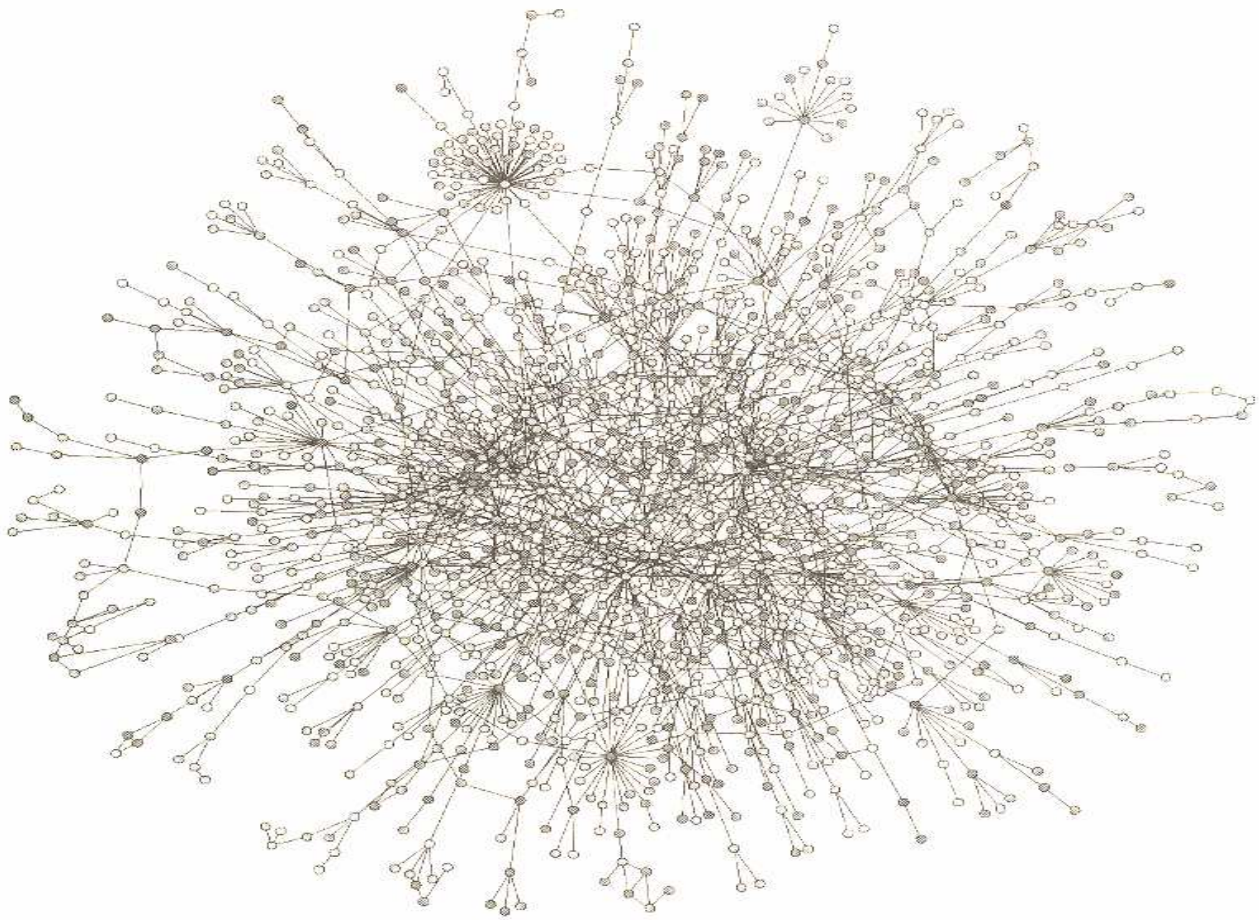


Figure 17. A diagram showing the network of interactions between the various proteins in the yeast *Saccharomyces cerevisiae*, more commonly known as brewer's or baker's yeast. (Image courtesy of Hawoong Jeong, reprinted by permission.)

Random vs. Scale-Free Networks

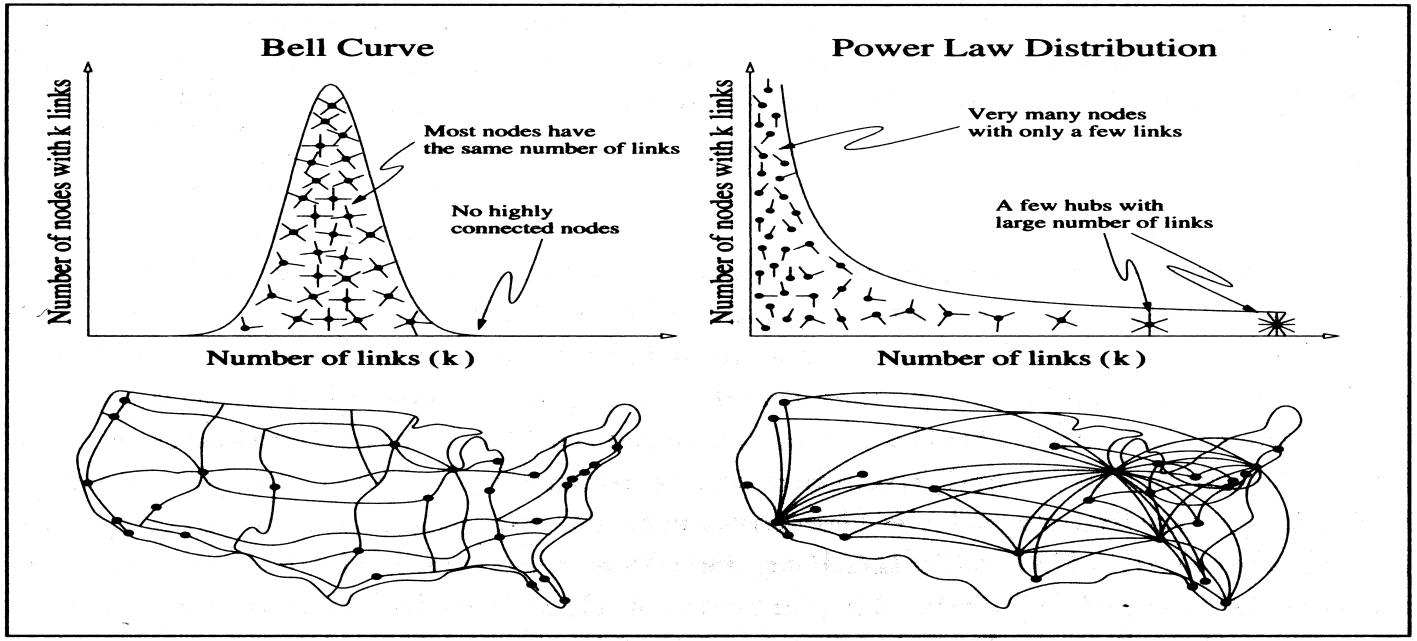


Figure 6.1 Random and Scale-Free Networks. *The degree distribution of a random network follows a bell curve, telling us that most nodes have the same number of links, and nodes with a very large number of links don't exist (top left). Thus a random network is similar to a national highway network, in which the nodes are the cities, and the links are the major highways connecting them. Indeed, most cities are served by roughly the same number of highways (bottom left). In contrast, the power law degree distribution of a scale-free network predicts that most nodes have only a few links, held together by a few highly connected hubs (top right). Visually this is very similar to the air traffic system, in which a large number of small airports are connected to each other via a few major hubs (bottom right).*

Example of Scale-Free Network: Network of Internet

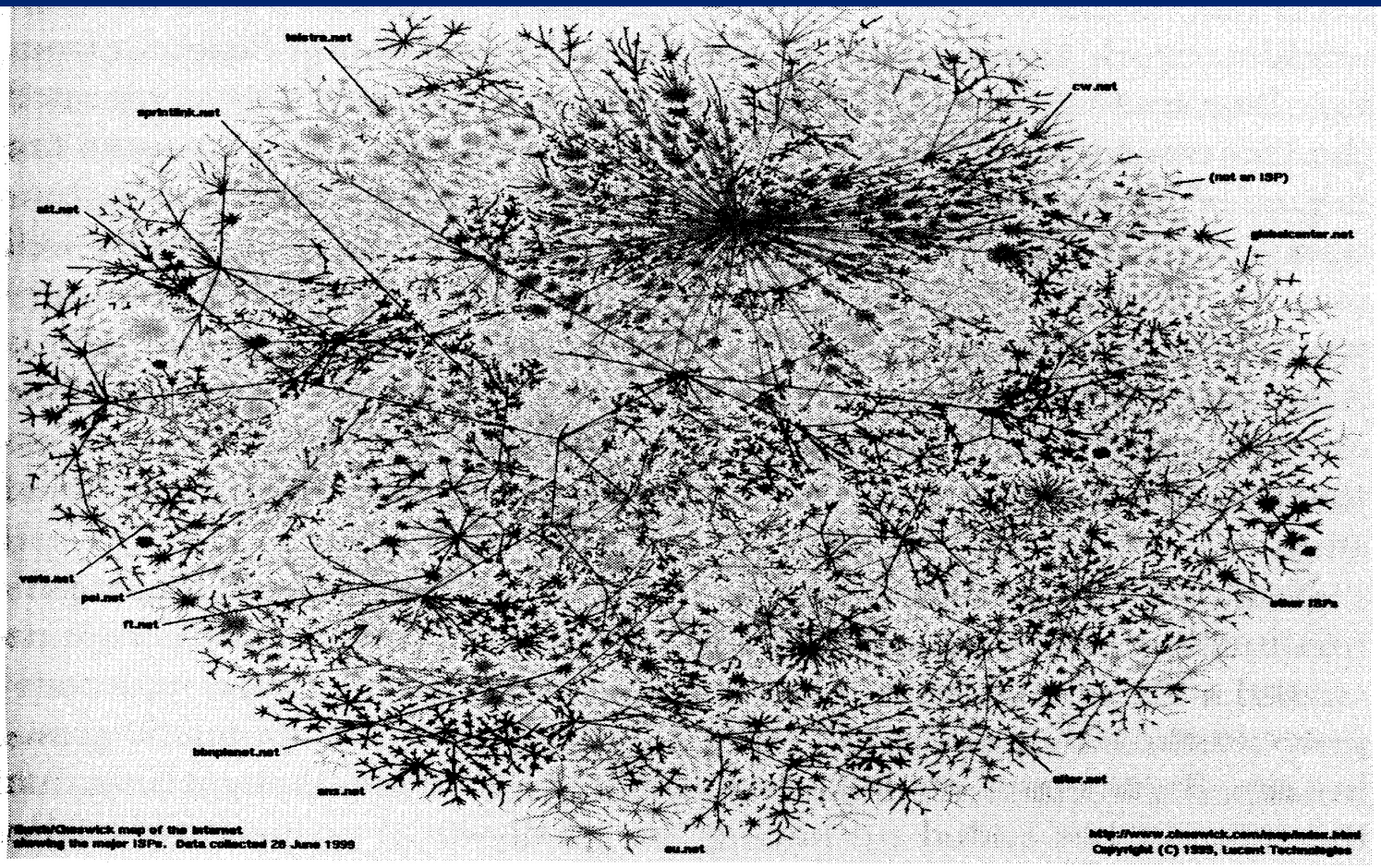


Figure 8. A map of the Internet. (Reprinted by permission of Bill Cheswick and Lucent Technologies.)

Connectivity of Links in Internet: Power Law Behavior

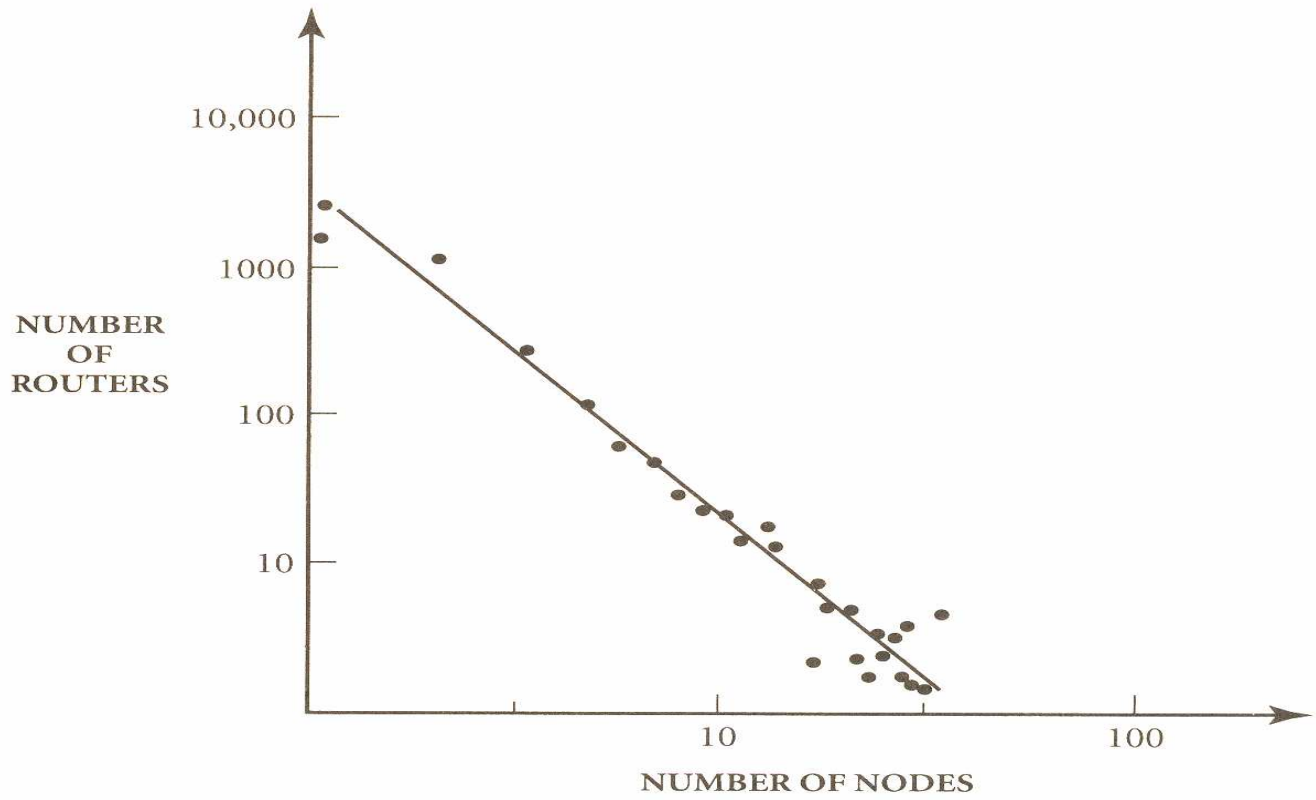


Figure 9. The distribution of Internet "nodes" according to how many links they possess. The curve follows a simple "power law" pattern. (Adapted from M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the Internet topology, *Comput. Commun. Rev.* 29, 251 [1999].)

Statistics of Communication Networks:

TABLE II. Communication networks. Data on the World Wide Web from http://www.nd.edu/~networks contains $N = 325\,729$ documents and $K = 1\,090\,108$ links [12], while the Internet database is taken from http://moat.nlanr.net and has $N = 6474$ nodes and $K = 12\,572$ links.

	E_{glob}	E_{loc}
WWW	0.28	0.36
Internet	0.29	0.26

Alternative Way to Characterize Network Connectivity

- Measuring network connectivity…
- Connectivity length D_local and D_global
- [Marchiori & Latora, Physica A285(2000)539.

- 1/E_global $\Leftrightarrow$ D_global $\Leftrightarrow$ L,
- 1/E_local $\Leftrightarrow$ D_local $\Leftrightarrow$ 1/C

- SW: C large + L small
- $\Leftrightarrow$ E_local + E_global both large
- $\Leftrightarrow$ D_local + D_global both small

Properties of Scale-Free Networks:

- Rapid propagation of computer viruses
- (no threshold).
- Stability against local errors.
- Vulnerable to attack in case of failure of hub.

