doi:10.20944/preprints201808.0251.v1

Peer-reviewed version available at *Entropy* **2018**, *20*, 824; doi:10.3390/e20110824

Article

The capacity for correlated semantic memories in the cortex

Vezha Boboeva [1], Romain Brasselet [1] and Alessandro Treves [1,2]*

[1] SISSA - International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, Italy
[2] Kavli Institute for Systems Neuroscience/Centre for Neural Computation, Norwegian University of Science and Technology, Trondheim, Norway
* Correspondence: ale@sissa.it

Abstract: A statistical analysis of semantic memory should reflect the complex, multifactorial structure of the relations among its items. Still, a dominant paradigm in the study of semantic memory has been the idea that the mental representation of concepts is structured along a simple branching tree spanned by superordinate and subordinate categories. We propose a generative model of item representation with correlations that overcomes the limitations of a tree structure. The items are generated through "factors" that represent semantic features or real-world attributes. The correlation between items has its source in the extent to which items share such factors and the strength of such factors: if many factors are balanced, correlations are overall low; whereas if a few factors dominate, they become strong. Our model allows for correlations that are neither trivial nor hierarchical, but may reproduce the general spectrum of correlations present in a data-set of nouns. We provide an estimate of the number of concepts that can be stored and retrieved by a large-scale cortical network, the Potts network, which is perhaps approximately 10^7 with human cortical parameters. When this storage capacity is exceeded, however, retrieval fails completely only for balanced factors; above a critical degree of imbalance, a phase transition leads to a regime where the network still extracts considerable information about the cued item, even if not recovering its detailed representation: partial categorization seems to emerge spontaneously as a consequence of the dominance of particular factors, rather than being imposed ad hoc. We argue this to be a relevant model of semantic memory resilience in Tulving's remember/know paradigms.

Keywords: Potts network; attractor neural networks; auto-associative memory; cortex; semantic memory

1. Introduction

One of the most fascinating aspects of the human brain is its ability to ascribe significance to and recognize meaning in objects and events, and to more generally make sense of the world. Semantic memory, comprising our acquired knowledge about the world, can be imagined to reflect, in its statistical structure, the complex, distributed, policentric structure of the neocortex where it resides. In contrast, the relatively much simpler network structure of the hippocampus, in particular of its CA3 field, where episodic memories have long been thought to be at least initially represented by unique patterns of neural activity, may lead to the limited set of outcomes of episodic memory retrieval: either the pattern is retrieved, or not. In the first case, retrieval, subjects *remember* what happened in the episode, in the second they do not, although they may still *know* many of the elements in the episode, likely as they reconstruct them with input from semantic memory. This is the basis for *remember/know* paradigms [1] that assess hippocampal contribution to memory retrieval. But how can the statistical structure of the memory representations themselves be characterized?

1.1. Correlations

In the case of episodic representations in the hippocampus, one straightforward hypothesis about their statistics is that they are largely uncorrelated: each representation is set up, e.g. in CA3, independently of other representations already stored there, under the influence of the Dentate Gyrus [2]. Then the representations are roughly at the same distance from each other in activity space, i.e. they are *ametric*: relations of being closer or farther away, or in the middle between another pair, lose their meaning. This may seem at odds with the best studied neural representations in the hippocampus, spatial representations in rodents, which reflect the continuity of space, where being close or distant is clearly defined. As soon as we move, however, from the representation of different locations in the same restricted spatial context to the representation of different contexts, the phenomenon of global remapping suggests that the notion of ametric representations is relevant. Indeed, it has been observed that even very similar spatial contexts are represented in rat CA3 by completely different, essentially uncorrelated representations [3]. Correspondingly, a measure of metric content has been shown to "increase" in human subjects who can rely *less*, due to incipient Alzheimer, on their hippocampal representations [4]. If hippocampal representations can be said to be ametric, what is the nature of the metricity observed, by contrast, in semantic representations in the neocortex?

Direct access to individual semantic representations through single unit recordings is of course very limited, and not just in the human brain, because of their very distributed nature. Multi-voxel pattern analyses from fMRI are consistent with a complex web of correlations [5,6], but their resolution is limited and so is the characterization of the statistical properties of those correlations. A simple alternative, however, is to assess the nature of the correlations among the semantic items themselves, rather than probing their representations in the brain. This can be done by utilizing any of a number of databases, where a set of semantic items have been described in terms of the features or attributes people associate with them. As a simple toy example, we took the $p = 60$ nouns used in a recent fMRI study [7]. We computed the pairwise correlation between these nouns, as measured by a set of intermediate or surrogate features, such as the co-occurrence with a set of verbs within a sentence in the corpus.

