Information and Topology in Attractor Neural Networks

D. Dominguez
david.dominguez@uam.es
K. Koroutchev
k.koroutchev@uam.es
E. Serrano
eduardo.serrano@uam.es
F. B. Rodríguez
F.Rodriguez@uam.es
EPS, Universidad Autónoma de Madrid, 28049 Madrid, Spain

A wide range of networks, including those with small-world topology, can be modeled by the connectivity ratio and randomness of the links. Both learning and attractor abilities of a neural network can be measured by the mutual information (MI) as a function of the load and the overlap between patterns and retrieval states. In this letter, we use MI to search for the optimal topology with regard to the storage and attractor properties of the network in an Amari-Hopfield model. We find that while an optimal storage implies an extremely diluted topology, a large basin of attraction leads to moderate levels of connectivity. This optimal topology is related to the clustering and path length of the network. We also build a diagram for the dynamical phases with random or local initial overlap and show that very diluted networks lose their attractor ability.

1 Introduction

Interest in attractor Amari-Hopfield neural networks, originally dealing with fully connected architectures (Amari, 1972; Hopfield, 1982), has been renewed with the study of more realistic topologies (Masuda & Aihara, 2004; Torres, Muñoz, Marro, & Garrido, 2004). Among them, relatively simple small-world (SW) graphs (Watts & Strogatz, 1998) can capture most facts of a wide range of realistic social, artificial or biological networks (Albert & Barabasi, 2002; Rolls & Treves, 2004). The SW connectivity properties interpolate between regular and a completely random network. The SW graphs are modeled by only two parameters: the average ratio of links per node $\gamma$ and the ratio of long-range random links $\omega$.

The Amari-Hopfield network, as an information processing system, stores information through specific configurations of the network activity (i.e., patterns that can be recovered in response to the stimuli). Of interest is the load $\alpha$, the ratio of patterns per links that the network stores. The most
common measure of the retrieval ability of these networks is the overlap, $m$, between neuron states and the memorized patterns (Amit, Gutfreund, & Sompolinsky, 1985). This measure is plotted as a function of the load $\alpha$ in the top panels of Figure 1 for the classical Amari-Hopfield fully connected (FC, $\gamma = 1$) and random extremely diluted (RED, $\gamma \ll 1, \omega = 1$) networks.

The left panel of Figure 1 show that the FC network has a critical capacity at $\alpha_{FC}^{c} \approx 0.145$: the overlap goes sharply to 0 for loads larger than $\alpha_{c}^{FC}$, defining well-pronounced phase transition with a small amount of error in the retrieval of patterns $m_{c} \approx 0.97$. This result was discussed long ago by Amit et al. (1985) and Crisanti, Amit, and Gutfreund (1986). However, for random extremely diluted networks, the transition of $m$ to zero occurs at $\alpha_{RED}^{c} \approx 0.64$ (Canning & Gardner, 1988), but it is a smooth transition: $m_{c} = 0$ (see the right panel of Figure 1). Therefore, the characterization of the system exclusively in terms of $m$ and $\alpha_{c}$ poses some problems.

Less attention has been paid to another parameter of the network: the mutual information between the stored patterns and the neural states, or its derivate, the information ratio (Okada, 1996; Dominguez & Bolle, 1998). The bottom panels of Figure 1 display the information ratio $i(\alpha)$, evaluated from the conditional probability of neuron states and the patterns. When
the network approaches its saturation limit $\alpha_c$, the neuron states cannot remain close to the patterns because the overlap is usually small. Thus, while the number of patterns increases, the information per stored pattern decreases. Therefore, the information $i(\alpha)$ for all topologies (except the FC) is a nonmonotonic function of the load, which reaches its maximum $i_{\text{max}} \equiv i(\alpha_{\text{max}})$ at some value $\alpha_{\text{max}} \leq \alpha_c$ of the load.

For the FC case, $\alpha_{\text{max}}$ and $\alpha_c$ coincide, and the critical information is about $i_{\text{FC}}^{\text{max}} \sim 0.13$, as can be seen in the left panel of Figure 1. However, for other long-range topologies, as RED, for instance, this does not hold (see the right panel of Figure 1). The RED network has null information at $\alpha_c$. If one looks for the maximum, one finds $i_{\text{RED}}^{\text{max}} \sim 0.22$ for $\alpha_{\text{RED}}^{\text{max}} \sim 0.32$. In conclusion, the parameters $i_{\text{max}}, \alpha_{\text{max}}$ are well defined and are more appropriate than $m_c, \alpha_c$, to evaluate the Hopfield network’s performance on various topologies.

In this letter, we address the problem of searching the topology that maximizes the information capacity of the network. Using the graph framework, we built networks with a connectivity ratio changing from fully connected ($\gamma = 1$) to extremely diluted networks ($\gamma \to 0$) and randomness ranging from purely local ($\omega = 0$) to totally random ($\omega = 1$) connectivity.