Propagation of Infectious Deseases:

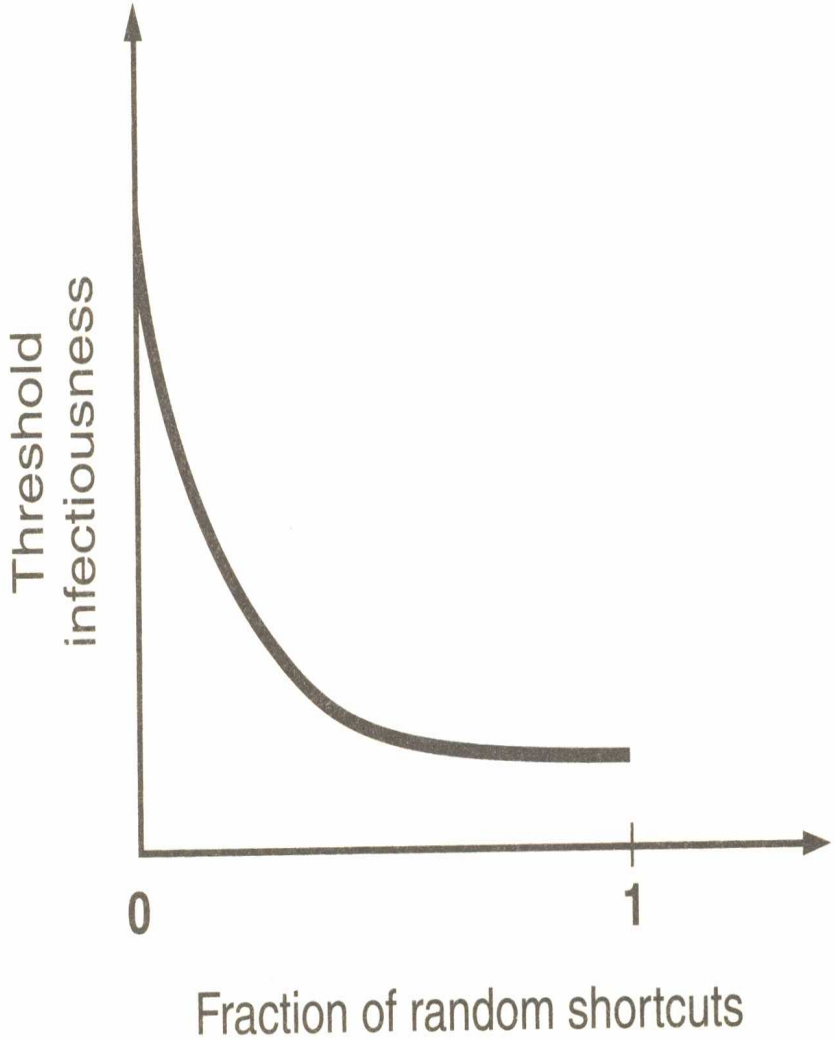


Figure 6.9. The threshold infectiousness required for an epidemic to occur decreases dramatically for small amounts of randomness in a network.

High Vulnerability

Most efficient SW networks: SW with single center node. T. Niskawa et al., Phys. Rev. E 66 (2002) 046139. Problem: High vulnerability by failure of central node.

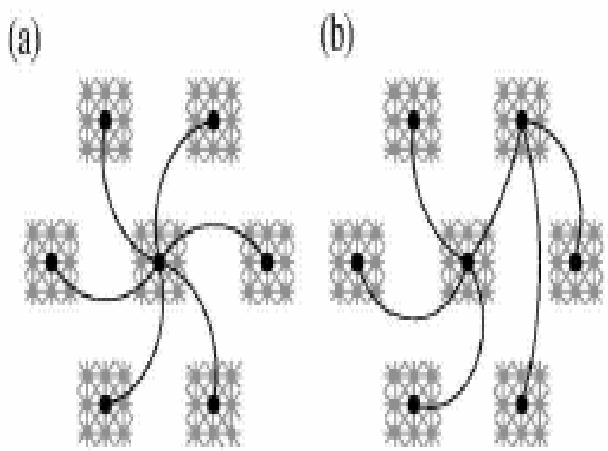


FIG. 1. Examples of shortcut configuration with (a) a single center and (b) two centers.

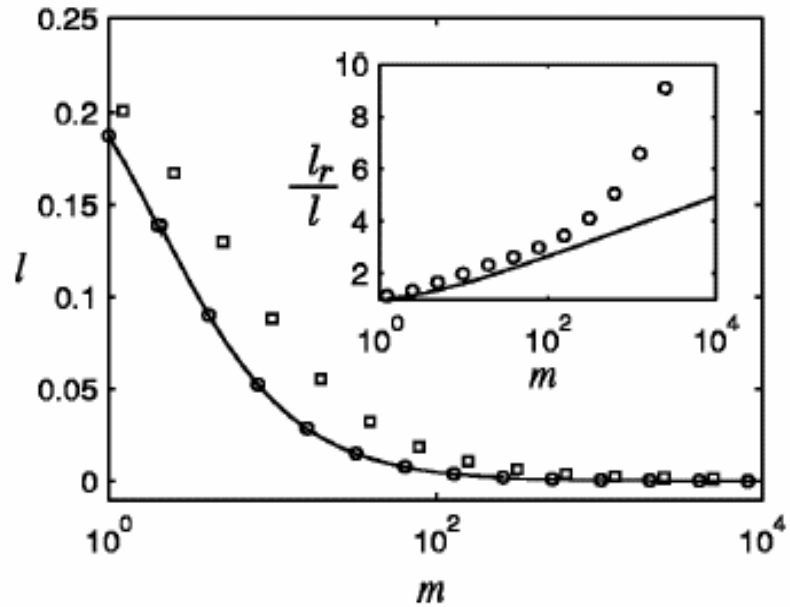


FIG. 4. Normalized path length of the network as a function of the number m of shortcuts for $k=1$. The continuous curve is Eq. (1). The circles and squares are the numerical computation of l for the configuration with a single center and of l_r over 10 random shortcut configurations, respectively. The inset shows the ratio l_r/l computed from numerical simulations (circles) and from theoretical results (1) and (2) for $N=\infty$ (continuous line). $N=10^4$ was used for numerical computations.

PART II. SWN in Neuroscience: Experiments

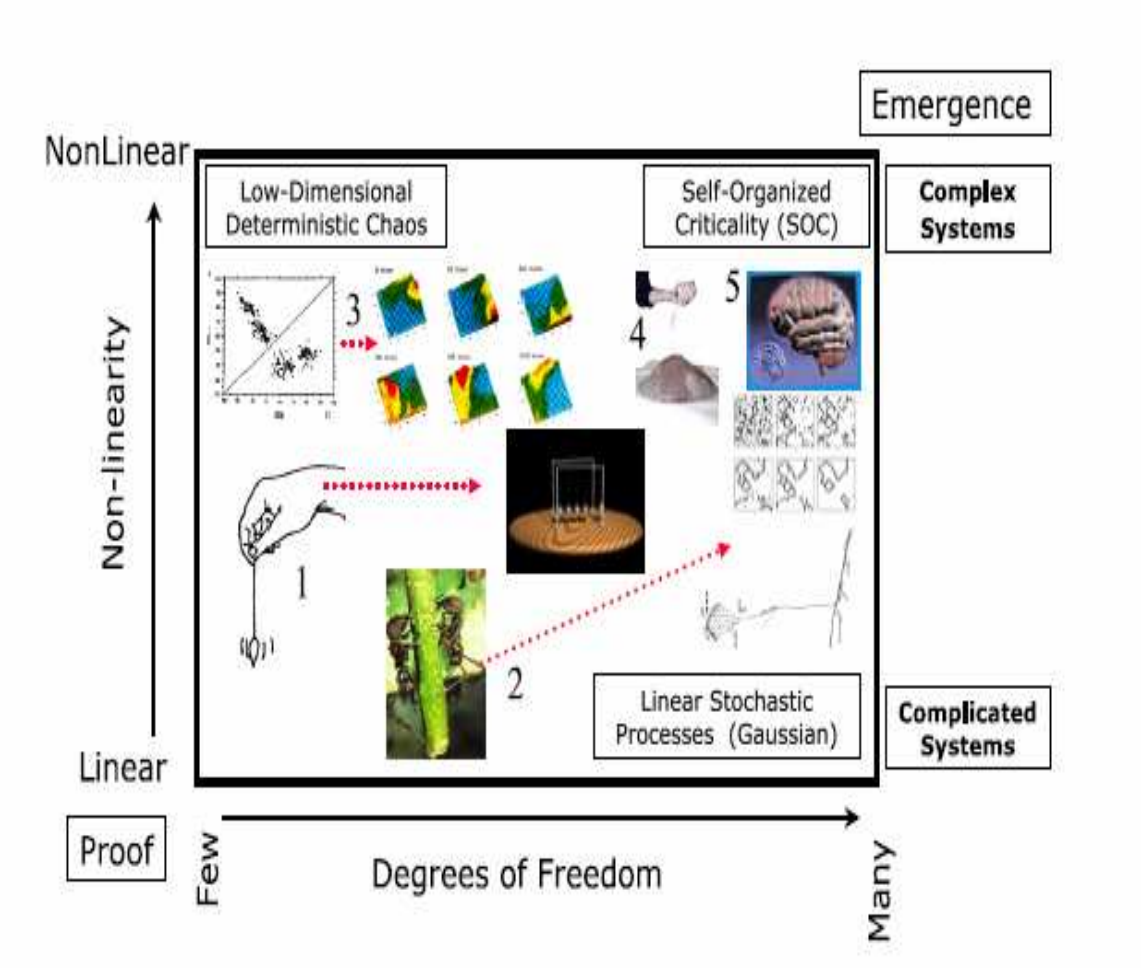
- Cat cortex
- Macaque visual cortex.
- Human brain: Activity network from magnetic resonance imaging.

TABLE I. Macaque and cat cortico-cortical connections [19]. The macaque database contains $N = 69$ cortical areas and $K = 413$ connections [20]. The cat database has $N = 55$ cortical areas (including hippocampus, amygdala, entorhinal cortex, and subiculum) and $K = 564$ (revised database and cortical parcellation from [21]). The nervous system of C. elegans consists of $N = 282$ neurons and $K = 2462$ links which can be either synaptic connections or gap junctions [24].

	E_{glob}	E_{loc}
Macaque	0.52	0.70
Cat	0.69	0.83
C. elegans	0.46	0.47

Dynamical Systems and Complexity

A cartoon representation of the parameter space for various classes of dynamical systems. The simplest ones ``live'' in the left bottom corner, where analysis and formal proofs are the techniques expected, but many fundamental problems in biology correspond to areas distant from that land. Relatively simple dynamics gets sophisticated as the nonlinear term acquires relevance (moving upward in the graph) or as the number of degrees of freedom increases (moving to the right). Pictorial examples include: (1)the transition from one to many coupled pendulums, (2) few foraging ants to the entire swarm [5],(3) from the chaotic dynamics of an isolated cardiac cell \cite{chialvo90} to the spatiotemporal spiral waves in the heart [7], (4) a sand pile and, of course, (5) the brain.



Scale-Free Brain Functional Netw.

V.M. Eguiluz et al., cond-mat/0309092

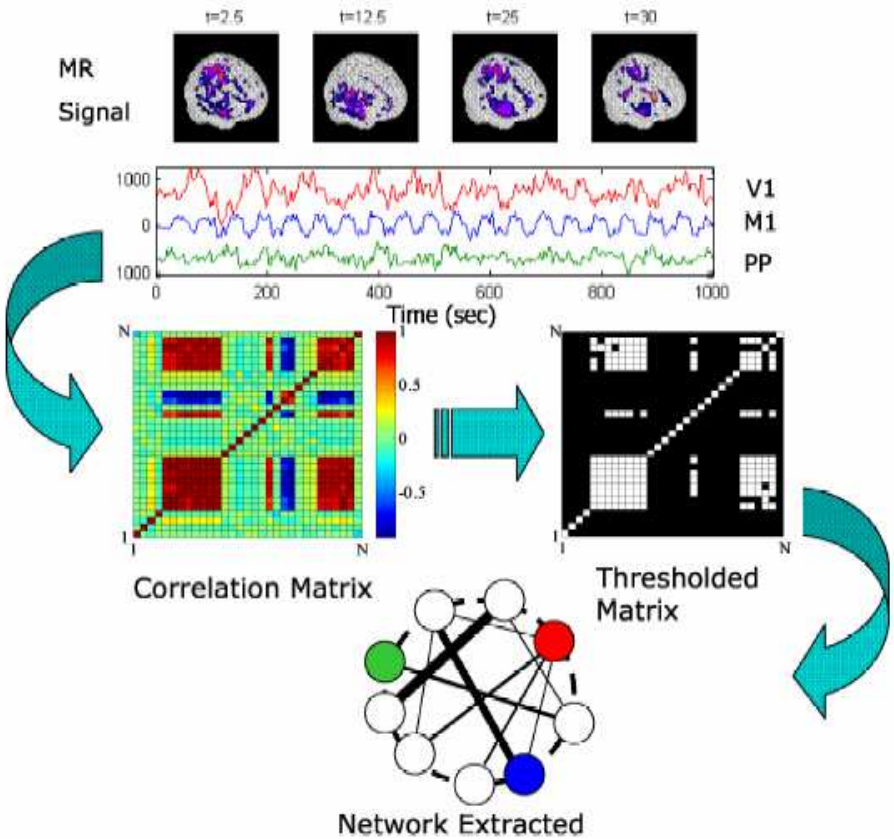


FIG. 1: Methodology used to extract the networks from the signals. The correlation matrix is calculated and then used to define the network among the highest correlated nodes. Top four images represent snapshots of activity and the three traces correspond to selected voxels from visual (V1), motor (M1) and posterio-parietal (PP) cortices.

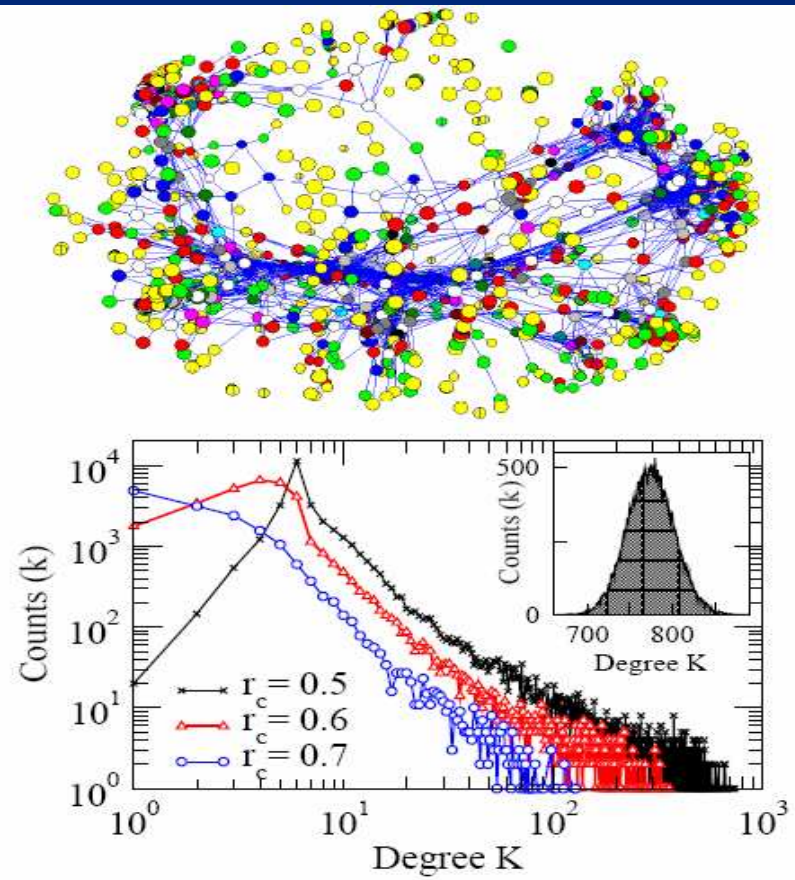


FIG. 2: Example of a network extracted using the methods described in Fig. 1. Top panel shows a pictorial representation of the network (Nodes colored according to its degree: yellow = 1, green = 2, red = 3, blue=4, etc). The bottom panel shows the degree distribution for three values of the correlation threshold. The inset depicts the degree distribution for an equivalent randomly connected network.