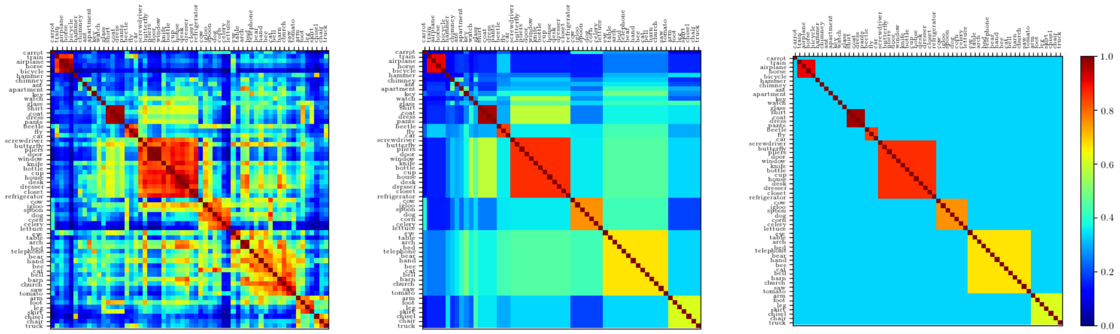
In Fig. 1(a), we report the correlation matrix of the nouns, from which it can be seen that the pattern of correlations cannot be described by any simple schema. One way of thinking about this organization, that has remained ubiquitous in the semantic literature, is to think of concepts organized in a hierarchy, or a tree [8,9]. Such models, in their descriptive and generative formulations, are dramatic oversimplifications that ignore important features of the data, such as the prevalence of concepts intermediate between other concepts (see Sect. 2.2.1). As an example, in Fig. 1(c), we report the correlation that an extreme hierarchical model would see in such data. It is apparent, from the comparison with Fig. 1(a), that this hierarchical model fails to represent off-block values with high correlation.

A less dramatic simplification consists in considering that the correlations between individual concepts belonging to distinct "clusters" can be well approximated by the mean correlation between clusters. Such a simplification yields Fig. 1(b). To quantify the validity of this simplification, one can measure to what extent the distance relations between the concepts match the fully hierarchical limit case [10]. This index, called the ultrametric content (see App. C) can be computed once correlations are translated into a measure of distance. The "soft" hierarchical structure of the matrix in Fig. 1(b) yields an ultrametric content index of 0.61, to be compared with the value 0.5 for the original data. The fully hierarchical matrix, Fig. 1(a), has an ultrametric content of 1. On the other hand, ametric, independently generated representations, as observed in CA3, have an ultrametric content close to 0. Semantic relations, we can conclude, are complex, and the ultrametric content of 0.5 is as far from the purely ultrametric as from the trivial ametric limit.

However, it is the statistical independence of memory patterns that had made available most of the mathematically sophisticated analyses. While these analyses have been successfully used to describe the CA3 circuit, it does not seem like they can be applied to semantic memory, which has in the shared structure between memories its raison d'être. Still, some progress in this direction has been made. In exploring variants of the Hopfield model, which initially featured uncorrelated patterns, the challenge of storing correlated patterns was eventually addressed. One of the earliest attempts to introduce correlations was through an algorithm that arranged patterns on a tree [11,12], in which the upper nodes correspond to classes of items and the lower nodes, each branching from a single upper node, correspond to exemplars of a class. Subsequently it was found [13] that beyond the storage capacity, initial states highly correlated with one individual pattern evolve to the corresponding class, while even the class categorization is lost at a higher critical loading. In [14], it was proposed that such a scheme could function as a model for *prosopagnosia*, an impairment in visual recognition in which the patient can correctly recognize the category of faces but is unable to recognize individual faces.

However, can such a simple scheme be relevant to describe semantic memory, too? It can be argued that a tree-like structure, while suited to capture a specific cognitive impairment, does not account for the complex relations of semantic memory. When dealing with the meaning of a concept, one typically accesses not only its identity and class membership, but also the stronger or weaker relations to other concepts, which span many dimensions and are not only contingent on common human experience but also on personal experience. As such, the complexity of semantic relations [15] can be argued to require a more sophisticated description than the one provided by an approximate tree-like model.

Valuable attempts to go beyond both uncorrelated memories and simple branching trees, for example within the parallel-distributed processing (PDP) framework [16], have remained largely data driven, focused on computer simulations that could qualitatively reproduce results in agreement with patterns of deficit seen in the neuropsychological literature [17–19]. No mathematical framework, however, has been proposed for theoretical questions of a quantitative scope. Such a theoretical perspective is necessary if one wants to approach the question of semantic memory in a more principled way. For example, what is the reliability and generalizability of such results from the small networks used in simulations to large-scale cortical networks such as those of the human brain?