We show that $i$ increases (decreases) with $\gamma$ for local (random) networks and remains about the same for SW topologies. A question arises about the optimal topology: If the randomness $\omega$ is fixed (by physical constraints), which is the best connectivity $\gamma$? To our knowledge, no one previously has answered this question. We approach this problem from two scenarios: the stability of the memorized patterns (related to the number of errors induced by crosstalk) and the domain of the attractors (related to the network’s ability to associate an corrupted input pattern with a stored pattern).

The structure of the letter is as follows. In section 2, we define the neural dynamics model and review the information measures used in the calculations. The results are shown in section 3, where we study the stationary states by simulation and theory. In section 4 the attractors’ sensitivity to different initial conditions is discussed. We present a diagram for the phases with local and random initial conditions and show the relation between topology and mutual information. The conclusions are drawn in the final section.

2 The Model

2.1 Topology and Dynamics. The network state at a given time $t$ is defined by a set of $N$ binary neurons, $\mathbf{\sigma}_t = \{\sigma^i_t \in \{\pm 1\}, i = 1, \ldots, N\}$. The network learns a set of independent patterns $\{\mathbf{\xi}_\mu, \mu = 1, \ldots, P\}$. Accordingly, each pattern $\mathbf{\xi}_\mu = \{\xi^i_\mu \in \{\pm 1\}, i = 1, \ldots, N\}$ is a set of site-independent random variables, binary and uniformly distributed: $p(\xi^i_\mu = \pm 1) = 1/2$. The synaptic couplings are $J_{ij} \equiv C_{ij} W_{ij}$, where $C$ denotes the topology matrix and $W$ are the learning weights.
The topology, defined by the matrix $C = \{C_{ij} \in \{0, 1\}\}$, with $C_{ii} = 0$, splits in local and random links. The local links connect each neuron to its $K_l$ nearest neighbors in a closed ring. The random links connect each neuron to $K_r$ others, uniformly distributed along the network. Hence, the network degree is $K = K_l + K_r$. The network topology is then characterized by two parameters: the connectivity ratio and the randomness ratio, defined respectively by

$$\gamma = K/N, \quad \omega = K_r/K,$$

where $\omega$ plays the role of a rewiring probability in the small-world model (SW) (Watts & Strogatz, 1998). The model presented here was proposed by Newman and Watts (1999), which has the advantage of avoiding disconnecting the graph. We named small-world the procedure to build the connectivity graph rather than the properties of a network (large clustering and small path length). Note that the topology can be implemented by an adjacency list connecting neighbors. So the storage cost of this network is given by the cardinality $|J| = N \times K$. Only long-range, $K \propto N$, $N \to \infty$, networks will be considered in this work, whatever $\omega$.

The learning algorithm updates $W$, according to the Hebb rule

$$W_{ij}^\mu = W_{ij}^{\mu-1} + \xi_i^\mu \xi_j^\mu.$$

The network starts at $W_{ij}^0 = 0$, and after $\mu = P$ learning steps, it reaches a value $W_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$. The learning stage has a slow dynamics, which is stationary in the timescale of the much faster retrieval stage.

The neural states, $\sigma_i^t \in \{\pm 1\}$, are updated according to the dynamics

$$\sigma_i^{t+1} = \text{sign}(h_i^t), \quad h_i^t \equiv \sum_j J_{ij} \sigma_j^t, \quad i = 1, \ldots, N,$$

where $h_i^t$ is the local field of neuron $i$ at time $t$. Note that in the case of symmetric synaptic couplings, $J_{ij} = J_{ji}$, an energy function can be defined, whose minima are the stable states of the dynamics, equation 2.3.

### 2.2 Information Measures.

The task of the neural channel is to retrieve a pattern (say, $\vec{\xi} = \vec{\xi}^\mu$) starting from a neuron state $\vec{\sigma}^0$, which is inside its attractor basin. This is achieved through the network dynamics, equation 2.3. The relevant order parameter is the overlap between the neural states and the pattern,

$$m_N^t \equiv \frac{1}{N} \sum_i \xi_i \sigma_i^t,$$

at the time step $t$. 
We consider a mean-field network, and the distribution of the states is assumed to be site independent (Newman, Moore, & Watts, 2000). Therefore, according to the law of large numbers, the overlap can be written, for \( K, N \to \infty \), as \( m' = \langle \sigma^t \xi \rangle_{\sigma, \xi} \). The brackets represent an average over the joint distribution \( p(\sigma, \xi) \), for a single neuron, understood as an ensemble distribution for the neuron states \( \{\sigma_i\} \) and pattern \( \{\xi_i\} \) (Dominguez & Bolle, 1998).