Brain Functional Networks …

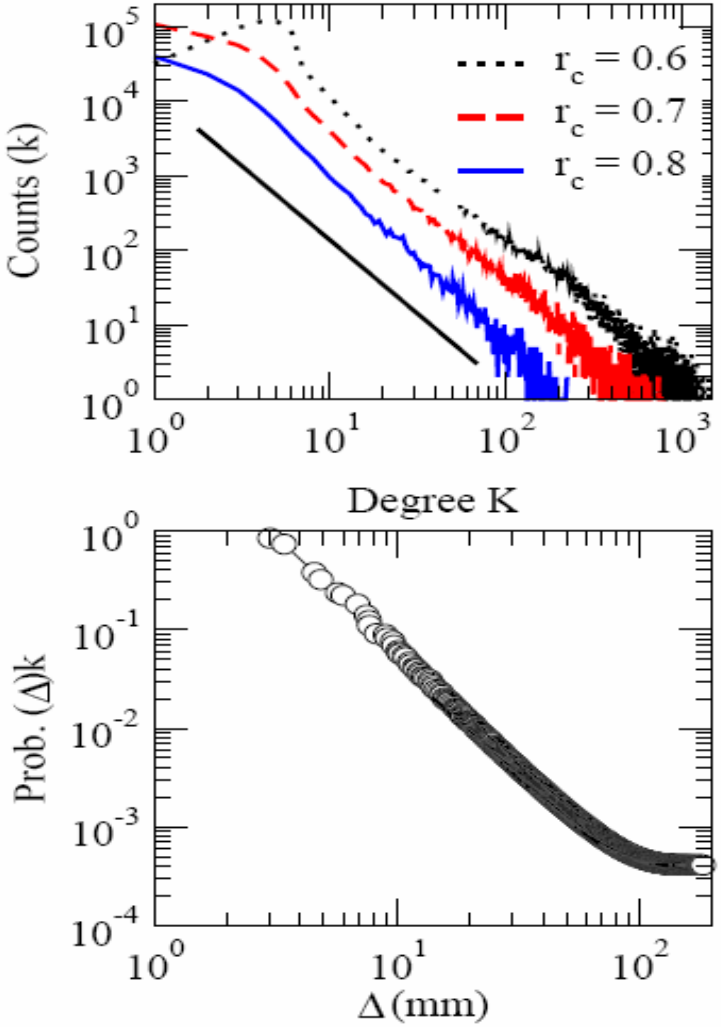


FIG. 3: Average scaling taken from 22 networks extracted from seven subjects. Top Panel: Average degree distribution. The straight line illustrates a decay of k^{-2}. Bottom panel: Average probability of finding a link between two nodes separated by a distance $x < \Delta$ (using $r_c = 0.6$).

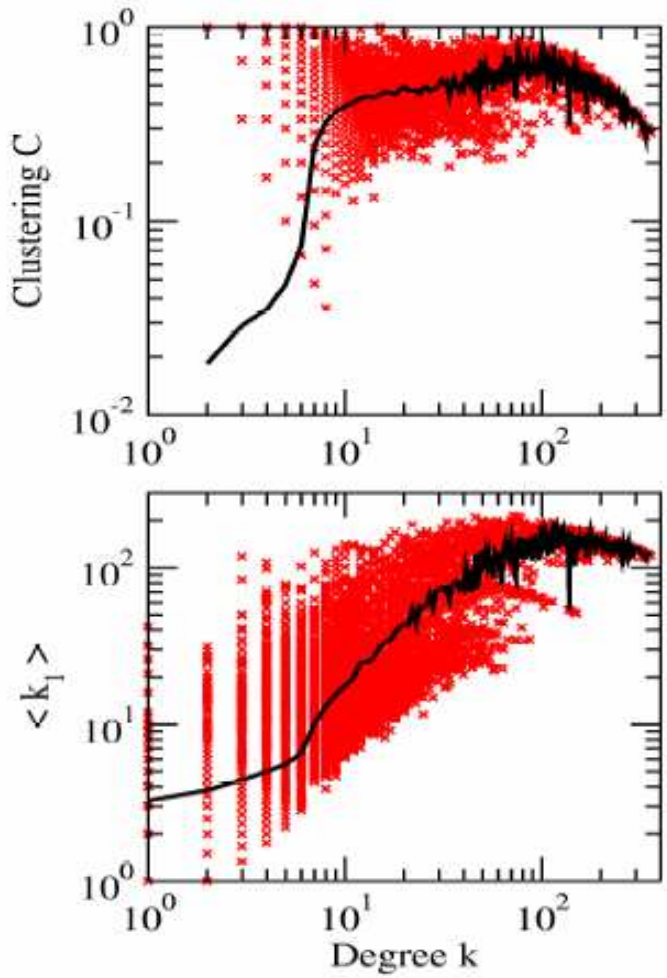
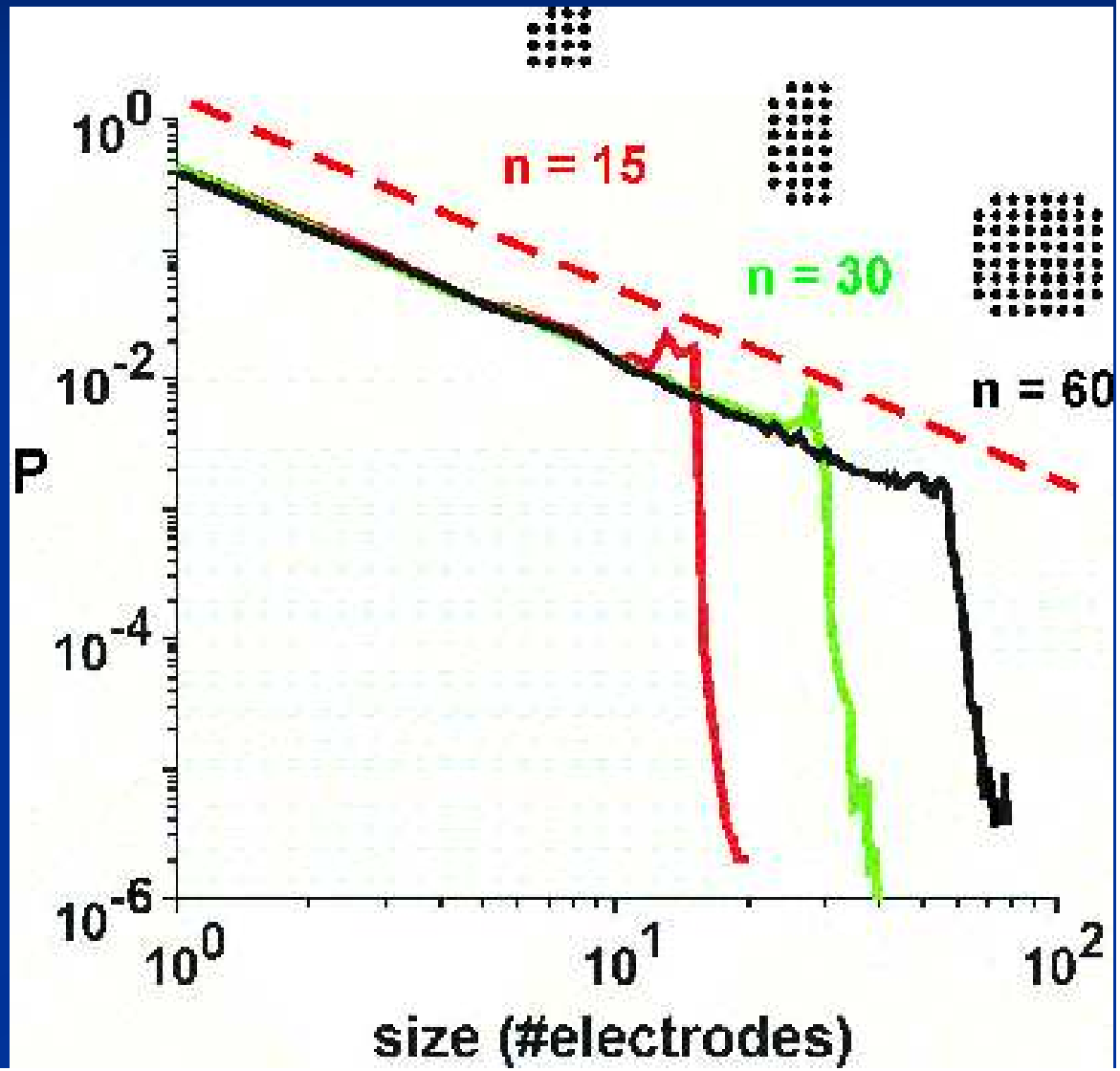


FIG. 4: Top Panel: Plot of clustering vs. degree. Bottom panel: Plot of a neighboring node's degree vs. degree illustrates the assortative feature. Symbols represents individual data and continuous lines the average values for nodes with the same degree. (Same subject shown in Fig. 2, with r_c=0.6).

Neural Avalanches: Size Distribution Follows Power Law

The size distribution of neuronal avalanches in mature cortical cultured networks follows a power law with an exponent ~3/2 (dashed line). The data, re-plotted from Figure 4 of [30] shows the probability of observing an avalanche covering a given number of electrodes for three sets of grid sizes shown in the insets with n=15, 30 or 60 sensing electrodes (equally spaced at 200 μm). The statistics is taken from data collected from 7 cultures in recordings lasting a total of 70 hours and accumulating 58000 (+- 55000)avalanches per hour (mean +- SD)



Statistics of Brain networks

r_c	N	C	L	$\langle k \rangle$	λ	C_{rand}	L_{rand}
0.6	31503	0.14	11.4	13.41	2.0	4.3×10^{-4}	3.9
0.7	17174	0.13	12.9	6.29	2.1	3.7×10^{-4}	5.3
0.8	4891	0.15	6.	4.12	2.2	8.9×10^{-4}	6.0

Average statistical properties of the brain functional networks

$Network$	N	C	L	$\langle k \rangle$	λ	C_{rand}	L_{rand}
C. Elegans	282	0.28	2.65	7.68	NA	0.025	2.1
Macaque VC	32	0.55	1.77	9.85	NA	0.318	1.5
Cat Cortex	65	0.54	1.87	17.48	NA	0.273	1.4

Statistics of relatively smaller networks. Because these networks are not scale-free λ is not indicated.

SWN in Neuroscience: Computational Models

- Fast response and coherence
- Efficient associative memory

SWN of Hodgkin-Huxley Neurons

Model: Hodgkin Huxley neurons in 1-D periodic network.

Result: Fast response and coherent oscillations.

L.F.Lago-Fernandez et al. Phys. Rev. Lett. 84 (2000) 2758.

Possible relevance in neuroscience: Binding problem

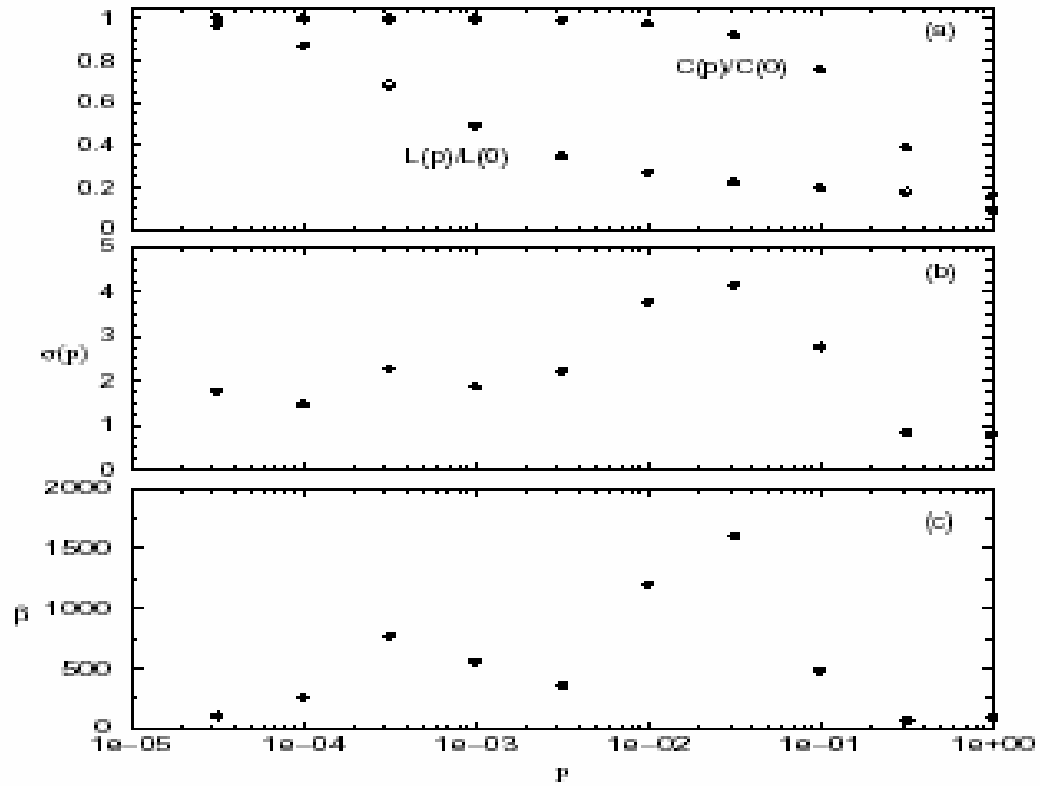


FIG. 1. (a) Characteristic path length $L(p)$ and clustering coefficient $C(p)$ for the family of randomly rewired graphs, normalized to the values $L(0)$ and $C(0)$ of the regular case. (b) Average activity oscillation amplitude $\sigma(p)$, and (c) degree of coherence $\beta(p)$ for the whole range of networks, calculated between $T_1 = 100$ and $T_2 = 200$. All curves are averages over ten realizations of the simulation with parameters $N = 797$, $k = 30$, and $g = 0.015$. An input signal $I_0 = 1.5$ was injected, at $t = 50$, to 80 contiguous neurons (10% of the total).

Efficient Associative Memory Model with SW Architecture

J.W. Bohland and A.A. Minai, Neurocomputing 38-40 (2001) 489.

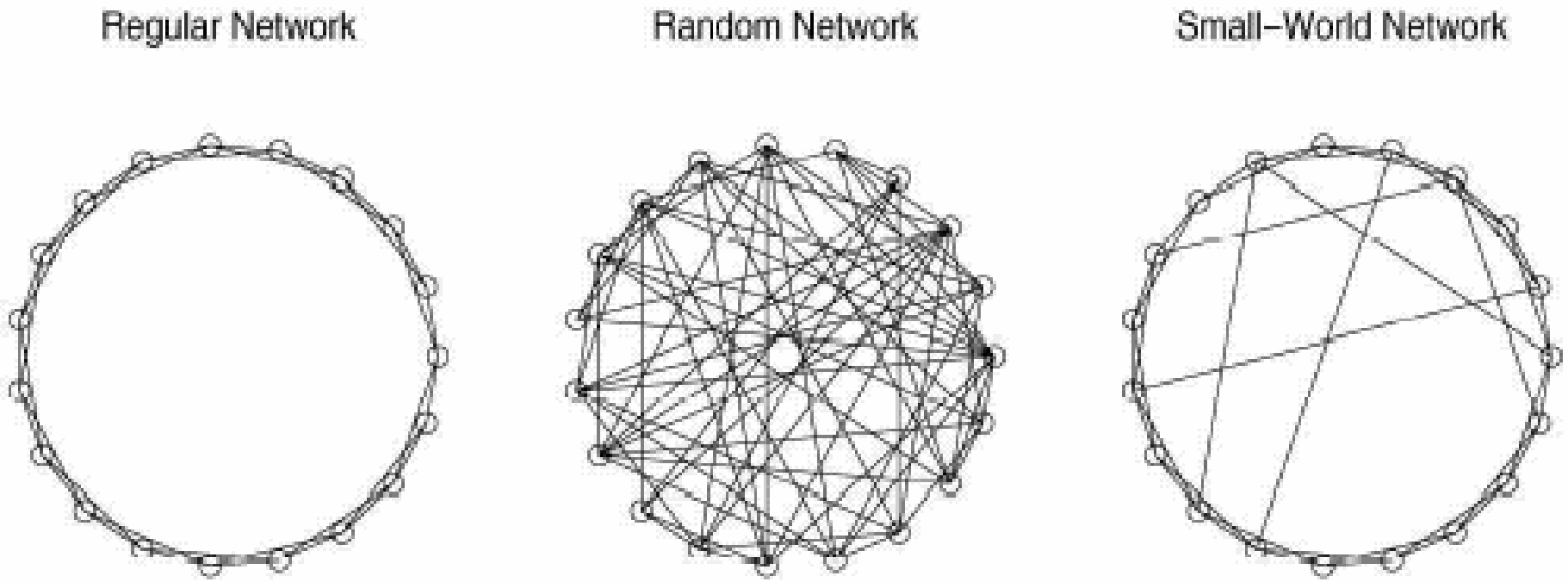


Fig. 1. Network connection topologies. Note that these graphs are undirected, and that in the associative memory networks, each edge is directed.