We calculate the mutual information \( \text{MI}[p(\sigma; \xi)] \), a quantity used to measure the prediction that an observer at the output, \( \sigma \), can do about the input, \( \xi \) (we drop the time index \( t \)). We consider solely nonspatially modulated states; hence, \( \text{MI} \) factorizes in sites. It reads

\[ \text{MI} = S[\sigma] - S[\sigma|\xi], \]

where the output entropy, \( S[\sigma] \), and the conditional entropy, \( S[\sigma|\xi] \), depend on the distribution \( p(\sigma, \xi) \). This distribution factorizes in conditional probability \( p(\sigma|\xi) = (1 + m \sigma \xi)/2 \) and input probability \( p(\xi) \). In \( p(\sigma|\xi) \), all types of noise in the retrieval process are enclosed (from both environment and the dynamical process itself). With the above expressions and \( p(\sigma) \equiv \delta(\sigma^2 - 1) \), the entropies read (Bolle, Dominguez, & Amari, 2000):

\[ S[\sigma|\xi] = -\frac{1 + m}{2} \log_2 \frac{1 + m}{2} - \frac{1 - m}{2} \log_2 \frac{1 - m}{2}, \]

\[ S[\sigma] = 1 \text{[bit]} \].

Moreover, together with the overlap, one needs a measure of the load, which is the rate of pattern bits per synapses used to store them. Since the synapses and patterns are independent, the load is given by \( \alpha = |\{\xi^\mu\}|/|J| \equiv P/K \).

We define the information ratio as

\[ i(\alpha, m) \equiv \text{MI}[\tilde{\sigma}; \{\tilde{\xi}^\mu\}]/|J| = \alpha \text{MI}[\sigma; \xi], \]

(2.6)

since for independent neurons and patterns, \( \text{MI}[\tilde{\sigma}; \{\tilde{\xi}^\mu\}] \equiv \sum_{i,\mu} \text{MI}[\sigma_i; \xi_i^\mu] \).

### 3 Stationary States

#### 3.1 Simulation. We deal with different architectures—three regimes of dilution: the fully connected (FC, \( \gamma = 1 \)), moderately connected (MD, \( 0 < \gamma < 1 \)), and extremely diluted (ED, \( \gamma \to 0 \)) symmetric networks. Also the level of randomness runs from purely local (\( \omega = 0 \)) to completely random (\( \omega = 1 \)) networks. We simulate the dynamics with a fixed number of total connections, \( |J| = K \cdot N = 3.6 \cdot 10^7 \), changing the size of the network from \( N = 6 \cdot 10^3 = K \) (\( \gamma = 1 \)) up to \( N = 6 \cdot 10^5, K = 60 \) (\( \gamma = 10^{-4} \)). McGraw & Menzinger (2003) use \( K = 50, N = 5 \cdot 10^3 \) (\( \gamma = 10^{-2} \)), with \( |J| = 25 \cdot 10^4 \), which is far from the asymptotic limit. All runs use smoothing average over a uniform window along the load axis \( P = \alpha \cdot N \).
We study the stability properties of the retrieval states, starting the network from a given pattern, with initial overlap $m^0 = 1$, and we let the neurons update to relaxation. The upper panels of Figure 2 plot the overlap after it converges to a fixed point ($m^*$), for each learned pattern. There is a transition from a retrieval phase ($R$, $m^* > 0$) to a zero phase ($Z$, $m^* = 0$) at a critical $\alpha_c$. For the FC network, $\alpha_{cFC} \sim 0.14$ (Amit et al., 1985; Crisanti et al., 1986). The load $\alpha_c$ for diluted networks depends on $\omega$. For the RED network, $\alpha_{cRED} \sim 0.64$ (Canning & Gardner, 1988), for the local ED network, $\alpha_{cLED} \sim 0.22$, as seen in right-most panel of Figure 2. By increasing the dilution (decreasing $\gamma$), the transition with $\alpha$ becomes smoother, but the $\alpha_c$ increases. On the other hand, we observe that the overlap is always larger for random networks than for local networks, increasing with $\omega$ for all dilution degrees.

Since there is a competition between critical load $\alpha_c$ and sharpness of the transition in $m$, we plot in the bottom panel of Figure 2 the information ratio, $i(\alpha, m)$, for the three architectures. It can be seen that $i(\alpha, m)$ is a non-monotonic function of $\alpha$ that reaches a maximum, $i_{\text{max}}(\gamma, \omega) \equiv i(\alpha_{\text{max}}, \gamma, \omega)$, for a load $\alpha_{\text{max}} \leq \alpha_c$, with $\alpha_{\text{max}} = \alpha_c$ only for the discontinuous transition.
of the FC network. The $i_{\text{max}}$ increases from the left to the right panel of Figure 2 ($\gamma \rightarrow 0$, ED networks) and from local ($\omega = 0$) to random connections ($\omega = 1$).

Overall, the optimal topology with respect to the stability of the retrieval states is the RED network, with $i_{\text{max}} \sim 0.23$, at $\alpha_{\text{max}} \sim 0.34$, and $m_{\text{max}} \sim 0.87$.

### 3.2 Replica Theory

We consider a symmetrical Hebb network with topology described by its connectivity matrix $C$, with parameters according to equation 2.1. The static thermodynamics of a symmetric Hebb neural network in the thermodynamic limit ($N \rightarrow \infty$) is exactly solvable for arbitrary topology, provided that each neuron is equally connected to the others and that a site-independent solution for the fixed point exists. This is the case for SW topology (Watts & Strogatz, 1998), which we will study in detail here. The energy of the symmetric network in the state $\bar{\sigma} \equiv \{\sigma_i\}$ is

$$H = -\frac{1}{2N\gamma} \sum_{i,j=1}^{N} C_{ij} W_{ij} \sigma_i \sigma_j,$$

with $W_{ij}$ given by the Hebb rule, equation 2.2. $H$ completely defines the equilibrium thermodynamics of the system.

We are looking only for spatially independent solutions; therefore, the order parameters are independent of the indexes. Following the literature (Amit et al., 1985; Canning & Gardner, 1988), we define the order parameters as

$$m^\rho = \langle \langle \xi_i | \sigma_i^\rho \rangle \rangle_{\xi},$$

$$q^{\rho \tau} = \langle \langle \sigma_i^\rho | \sigma_i^\tau \rangle \rangle_{\xi}, \quad \tau \neq \rho,$$

$$r^{\rho \tau} = \gamma \alpha^{-1} \left\langle \left( \sum_{\nu} m_i^\rho m_i^\tau \right) \right\rangle, \quad \tau \neq \rho,$$

where $\langle \langle \rangle \rangle_{\xi}$ is a spatial average, $\langle \langle \rangle \rangle_{\xi}$ is the pattern average, and $\langle \rangle$ is the thermal average. The $\xi$ is the only pattern with macroscopic overlap; the index $\nu$ runs over the rest of the patterns; $q$ and $r$ are the configuration’s correlation between the replicas of $\langle \sigma \rangle$ and $m^\nu$, respectively.

Further, we calculate the free energy and get an equation, analogous to equation 3.2 in Canning and Gardner (1988), the extremum in $m, r, q$ of:

$$f = \frac{\gamma}{2} \alpha + \frac{\gamma}{2} m^2 \text{Tr} C^{-1} + \frac{\gamma}{2} \alpha \beta r (1 - q)$$

$$+ \frac{\alpha}{2\beta} \left\{ \text{Tr} \ln (1 - \beta (1 - q) C / N) \right\}.$$
\[
- \text{Tr} \left[ \beta q \left( 1 - \beta (1 - q) C \right)^{-1} C/N \right] \\
- \frac{1}{\gamma \beta} \int Dz \left\langle \left\langle \ln 2 \cosh \gamma \beta \left( \sqrt{\alpha r z + m \xi} \right) \right\rangle \right\rangle, \tag{3.5}
\]

where \( \beta = 1/T \) is the inverse temperature, \( 1 \) is the unity \( N \times N \) matrix, and \( \int Dz \) means gaussian integral. We denote the trace of a matrix as \( \text{Tr} \ M = \sum_i M_{ii} \) and not the trace over the configurations of the system.

All terms of the free energy scale well with \( N \to \infty \). The only problem is in the term \( \text{Tr} C^{-1} \), for which the sum needs to provide \( 1/\gamma \) factor. Let us note that for the SW connectivity considered here, the degree of each node is \( \gamma N + O(\sqrt{N}) \), the major eigenvalue of \( C \) is also \( \gamma N + O(\sqrt{N}) \), and each eigenvector component is \( v_{0i} = 1/\sqrt{N} + O(1/N) \).

The term in question can be calculated decomposing \( C^{-1} \) by its eigenvalues,

\[ C^{-1} = \sum_k \bar{v}_k \dagger \lambda_k^{-1} \bar{v}_k, \]

where \( \lambda_k \) is the \( k \)th largest eigenvalue and \( \bar{v}_k \dagger \) its corresponding transposed eigenvector. Note that \( \lambda_0 = \gamma N + O(\sqrt{N}) \). The approximation terms will vanish in the limit \( N \to \infty \).