Efficient Associative Memory …

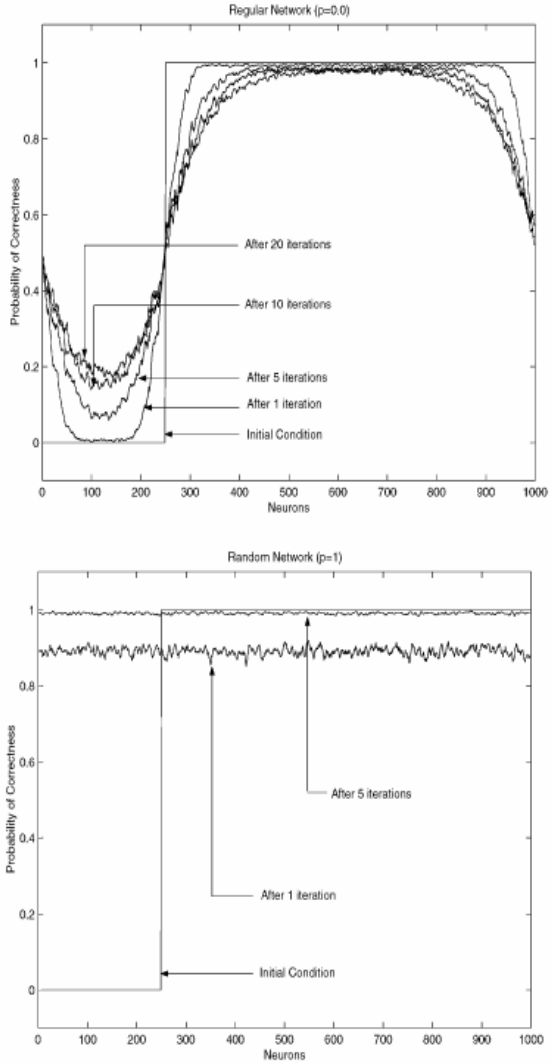
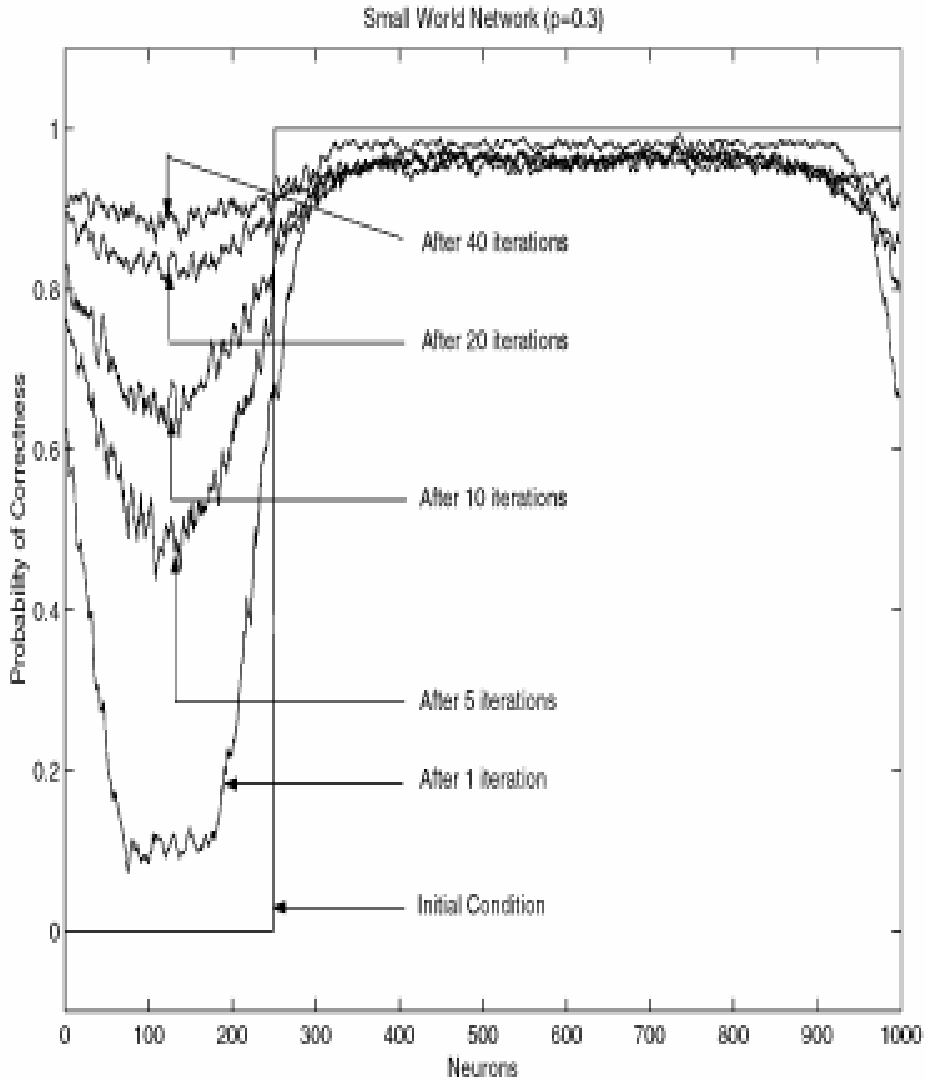
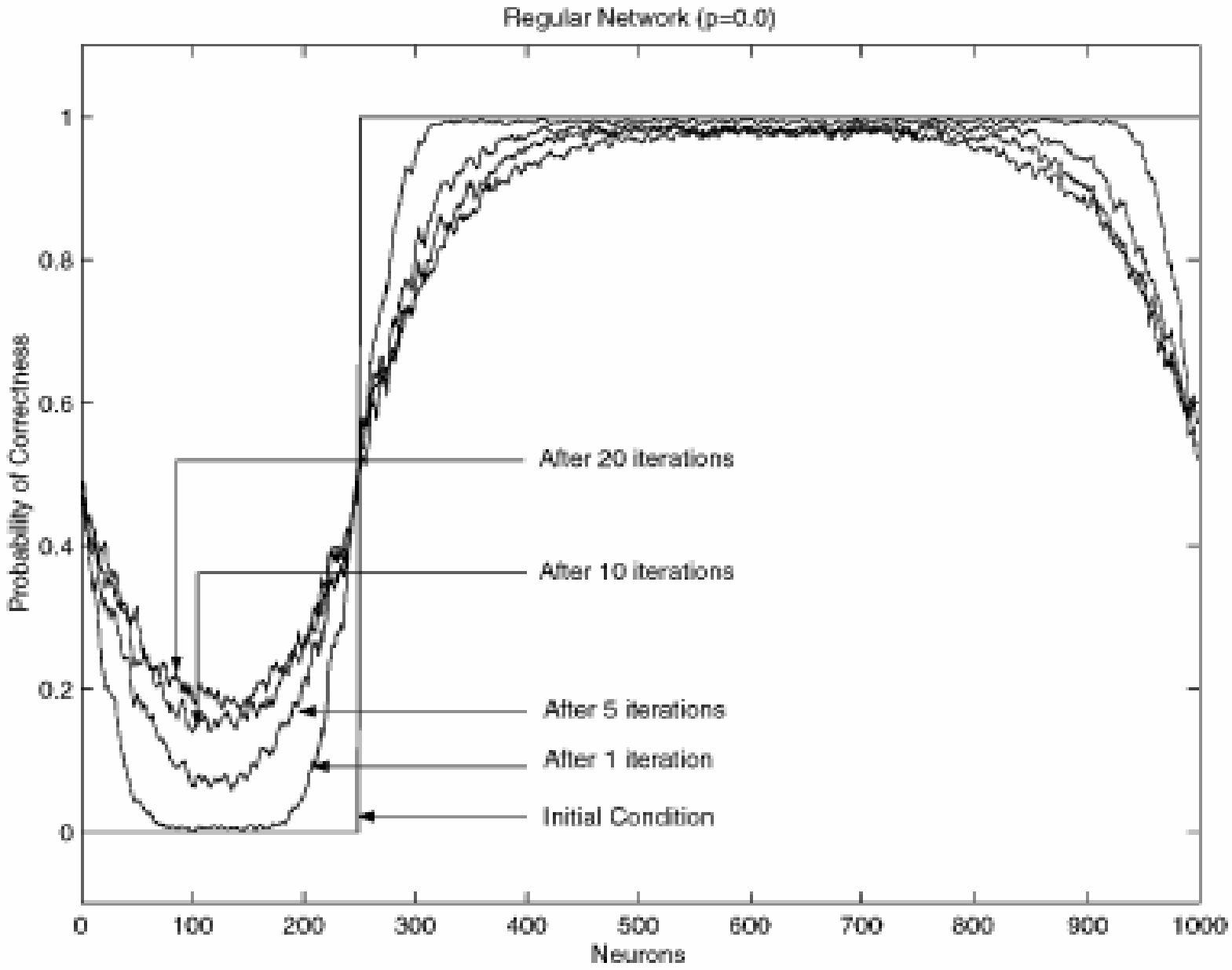


Fig. 3. (left) Regular network recall dynamics: Continuous curves for this and other corresponding graphs are obtained by using a nearest neighbors averaging filter; (right) Random network recall dynamics.




Regular Network (p=0.0)
Probability of Correctness
Neurons
After 20 iterations
After 10 iterations
After 5 iterations
After 1 iteration
Initial Condition
0
0.2
0.4
0.6
0.8
1
0
100
200
300
400
500
600
700
800
900
1000

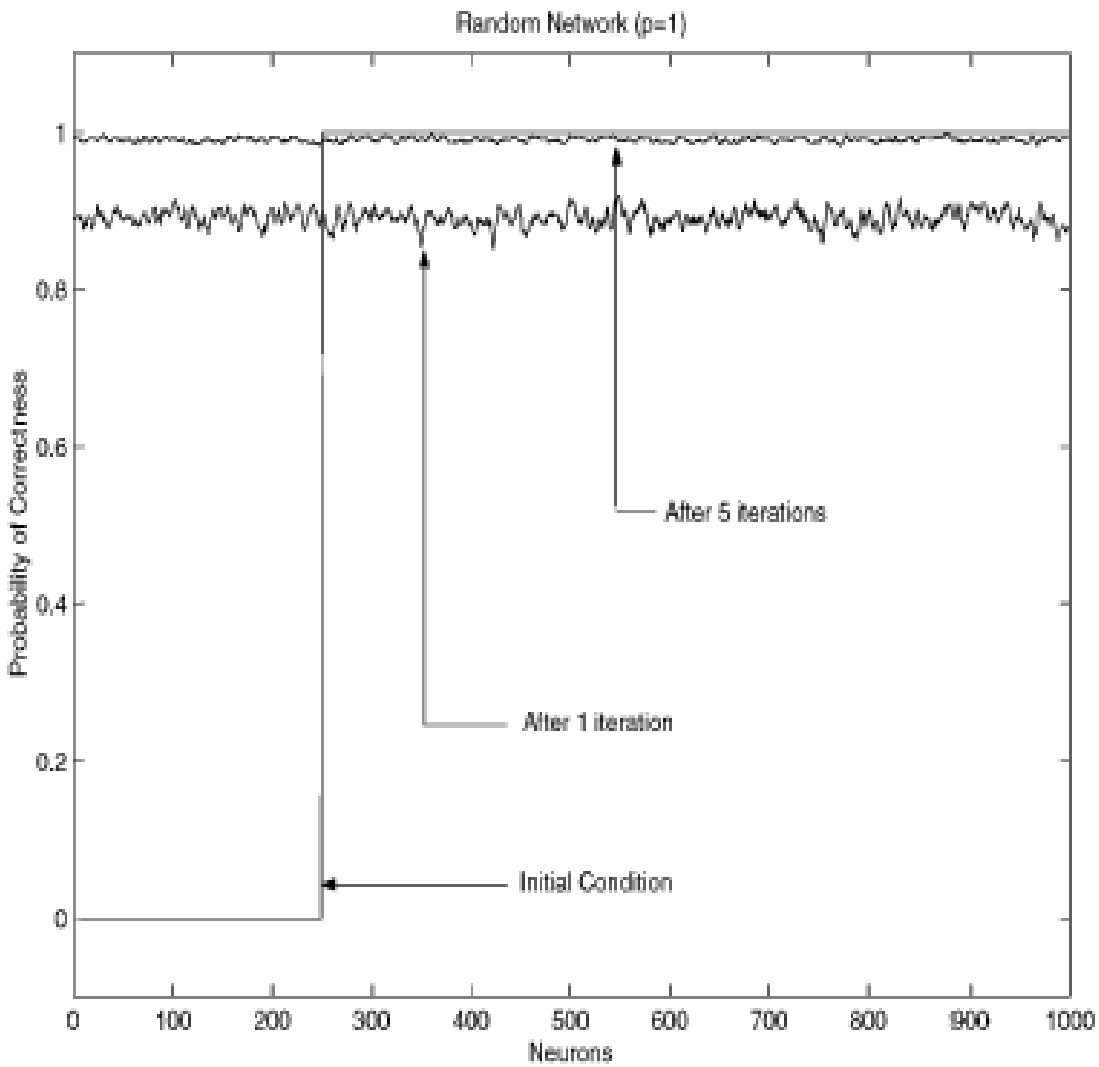
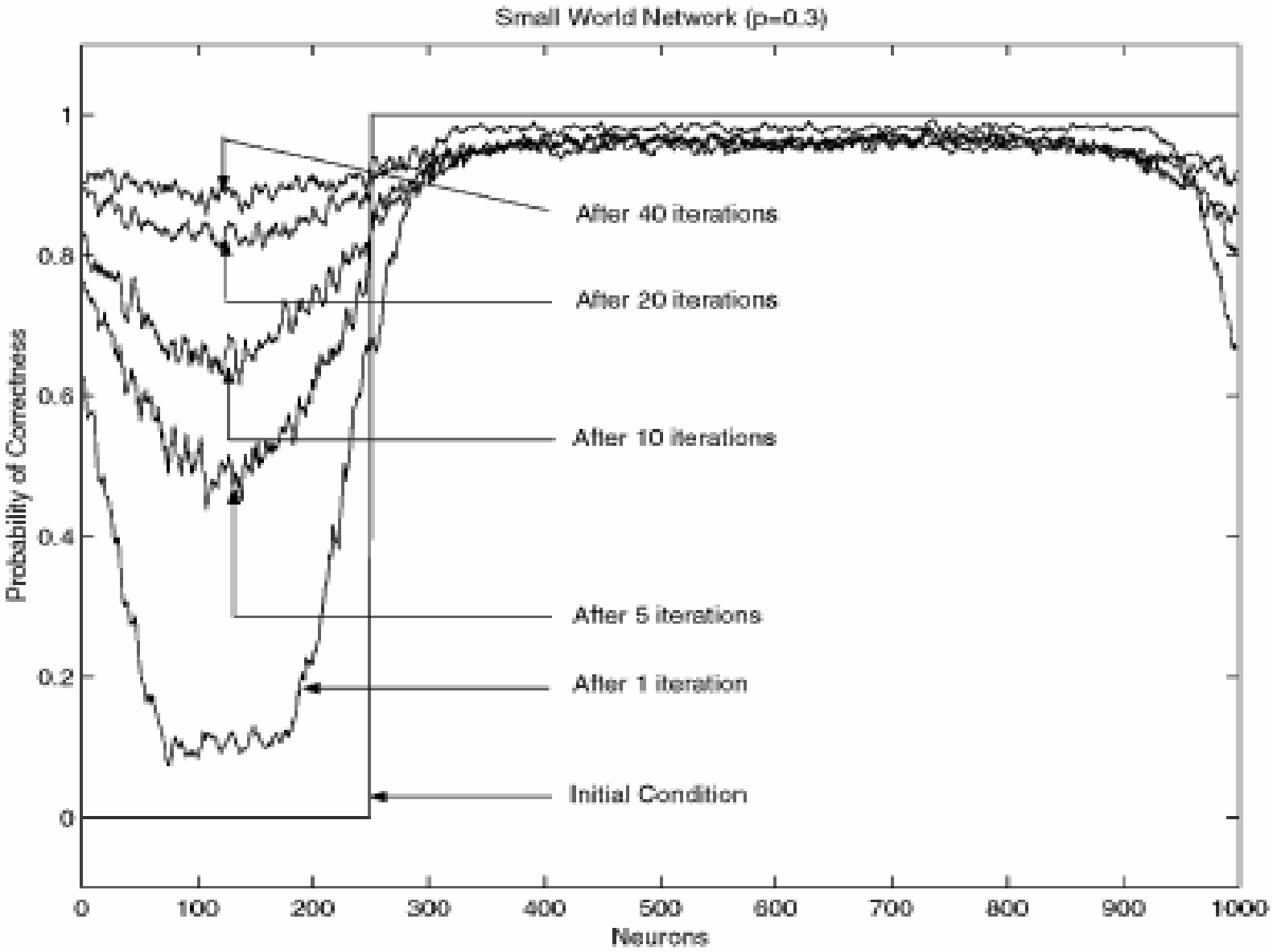


Fig. 3. (left) Regular network recall dynamics: Continuous curves for this and other corresponding graphs are obtained by using a nearest neighbors averaging filter; (right) Random network recall dynamics.


Small World Network (p=0.3)
Probability of Correctness
Neurons
After 40 iterations
After 20 iterations
After 10 iterations
After 5 iterations
After 1 iteration
Initial Condition
0
0.2
0.4
0.6
0.8
1
0
100
200
300
400
500
600
700
800
900
1000

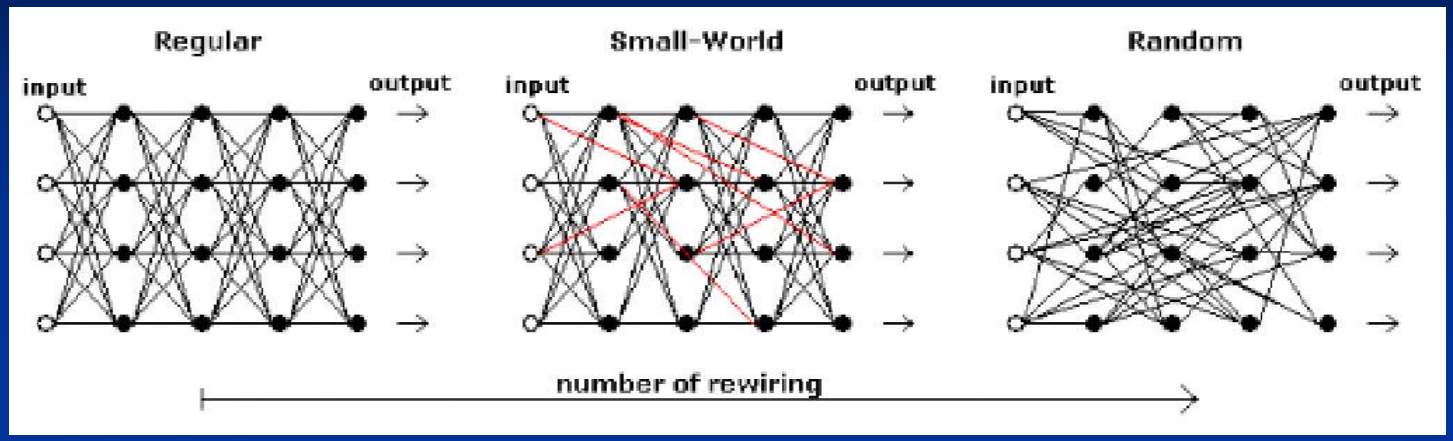
More recent modeling using SWN and Scale-Free Networks:

- Self-Sustained Activity in a Small-World Network of Excitable Neurons, A. Roxin et al. (Northwestern Univ.), Phys. Rev. Lett. 92 (2004) 198101 [simple network model of working memory].
- Efficient Hopfield pattern recognition on a scale-free neural network, D. Stauffer et al., cond-mat/0212601.
- Epilepsy in Small-World Networks, T.I. Netoff et al., J. of Neuroscience 37 (2004) 8075 [model explains short and long bursts from region CA1 and CA3 of hippocampus].

Part III. Fast Learning in Small-World Networks

Multi-layered feed-forward network (visual cortex)

Design of the feed-forward NN



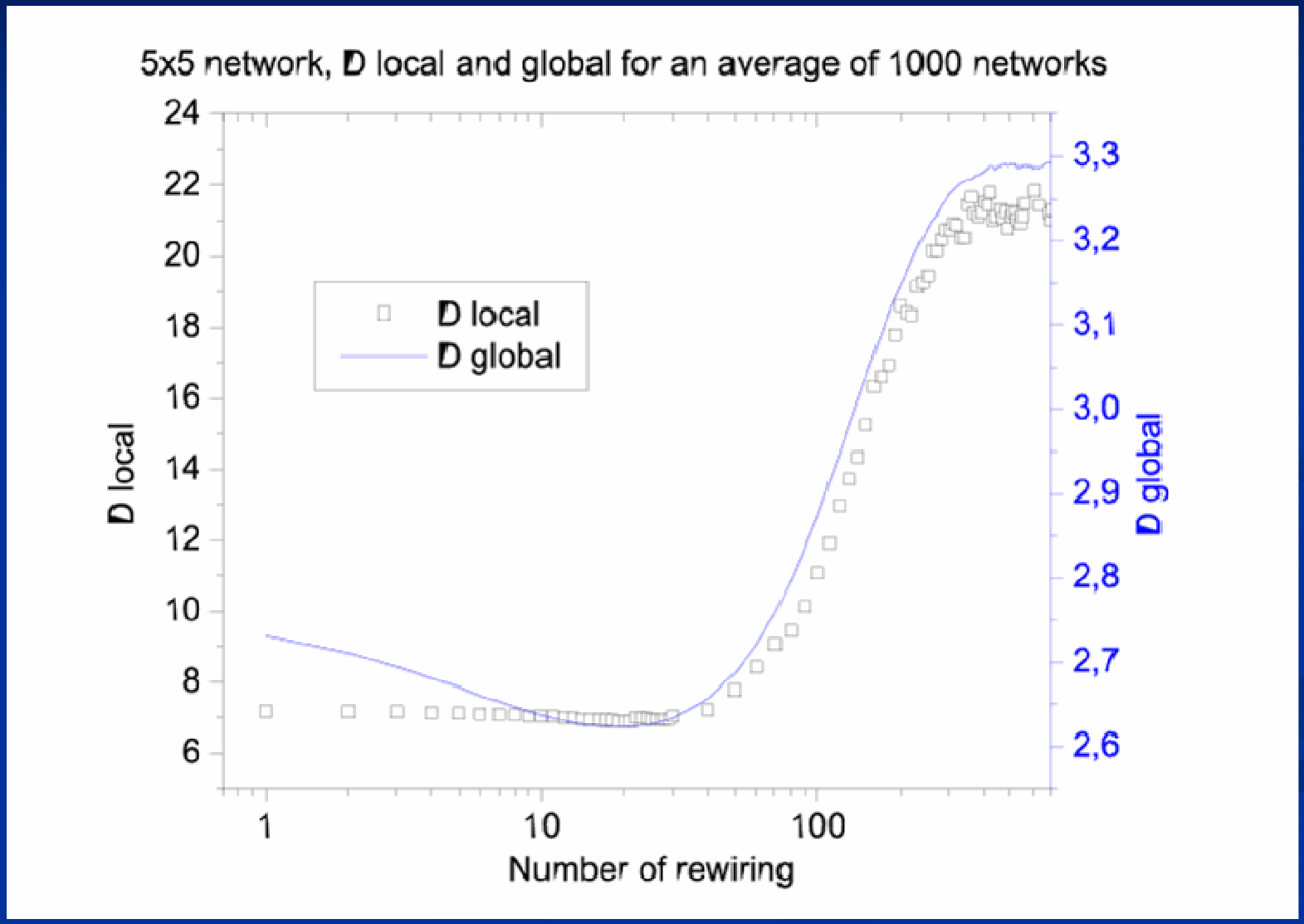
The definition of « Regular », « Small-World » and « Random » are not the same as usual

•Regular : Each neuron is connected to all neurons in the next layer

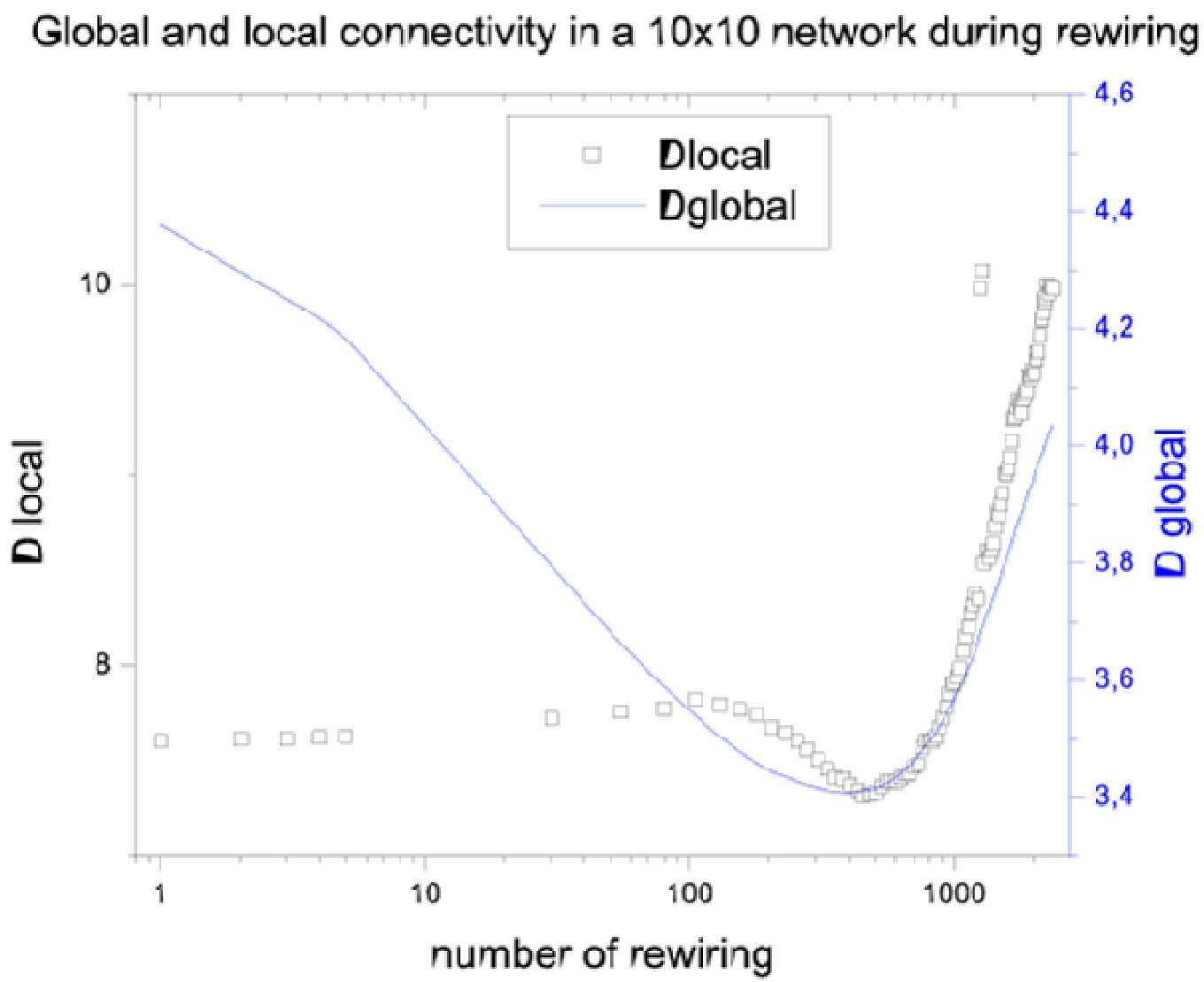
•Random: Each neuron is connected randomly to a forward neuron, no backward connection is allowed

•Small-World: Starting from the regular architecture, some connections to the next layer are rewired to some forward layer