Changing the SW rewiring scheme, one can achieve results without approximation terms, but to match the computer simulations, we do not do that. We have:

\[
\sum_{ij} (C^{-1})_{ij} = N \bar{v}_0 \cdot C^{-1} \cdot \bar{v}_0 \dagger = N \bar{v}_0 \cdot \sum_k \bar{v}_k \dagger \lambda_k^{-1} \bar{v}_k \cdot \bar{v}_0 \dagger = 1/\gamma.
\]

From this point on, the derivation of the equations for the order parameters is straightforward, and in the thermodynamic limit (\( N \to \infty \)), one obtains the same result as Canning and Gardner (1988):

\[
m = \left\langle \left\langle \int Dz \xi \tanh \beta \gamma (\sqrt{\alpha r z + m \xi}) \right\rangle \right\rangle, \tag{3.6}
\]
\[
q = \left\langle \left\langle \int Dz \tanh^2 \beta \gamma (\sqrt{\alpha r z + m \xi}) \right\rangle \right\rangle, \tag{3.7}
\]
\[
r = \text{Tr} \left[ \frac{q}{\gamma} \left( 1 - \beta (1 - q) C \right)^{-2} \cdot \left( \frac{C}{N} \right)^2 \right]. \tag{3.8}
\]

Having spatial-dependent connectivity, it is unclear whether the site fluctuations will not provide solution with local macroscopic dependence of
$m$, for example, spatially dependent solutions. However, when the neurons are unbiased, $\sigma \in \{\pm 1\}$, with zero threshold, only spatially symmetric states can be observed in the thermodynamic limit in the case of finite $\gamma > 0$, as can be seen in Koroutchev and Korutcheva (2005).

At $T \to 0$, keeping the quantity $G \equiv \gamma \beta (1 - q)$ finite, in equation 3.6, the $\tanh(\cdot)$ converge to $\text{sign}(\cdot)$, in the next equation $\tanh^2(\cdot)$ behaves as $1 - \delta(\cdot)$, and in equation 3.8, the expression can be expanded in series of $G$, giving:

$$m = \text{erf}(m/\sqrt{2r\alpha}) \quad (3.9)$$
$$G = \sqrt{2/(\pi r\alpha)} e^{-m^2/(2r\alpha)} \quad (3.10)$$
$$r = \sum_{k=0}^{\infty} (k + 1)a_k G^k, \quad (3.11)$$

where $a_k \equiv \gamma \text{Tr}[(C/K)^{k+2}]$. Note that $a_k$ is the probability of the existence of (eventually self-crossing) cycles of length $k + 2$ in the connectivity graph. According to the definition, $a_0 = 1$, because $C$ is symmetric, and $a_1$ is the probability of having 3-cycle, that is, the clusterization index of the graph. The series converge, because $G \in [0, 1)$ and $a_k \in [0, 1]$. For random extremely diluted and fully connected networks, one recovers the known results for $r_{\text{RED}} = 1$ and $r_{\text{FC}} = 1/(1 - G)^2$, respectively (Bolle, Jongen, & Shim, 1999).

The only equation explicitly dependent on the topology of the system is equation 3.11, and the only important topology features for the network equilibrium are $\{a_k\}$. As $k \to \infty$, $a_k$ tends to $\gamma$. Cyclic evaluation of the Ising model was studied by Parisi (1988). It was observed that compared with the simulations, with finite $N$, equations 3.9 to 3.11 overestimate the role of long loops. In order to close the system of these equations, one must estimate the $\{a_k\}$, which are topology dependent. This is done in the appendix.

We plot with lines in Figure 3 $i_{\text{max}}$ as a function of the dilution $\gamma$ for different values of the randomness $\omega$, obtained from theoretical equations 3.9 to 3.11. The results from the simulation for $m^r=0 = 1$ are represented in the Figure 3 with symbols. A comparison of theoretical predictions and the simulations yields, in the experimental range we study, no significant deviation from the replica symmetric solution. Therefore, no replica symmetry breaking is expected. Figure 3 shows that both theory and simulation results agree for most $\omega > 0$, with some fluctuations for $\omega \sim 0$.

4 Attractors

4.1 Sensitivity to Initial Overlap. The theoretical equations for the stationary states, equations 3.9 to 3.11, account only for the existence of the retrieval (R) solution $m > 0$. They say nothing about its stability. The zero
states \( (Z), m = 0, \) are also a solution of these equations, so both \( R \) and \( Z \) may coexist in some region of topological parameters \( \gamma, \omega. \) In order to study the stability of the attractors, we simulate equation 2.3, and see how the network behaves under different initial conditions. It is worth noting that no theory for the dynamics of our topological model is yet known (Bolle et al., 1999).