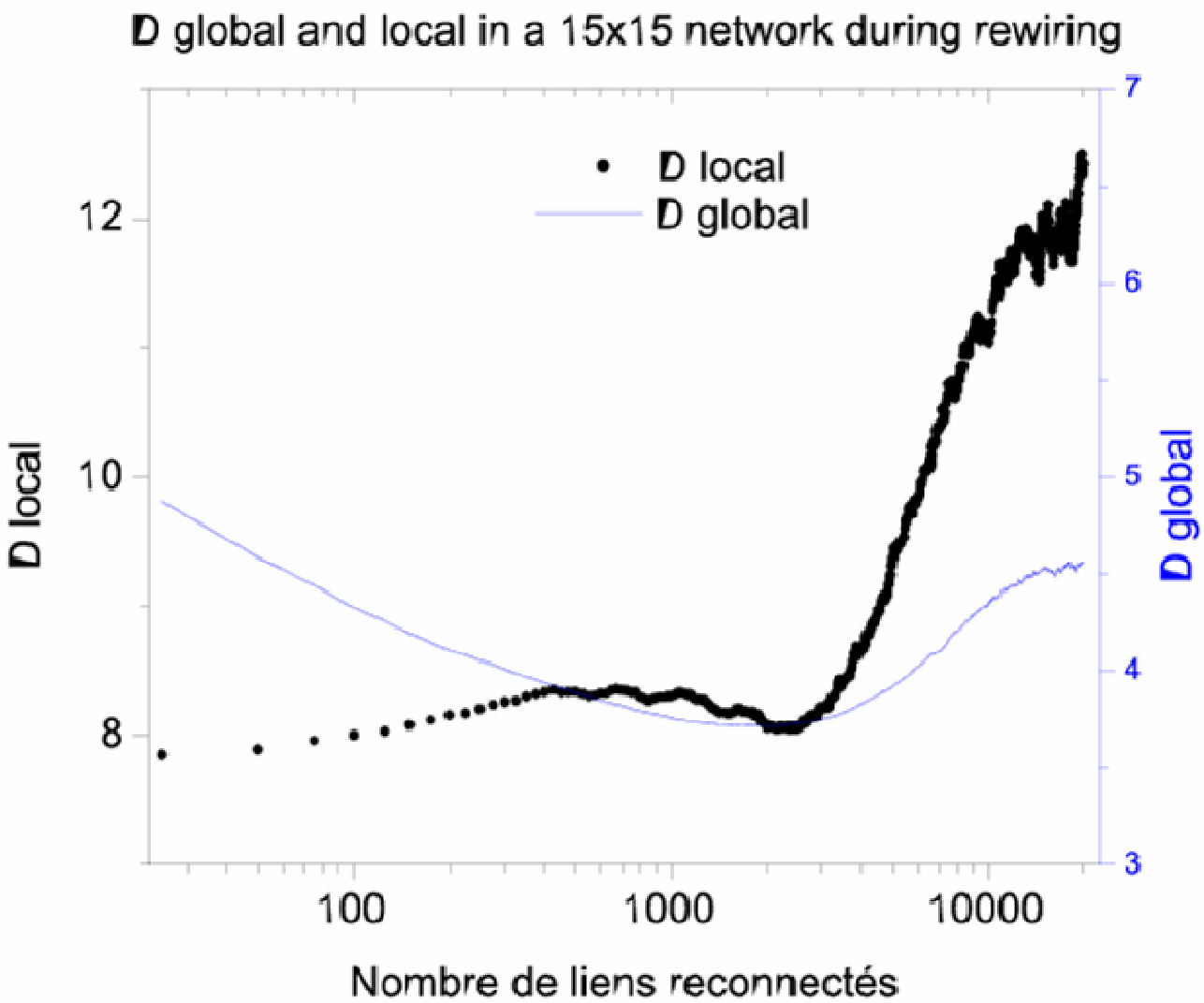
Connectivity: 5x5 NN



Connectivity: 10x10 NN



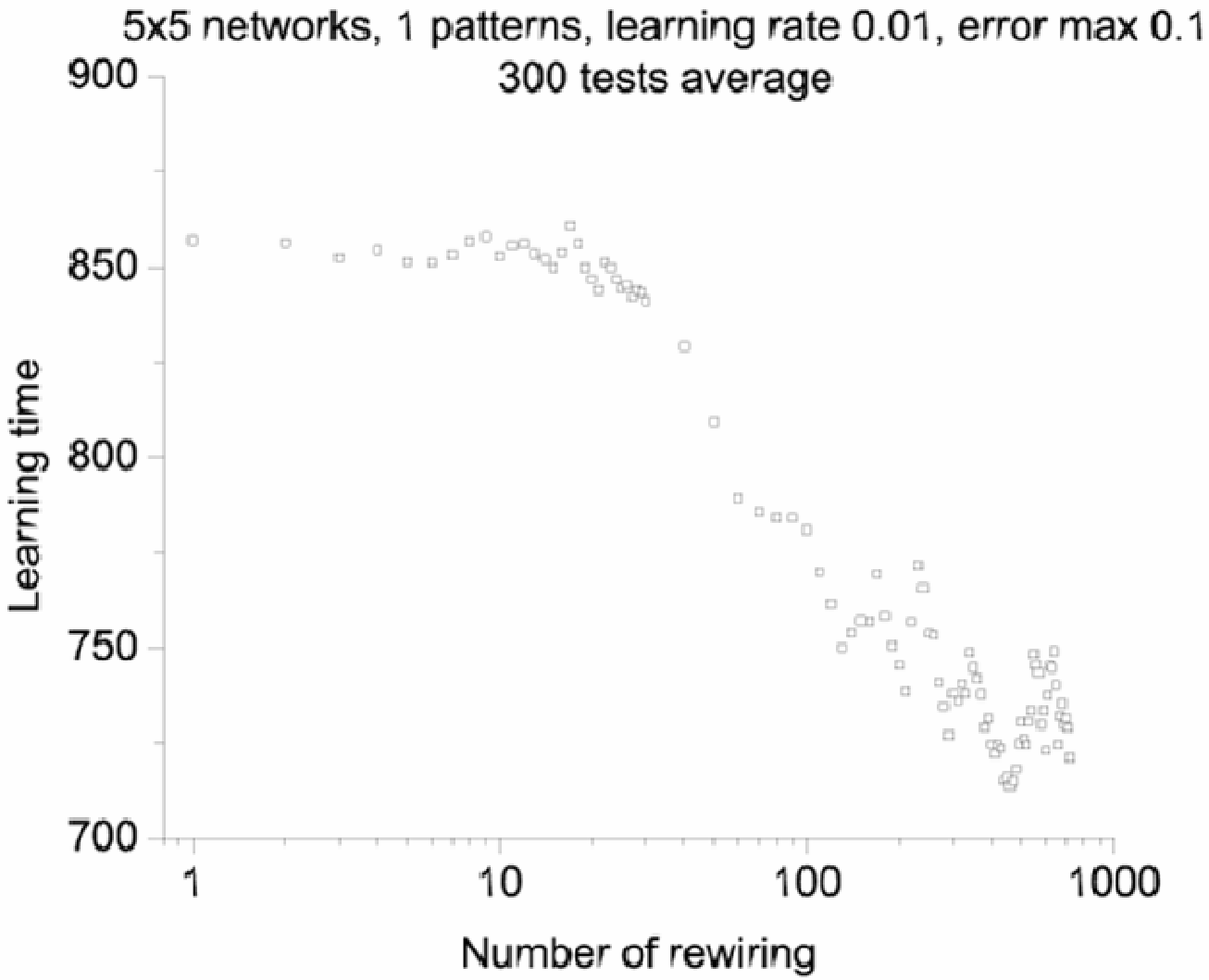
Connectivity: 15x15 NN



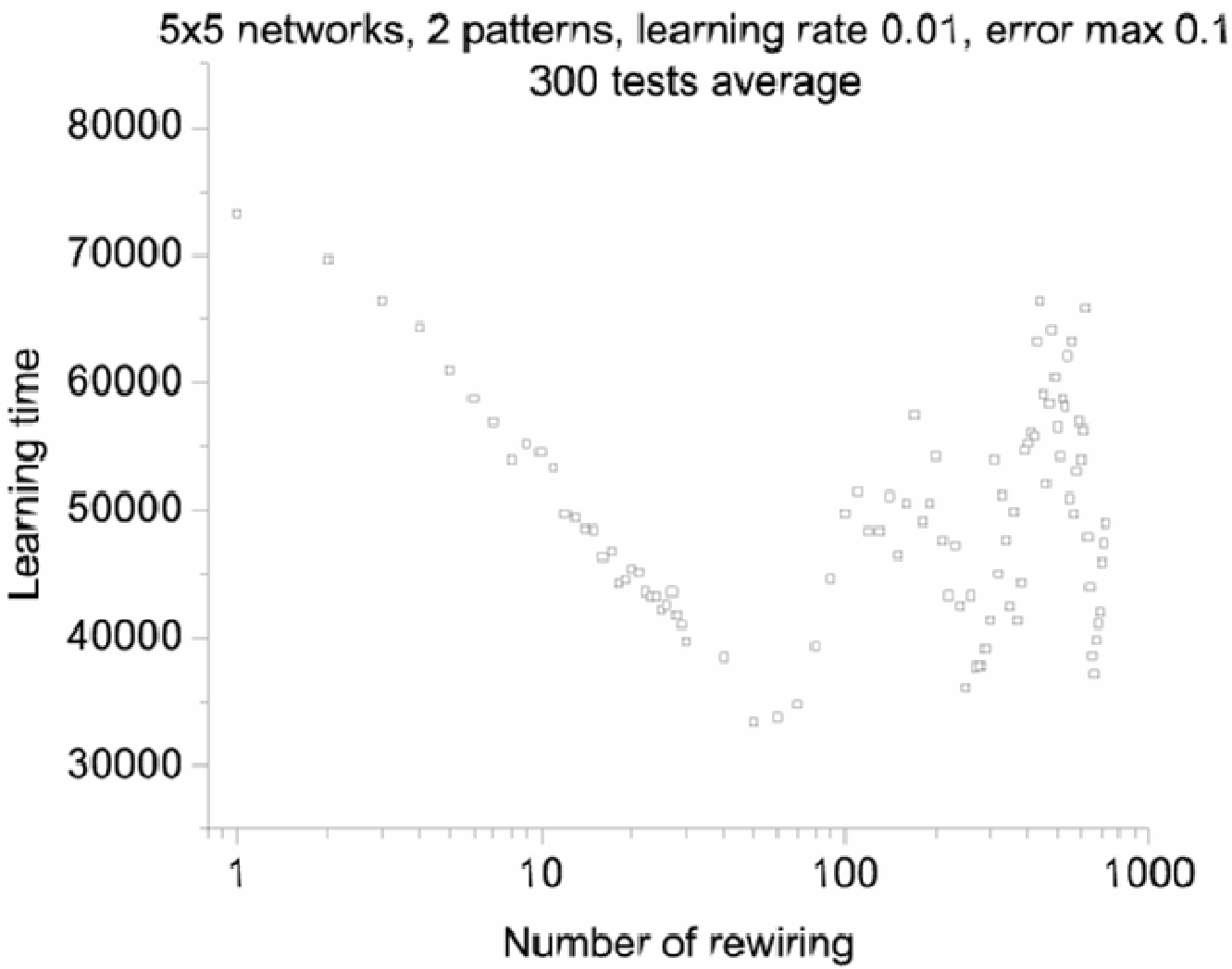
Definition of Learning Time:

Measures how often a pattern as be presented as input to the network, in order that the error of the output pattern becomes smaller than a given error limit.

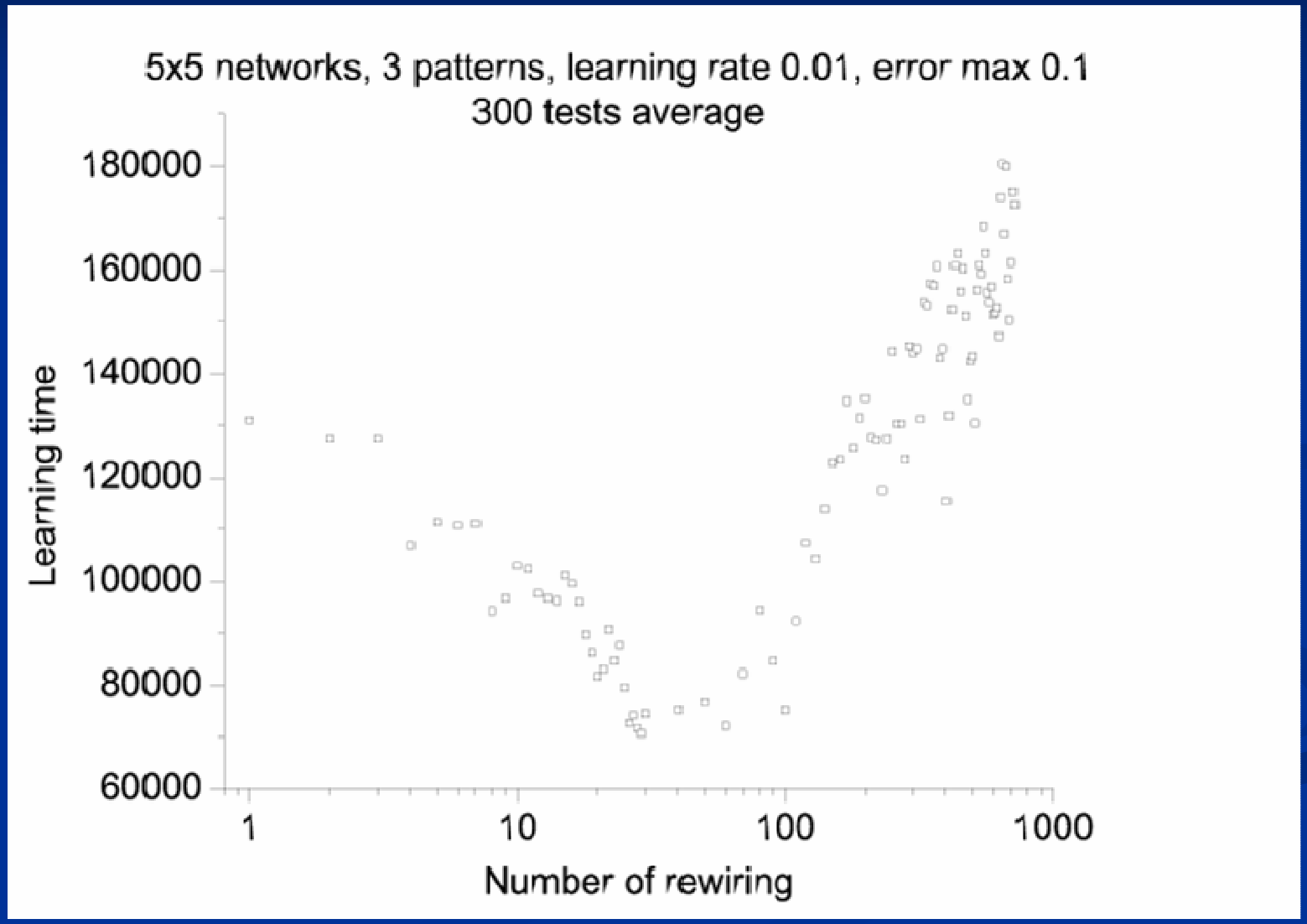
Learning time: 5x5 NN, 1 pattern



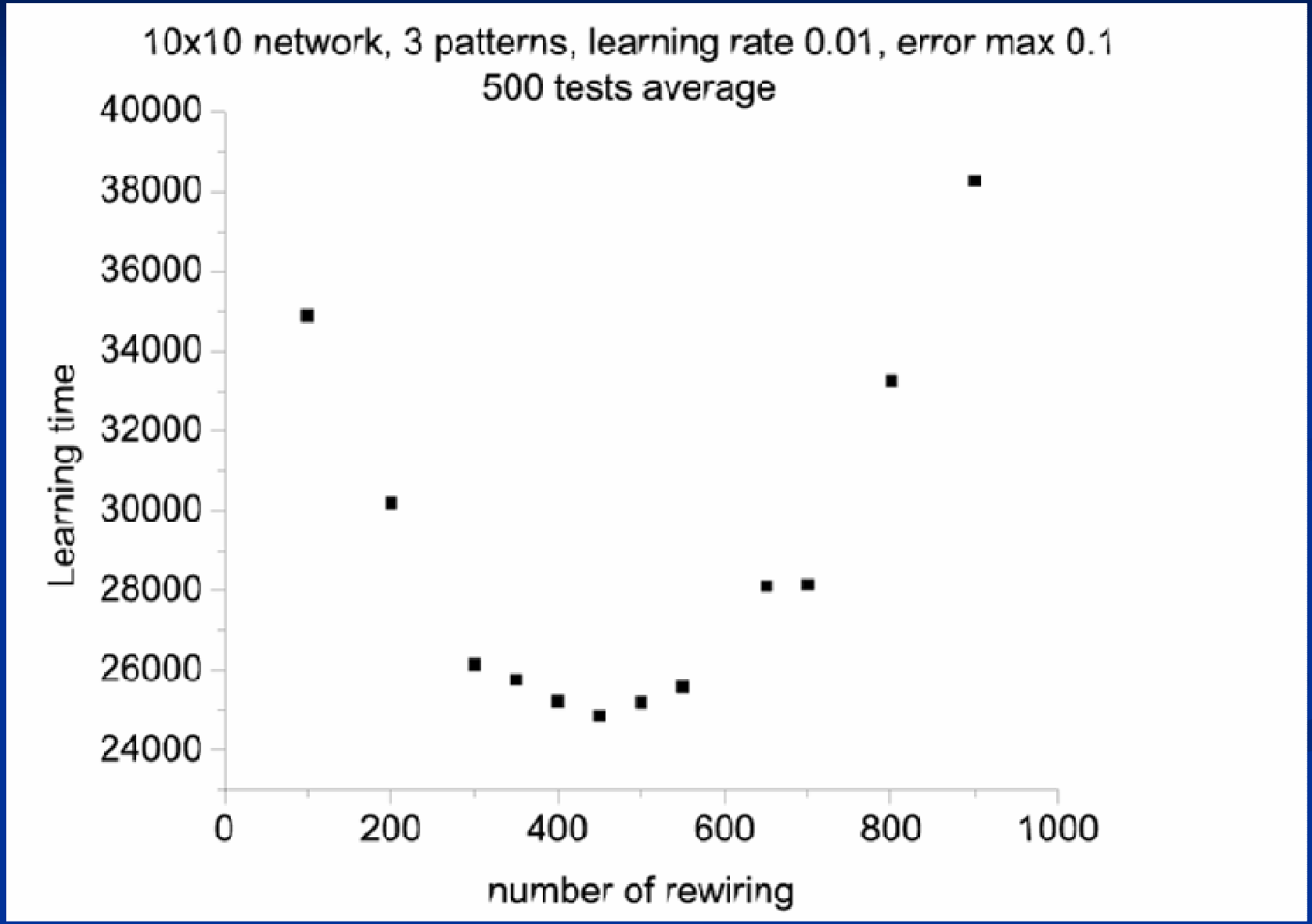
Learning time: 5x5 NN, 2 patterns



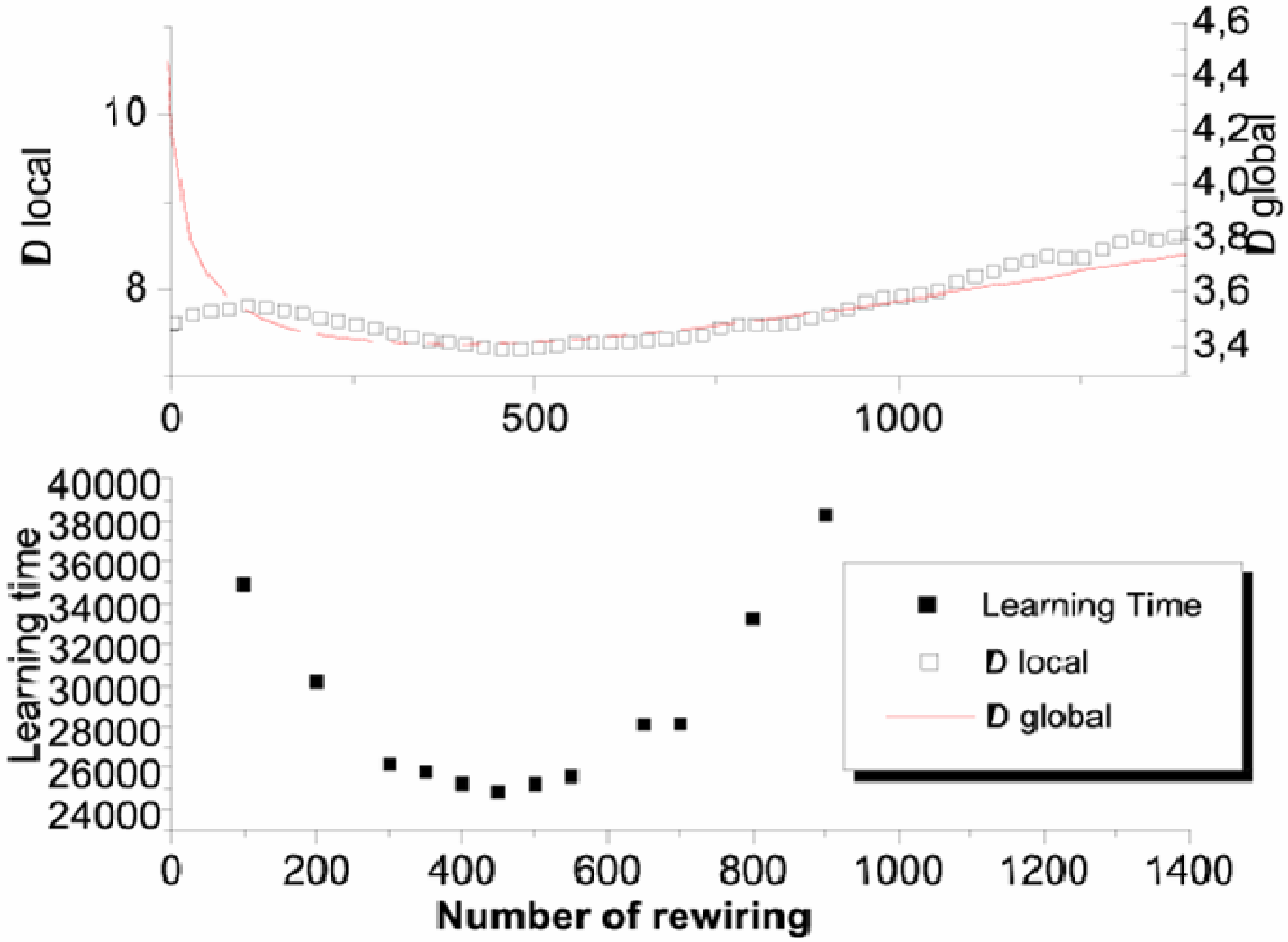
Learning time: 5x5 NN, 3 patterns



Learning time: 10x10 NN, 3 patterns



Coincidence of minima of D_loc, D_glob and T_learn: SWN learns fastest.

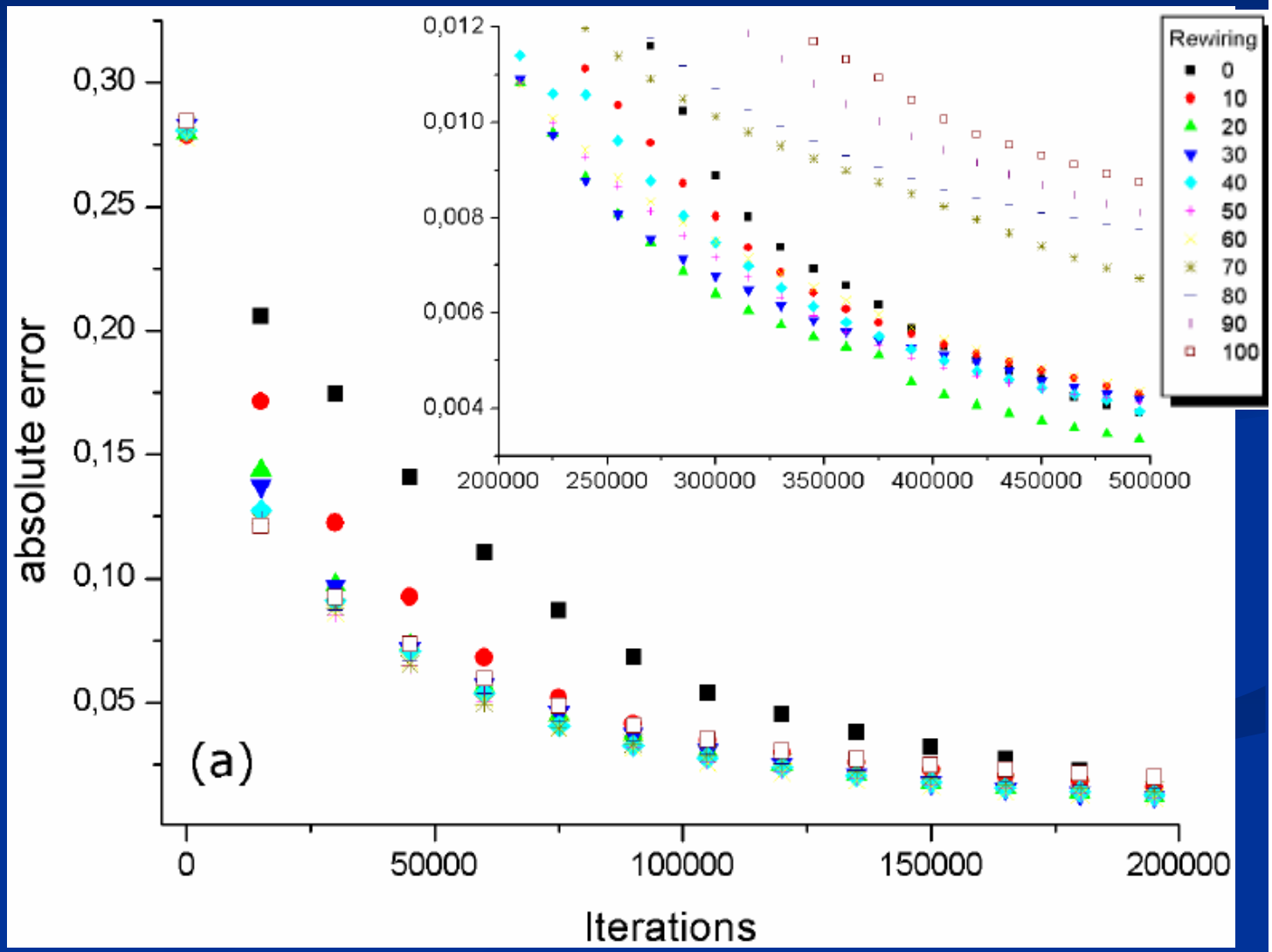


Impass if too
many layers. But
a few short-cuts
do help.

Dependence on number of patterns: Few patterns

Simulation with 5 Neurons per layer and 5 layers. Learning of 5 patterns.

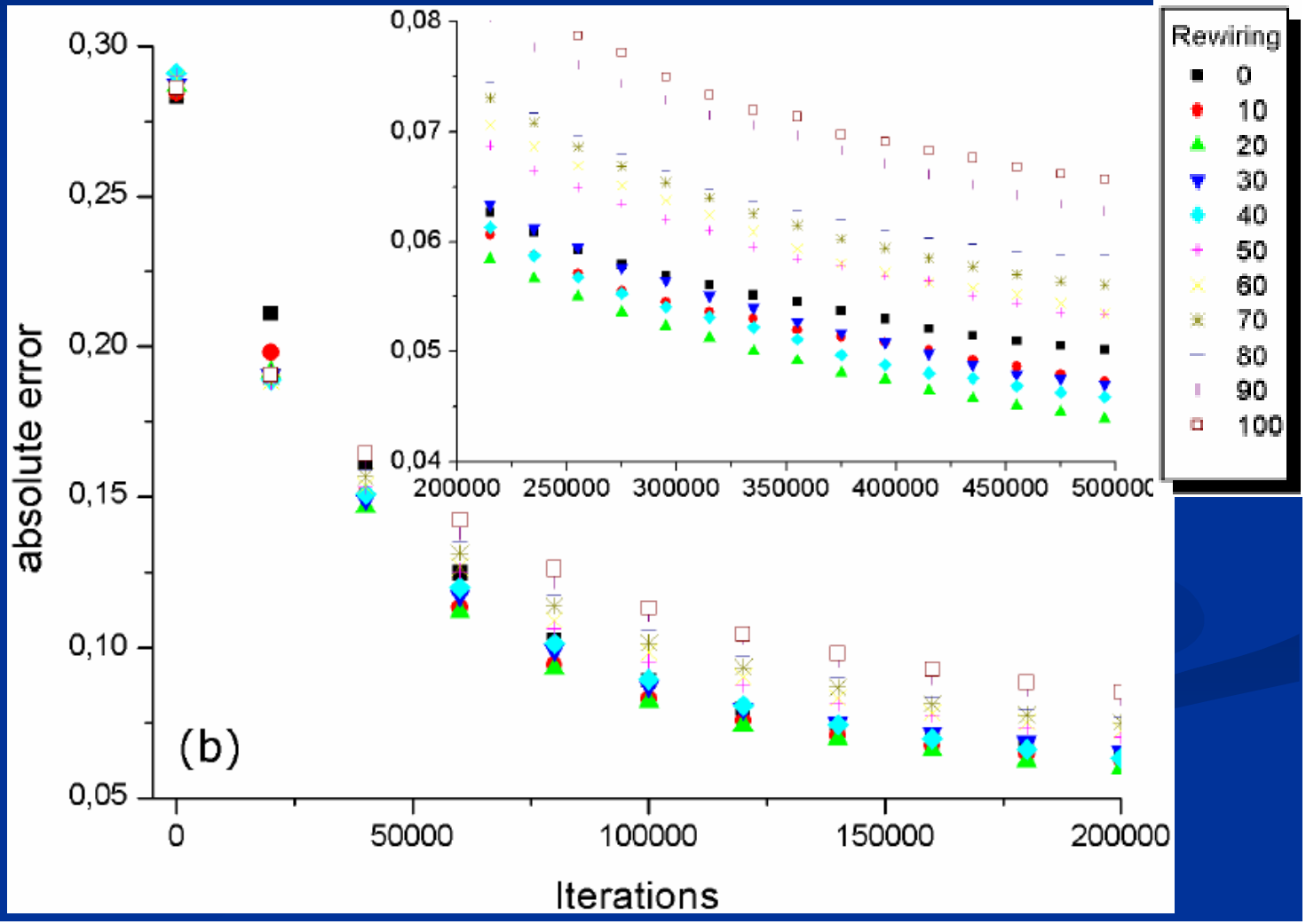
Learning 5 patterns (far from maximal capacity). SWN gives small advantage.



Increasing load of patterns

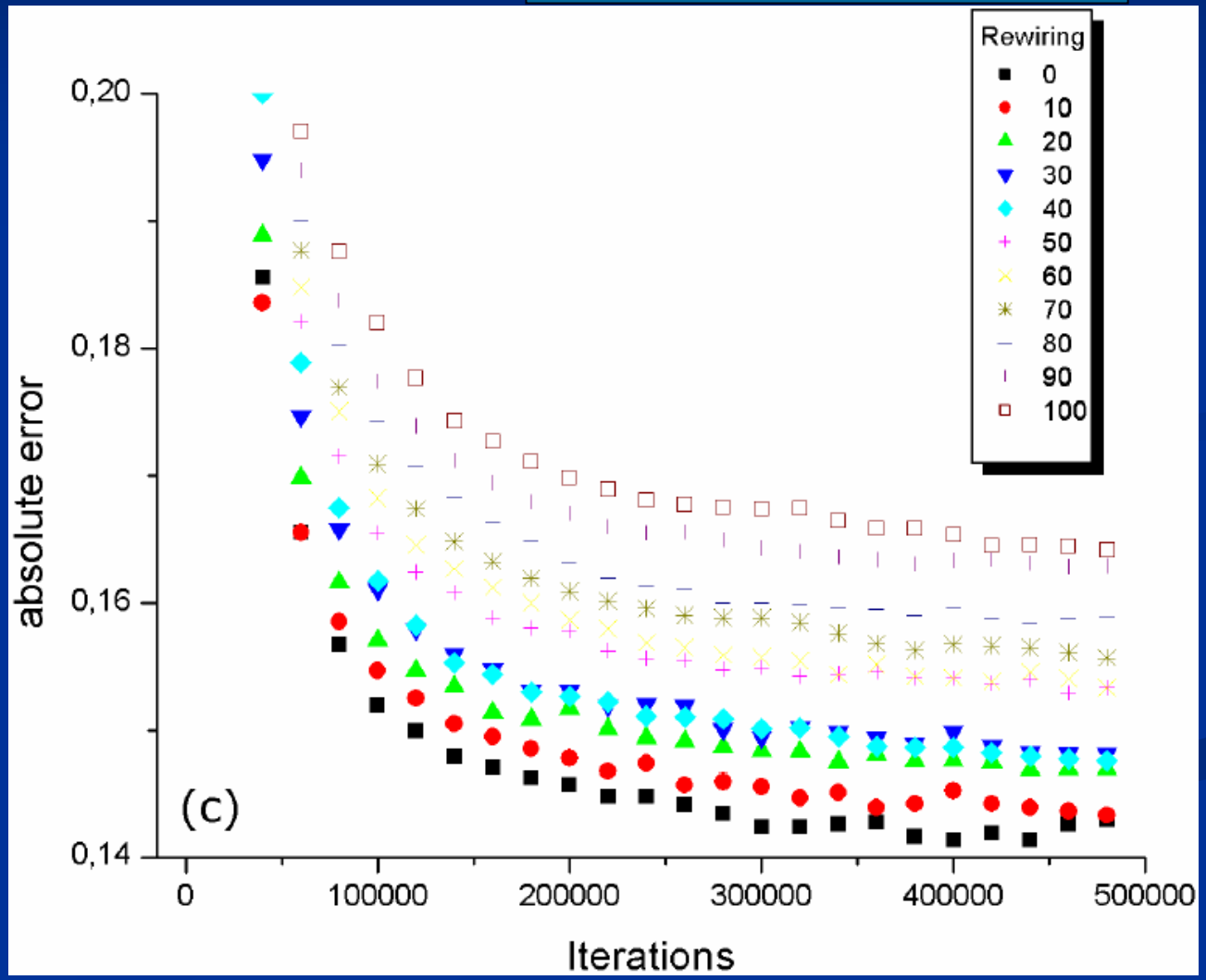
Simulation with : 5 Neurons per layer with 5 layers. Learning 20 patterns.

Learning 20 patterns. Load has not reached max. capacity. SWN learns better and faster than any other architecture.



Increasing load of patterns beyond capacity

Simulation with 5 neurons per layer and 5 layers. Learning 80 patterns.