To analyze the attractor properties of the retrieval, we made the neuron states start at a configuration \( \vec{\sigma}^0 \) weakly correlated with a learned pattern, \( \vec{\xi} \equiv \xi^\mu. \) If they are inside its basin of attraction, \( \vec{\sigma}^0 \in B(\vec{\xi}), \) the pattern will be recovered, \( \vec{\sigma}^t \rightarrow \vec{\xi}. \) First, we choose an initial configuration given by a random correlation with the learned pattern, \( p(\sigma^0 = \pm \xi | \xi) = (1 \pm m^0)/2, \) for all neurons (so we avoid a bias between local and random neighbors). We call this the \( m_R \) initial overlap. The retrieval dynamics starts with an overlap \( m^0 = 0.1 \) and stops after it converges to a fixed point \( m^*. \) Usually \( t_f = 80 \) sequential updates are a large enough delay for retrieval. The information \( i(\alpha, m; \gamma, \omega) \) is calculated according to equations 2.5 and 2.6 for \( m^*. \)

The results are depicted in the bottom panel of Figure 4, where \( i(\alpha) \) is plotted for different values of \( \gamma, \omega. \) Unlike the theoretical results for the storage capacity, there are MD topologies that perform best whenever the attractor properties are considered. We define the optimal topology, for a
fixed $\omega$, as the value of $\gamma$ that yields the most information. Starting with $m^0 = 0.1$ yields optima $i(\gamma_{opt}, \omega)$ for moderate dilutions; for instance, with $\omega = 0.1$, it holds $\gamma_{opt} \sim 10^{-2}$.

Next, we compare $m_R$ with another type of initial distribution. The neurons start with local correlations: $\sigma_i^0 = \xi_i$, $i = 1, \ldots, (Nm^0)$, and random $\sigma_i^0 = \pm 1$ otherwise. We call it the $m_L$ initial overlap. The results are shown in the top panel of Figure 4. The first observation is that the maximum information $i_{max}(\gamma; \omega)$ has the same nonmonotonic behavior with $\gamma$ in the region of $\omega$ considered; there is still a moderate $\gamma_{opt}$ for which the information $i(\gamma_{opt})$ is optimized. However, the $i_{max}$ for the $m_L$ are a little smaller than for the $m_R$ overlap, and the optima shift a bit to the left panels, where the less diluted topologies are. For instance, with $\omega = 0.1$, now it holds $\gamma_{opt} \sim 10^{-1}$.

The comparison between the upper ($m_L$) and lower ($m_R$) panels of Figure 4 shows that the initial $m_R$ allows for an easier retrieval for any $\omega$. Local topologies ($\omega = 0$) are especially sensitive to the type of initial overlap, losing their retrieval abilities for $m_L$ if the connectivity is $\gamma_{opt} \leq 10^{-2}$. This sensitivity to the initial conditions can be understood in terms of the basins of attraction. Random topologies have very deep attractors, especially if the network is diluted enough, while regular topologies almost lose their

Figure 4: Mutual information versus $\alpha$ for the local (upper) and random (lower) initial overlap $m^0 = 0.1$, with $\omega = 0.0, 0.1, 0.2$. 

$\gamma=1.00$ $\gamma=10^{-1}$ $\gamma=10^{-2}$ $\gamma=10^{-3}$ $\gamma=10^{-4}$

$t=80s; |J|=40M; Sim; m_L \times m_R=0.1$
retrieval abilities with dilution. However, since the basins become rougher with dilution, the network takes longer to reach the attractor and can be trapped in meta-stable states. Hence, the competition between depth and roughness is won by the more robust MD networks.

The retrieval capability of the network starting at condition \( m_R \) or \( m_L \) is plotted in Figure 5. We represent as \( R \) the phase where the retrieval reaches the information \( i_{\text{max}} \geq 0.05 \), starting from \( m^0 = 0.2 \), with \( m_R \). The phase \( L \) is the same, but starting with \( m_L \). Efficient retrieval is not allowed with the \( m_L \) condition for very connected or local topologies. On the other hand, only local diluted topologies exclude retrieval with the \( m_R \) initial overlap, shown below the dashed curve in Figure 5.

Next we study the role of the symmetry constraints in the synaptic weights. By asymmetric topology, we mean that no condition \( C_{ij} = C_{ji} \) is assumed, and the local connections are chosen for only one direction in the network ring. The maximum of information against the dilution, \( i_{\text{max}}(\gamma) \), is plotted in Figure 6, for several degrees of randomness \( 0 \leq \omega \leq 1 \). The initial overlap is the random condition \( m_R \), with \( m^0 = 0.1 \). The left panel shows the results for the symmetric network and the right panel the results for the asymmetric network. We observe that the nonmonotonic behavior of \( i_{\text{max}}(\gamma) \) still holds for the asymmetric network, only for \( \omega \leq 0.2 \). The asymmetric topology has more robust attractors than the symmetric topology for strong dilution. The symmetric topology keeps its optimal moderate dilution, \( 0 < \gamma < 1 \), up to \( \omega \sim 0.4 \). Both types of symmetry have the same asymptotic behavior with extreme dilution, for random topologies, saturating the information at \( i_{\text{max}}(\gamma \to 0) \sim 0.22 \). Nevertheless, the

Figure 5: Diagram \((\omega \times \gamma)\) with the phases \( R \) and \( L \), for initial overlap \( m^0 = 0.2 \).
nonmonotonic behavior is qualitatively similar in spite of the rather different topologies.