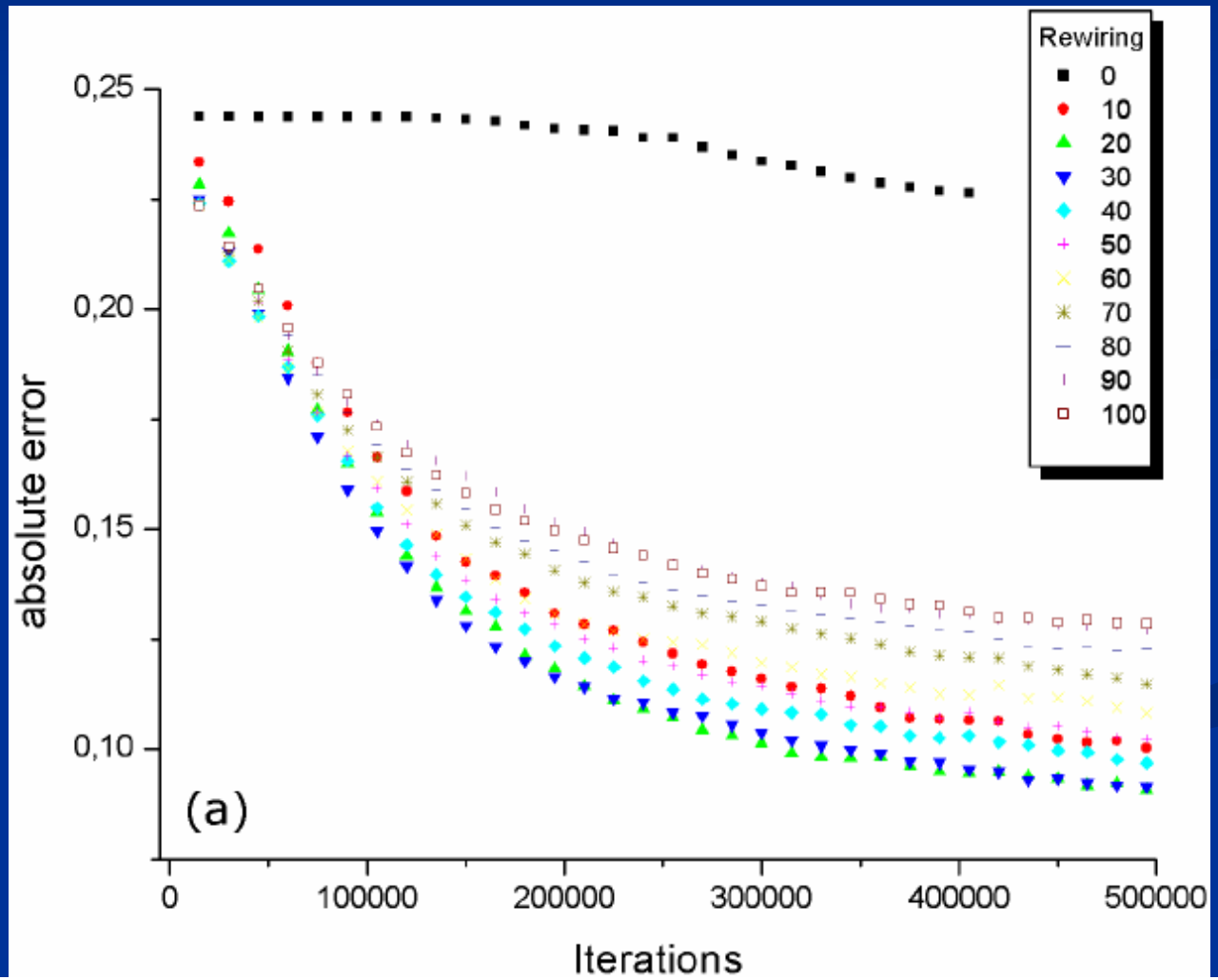


Learning 80 patterns. The average error is high. Here the regular network is optimal.

Network of 5 Layers by 8 Neurons

Simulation with 5 neurons per layers and 8 layers. NN was trained with 40 patterns for 50 different runs.

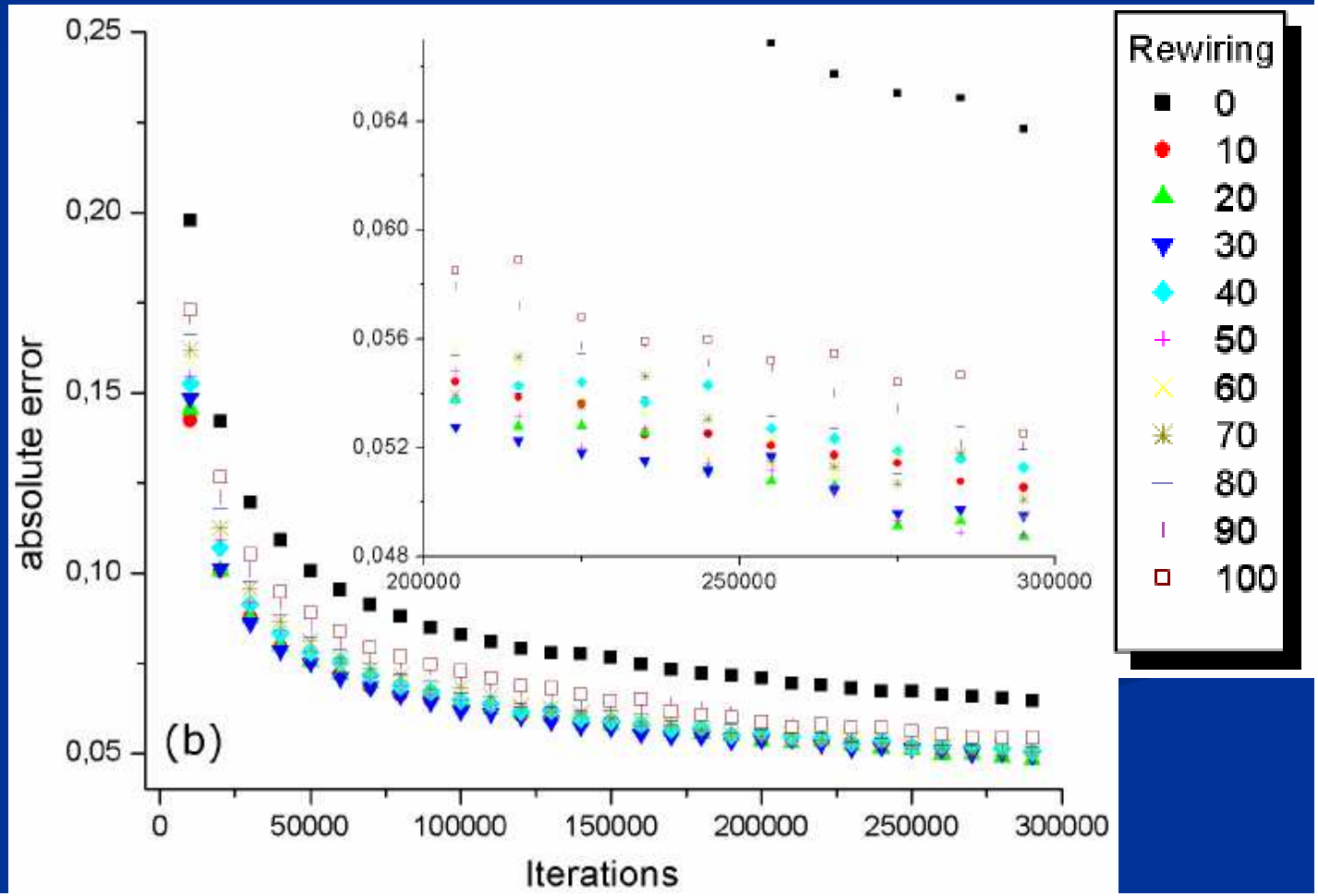
Learning 40 patterns. The regular network almost fails to learn. But with a few short-cuts the network learns well. The best architecture is not random but a Small-World architecture.



Network of 15 Layers by 8 Neurons

This simulation was made with a networks of 15 neurons per layers with 8 layers. It was trained with 100 patterns for 20 different runs.

Learning 100 patterns. More neurons and more patterns than before (and the ratio patterns/Max patterns changed) Observation: Small-World architecture performs better. Random and regular architectures are worse in this case.



Network Connectivity

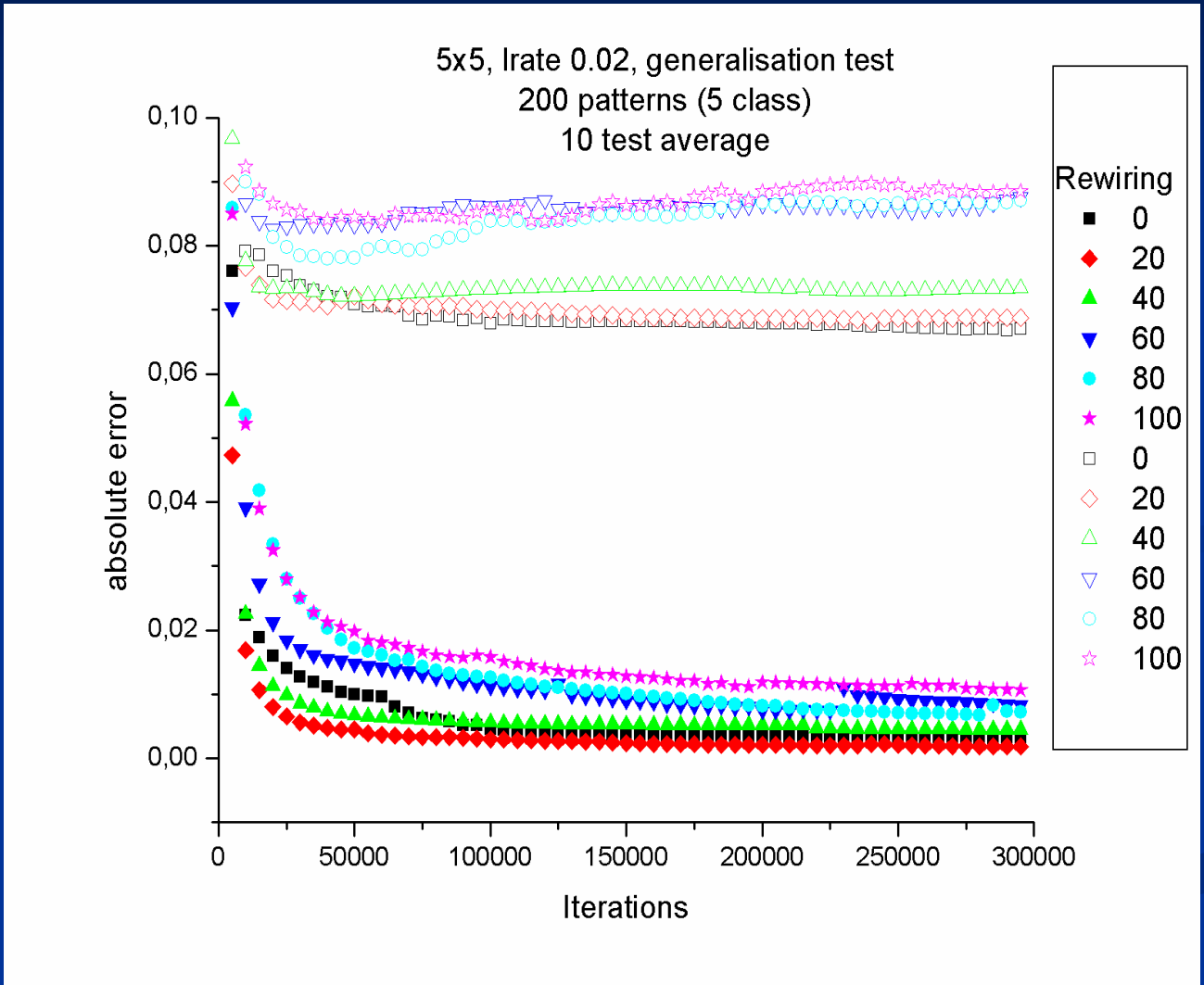
This simulation was made with a networks of 15 neurons per layers with 8 layers. It was trained with 100 patterns for 20 different runs.

Optimal number of short-cuts vs. no of iterations

# of rewiring	5 pat	20 pat	40 pat
0	70000	85000	220000
10	45000	75000	220000
20	30000	75000	160000
30	30000	80000	215000
40	30000	85000	200000
50	25000	95000	335000
60	25000	100000	340000
70	25000	105000	480000
80	25000	115000	-
90	25000	130000	-
100	30000	135000	-
nb of patterns	**5**	**20**	**40**
min (learning time)	25000	75000	160000
min(position)	~70	~20	~20

Test of Generalization

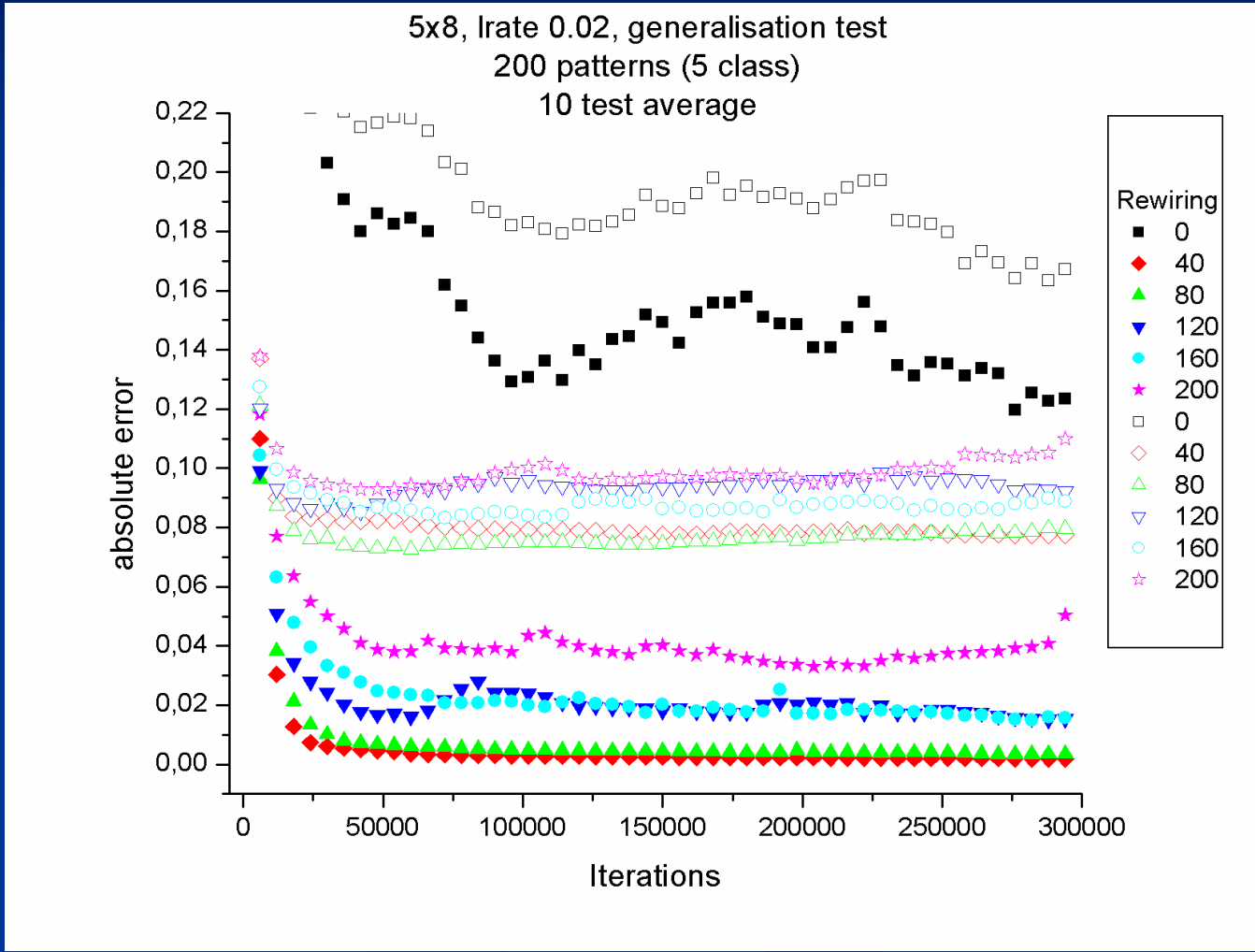
5 neurons by 5 layers network. Task: Classification into 5 classes. Generalization error is best for 0-20 rewirings (few short-cuts).



- Full symbols: Error of trained patterns.
- Empty symbols: Error of generalisation patterns (untrained patterns)

Generalization …

5 neurons by 8 layers network. Task: Classification into 5 classes. Generalization error is best for 40-80 rewirings (few short-cuts). Regular architecture gives much larger error.



- Full symbols: Error of trained patterns.
- Empty symbols: Error of generalisation patterns (untrained patterns)

Summary

- With a small load of patterns (in comparison with the capacity of the NN), the gain of speed and quality of learning is small.
- With too many patterns, the NN needs a maximum of capacity, and the regular architecture has the largest capacity. In this range the average error is high.
- With a number of patterns between those extremes, Small-World networks are the best with respect to errors and speed of learning.

Implications in Neurobiology

- Better understanding the basis of neural learning.
- Insight into complexity in organization of brain vs. functional tasks.
- Fastest learning possible constraint on evolution of nervous system of biological species?
- Assumption of fastest learning: Model predicts

C and L of visual cortex (layered feed-forward):

C = 0.125(8) and L = 3.7(5) [C.elegans: C=0.28, L=2.65]

Future Work

- Artificial Neural Networks, Data Mining: Do our results carry over for learning in NN with very many neurons and patterns?
- How about unsupervised learning (Hebbian learning) on SWN and Scale-Free NN architecture?
- Memory on SWN and Scale-Free NN architecture?