4.2 Clustering and Mean-Length Path. We describe here the topological features of the network as a function of its parameters: the clustering coefficient, $c$ (the average number of neighbors of each node that are neighbors between them), and the mean-length path between neurons, $l$ (the average distance between nodes measured as the minimal path length between them). When $\gamma$ is large, the net has $c$ large, $c \sim 1$ and $l$ small, $l = O(1)$, whatever $\omega$ used. When $\gamma$ is small, then if $\omega \sim 0$, the net is clustered and has large paths: $c = O(\gamma)$, $l \sim N/K$. If $\omega \sim 1$, the net becomes random: $c \ll 1$ and $l \sim \ln N$. However, if the randomness is about $\omega \sim 0.1$, then $c = O(\gamma)$, but with $l \sim \ln N$, and the network behaves as an SW: clustered but with short paths.

The dependence of $c$, $l$, and $i$ on randomness $\omega$ is plotted in Figure 7. In all panels, for connectivity $\gamma = 10^{-1}, 10^{-2}, 10^{-3}$, we see a decrease of $c$ and $l$ and an increase of $i$ with $\omega$. For these ranges of $\omega$, the path length $l$ has already
Figure 7: The maximal information $i(\alpha_{\text{max}})$ (bottom), clustering coefficient $c$, and mean-path length $l$ (top) versus $\omega$, for $\gamma = 10^{-1}, 10^{-2}, 10^{-3}$ (from left to right). Simulations with $N.K = 10^M, m^0 = 0.1$.

decreased to small values, and the networks have entered the SW region. However, in the right panel, there is still a slowdown of $l$. The clustering $c$ decreases quickly around $\omega = 0.2$, after which the network is random-like.

The comparison of the behavior of $c$ (top panels) with $i$ (bottom panels) is in agreement with Kim (2004), which allows us to conclude that the performance of the Hopfield model on various networks can be enhanced by decreasing the clustering coefficient.

This region $0.001 \leq \omega \leq 0.2$ is the SW regime for the network we study. Looking at the bottom panel, we see that at the end of the SW graph, there is a fast increase of $i$ between $0.05 \leq \omega \leq 0.20$. We conjecture that after the SW region, a further increase of the randomness $\omega$ is not worth its wiring cost to gain a little extra information $i$.

5 Conclusions

We have discussed the information capacity of an attractor neural network as a function of its topology. We calculated the mutual information for an Amari-Hopfield model, with Hebbian learning, varying the connectivity ($\gamma$) and randomness ($\omega$) parameters, and obtained the maximum respect to $\alpha$, 

$\omega = K_r/K, \gamma = K/N$

$T = 0; m_0 = 0.1$

$\gamma = 10^{-1}$

$\omega = N_x K = 10 \times 10^6$

$\gamma = 10^{-1}$

$\omega = 10^{-2}$

$\omega = 10^{-3}$
$i_{\text{max}}(\gamma, \omega) \equiv i(\alpha_{\text{max}}; \gamma, \omega)$. The information $i_{\text{max}}$ always increases with $\omega$, but for a fixed $\omega$, an optimal topology $\gamma_{\text{opt}}$, in the sense of the information, $i_{\text{opt}} \equiv i_{\text{max}}(\gamma_{\text{opt}}, \omega)$, can be found. We studied the stability and attractor properties.

From the point of view of stability, the calculations show that the optimal topology regarding the storage is the random extremely diluted (RED) network. Indeed, if no pattern completion is required, the patterns can be stored statically, which is achieved with a vanishing connectivity. As long as the dynamics is concerned (starting from a noisy pattern, $m^0 < 1$), however, the latter is not true: we found there is an moderate $\gamma_{\text{opt}}$, which optimizes the retrieval performance, whenever $0 \leq \omega < 0.3$ (see Figure 6). Moreover, the memory behavior of local diluted networks is even more damaged if they start with local overlap than if the initial overlap is random. This can be understood regarding the shape of the attractors. The ED network waits much longer for the retrieval than more connected networks do, so the neurons can eventually be trapped in spurious states with vanishing information.

We found a relation between the fast increase of information with $\omega$ and the SW region of the topology. This implies that it is worth increasing the wiring length just to the end of the SW zone. In both nature and technological approaches to neural devices, dynamics is an essential issue for information processing. So an optimized topology holds for any practical purpose, even if no attention is paid to wiring or other costs of the random links (Morelli, Abramson, & Kuperman, 2004). We conjecture that the reason for the intermediate optimal $\gamma_{\text{opt}}$ is a competition between the broadness (larger storage capacity) and roughness (slower retrieval speed) of the attraction basins.

We believe that the maximization of the information with respect to the topology could be a biological criterion (where nonequilibrium phenomena are relevant) for building real neural networks. We expect that the same dependence can be found for more structured networks and learning rules. More complex initial conditions may also play a role in the retrieval and deserve dedicated study.

Appendix: The Probability of Cycles

The small-world (SW) topology can be defined by the probability of having the nodes $i$ and $j$ connected:

$$P(C_{ij} = 1) \equiv \omega \gamma + (1 - \omega) \Theta\left[\gamma N - \left(i - j + N + \frac{\gamma}{2} N\right) \mod N\right],$$

(A.1)

where $\omega$ is the SW parameter defined as the rewiring probability in Watts and Strogatz (1998) and $\Theta$ is the unit step function.

One can estimate $a_k$ by generating random SW graphs with a given size, performing $k + 1$ random walk steps starting from some arbitrarily chosen origin node and then counting the trail as successful if there is a link from
that node and as failing otherwise. Using this Monte Carlo procedure, we have a Bernoulli process for estimating $a_k$.

However, this estimation will suffer finite-scale effects that are difficult to estimate, especially in the case of small $\gamma$. Therefore, first, one can try to make the limit $N \to \infty$ in equation A.1 and then estimate the probabilities $a_k$. This continuous topology estimation will not suffer finite-scale effects, with the exception of the larger error in estimating $a_k$ when $\gamma$ is small.

However, the topology in the continuous case is not exactly the same as in the discrete case (the probability of repeating the movement along any edge in the graph is exactly zero, and the symmetry is loosely defined in the continuous approximation), so the question of whether the discrete estimation with $N \to \infty$ and the continuous estimation coincide is legitimate. We have checked the discrepancies with simulations, and when $\gamma N > 30$, there is a very good match between the continuous and the discrete case. Similar questions are discussed by Newman et al. (2000).

The conclusion is that the coefficients $a_k$, $k > 0$ can be calculated using a modification of the Monte Carlo method originally proposed by Canning and Gardner (1988). Let us consider $k$ as a number of iterations and regard a particle that at each iteration changes its position among the network connectivity graph with probabilities defined by the topology of the (SW) graph (see Figure 8). If the particle at step $k$ is in node $j(k)$, we say that the coordinate of the particle is $x(k) \equiv j(k)/N$. When $N \to \infty$, $x(k)$ is a continuous variable. If $y$ is a uniformly distributed random variable between $-1/2$ and $1/2$, then in step $k + 1$, the coordinate of the particle changes according to the rule:

$$
\begin{align*}
  x(k + 1) &= [x(k) + \gamma y] \text{mod } 1 \text{ with probability } 1 - \omega, \\
  x(k + 1) &= [x(k) + 1y] \text{mod } 1 \text{ with probability } \omega.
\end{align*}
$$

If the particle starts at moment $k = 0$ from $x(0) = 0$ and is located in position $x(k)$ at step $k$, then the probability of having a loop of length $k + 1$ according to the SW topology is

$$
  a_{k-1} = (1 - \omega) \theta(\gamma/2 - |x(k)|) + \omega \gamma,
$$

where it has been assumed that $x(k) \in [-1/2, 1/2]$. From this Bernoulli process one can estimate $a_k$. (See the precision’s estimation at Koroutchev, Dominguez, Serrano, & Rodriguez, 2004.)

Alternatively to the Bernoulli process, one can get an expression to estimate $a_k$, $k > 0$. Suppose that we have unity length of the graph; then the
Figure 8: Calculus of $a_k$ by random walk in SW topology. The network with its connection is represented by the gray circle. The height of the surrounding circle represents the probability of connection of the neuron at position 0. Movements to distances shorter (larger) than $\gamma/2$ are represented by bold (thin) lines. After five steps, the probability of the existence of a cycle with length 6 ($a_4$) is given by the probability of a connection between $x_5$ and $x_0$ (dotted line).

The probability of having a loop is

$$a_k = \gamma \sum_{R=-k}^{k} \int \prod_{i=1}^{k+2} \frac{dx_i}{\gamma} p(x_i) \left[ \delta \left( \sum x_i - R \right) \right], \quad (A.4)$$

where $x_i$ represents the movement in step $i$, $p(x)$ is the probability of a link at displacement $x$, $x \in [-1/2, 1/2]$, and $R$ is the number of revolution performed by the loop. Although this integral can be solved analytically for small $k$, the calculation of $a_k$ using the Monte Carlo method seems more efficient.

The calculation of $a_k = a_k(C)$ closes the system of equations 3.9 to 3.11, giving the possibility of solving them with respect of $(G, m, r)$ for every $(\omega, \gamma, \alpha)$ values.

**Acknowledgments**

This work was supported by grants TIC01-572, TIN2004-07676-C01-01, BFI2003-07276, TIN2004-04363-C03-03 from MCyT, Spain. D.D. thanks a Ramon y Cajal grant from MCyT.

**References**


Received February 3, 2006; accepted June 22, 2006.