(a) (b) (c)

Figure 1. (**a**) Original correlation matrix. (**b**) Correlation matrix obtained by replacing each within-block entry with the mean correlation value of that noun cluster and each off-block entry with the mean correlation between clusters. The clusters are obtained through the application of a standard clustering algorithm to the original correlation matrix, Fig. 1(a). (**c**) Strictly ultrametric correlation matrix obtained by again replacing each within-block entry with the mean value within that block, and now each off-block entry with the overall off-block mean.

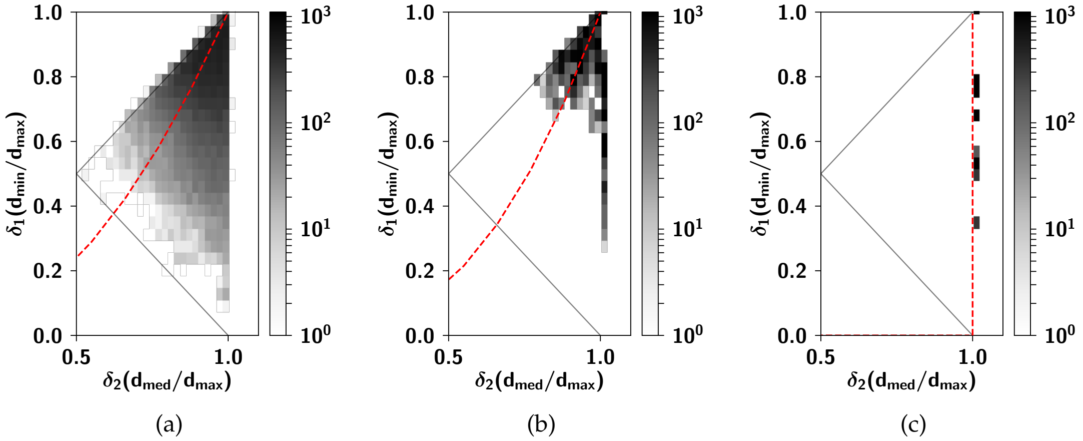


Figure 2. (**a**) Two-dimensional logarithmic density plot of the ratio between the intermediate and the longest edge vs. the ratio between the shortest and the longest edge, in the triangles created by extracting quasi-distances for the triplets of nouns taken from Fig. 1(a). The triplets are scattered, with an ultrametric content of 0.5. (**b**) Same as Fig. 2(a), with triplets taken from Fig. 1(b). The triplets have less scatter and yield an ultrametric content of 0.61. (**c**) Same as Fig. 2(a), with triplets taken from Fig. 1(c). Here triplets constitute isosceles triangles with two long sides, as can be seen from the alignment of the ratios along the vertical line $d_{med} = d_{max}$. The ultrametric content (see App. C) is exactly 1. In all three panels, the dashed red line corresponds to the line of constant ultrametric content index.

1.2. Connectivity

The analytical tools allowing for a complete analysis have been applied to fully connected or else very sparsely connected networks, in which the average connectivity between the units vanishes. These models have been thoroughly analyzed and scaling relations have been found for the storage capacity as a function of the mean connectivity and the coding sparsity in the network. Remarkably, such scaling relation holds, when coding is sparse, for both limit cases of full connectivity and extremely sparse connectivity. Does it mean that it holds also for any connectivity in between, including realistic models of cortical connectivity?

From the point of view of plausibility, such studies of randomly wired networks fall short of describing some features of the anatomy of cortical connectivity. For example, it has been shown [20] that in layers II and III of mouse visual cortex the probability of connection falls from $50-80$ percent for directly adjacent neurons to $0-15$ percent at a distance of 500 micrometers. Building on such observations, the properties of an autoassociative network of threshold-linear units whose synaptic connectivity is spatially structured has also been investigated [21]. Other studies however, have shown that at a larger scale, cortical connectivity is not randomly distributed, not even after allowing for a distance-dependent parameter. For example, it has been shown that in the prefrontal cortex of monkeys, patches of a hundred microns make connections to and from other discrete patches of cortex of the same size [22]. A patch is connected to about $15-20$ other patches in its proximity via grey matter connections, and to at least $15-20$ more distant patches connected via white matter connections.

Braitenberg and Schuz have elegantly synthesized this dual local and global characteristic of the cortex in terms of the A and B systems (referring to apical and basal dendrites [23]). They suggest regarding the whole cortex as a memory machine, in which the B-systems encode a set of memories as local attractors and the A-system encodes global attractors, by virtue of long-range connections. Variant models of associative memory networks that implement this separation of scale between dense local connectivity and sparse long-range connectivity have been studied [24–27]. This study is in line with such an approach, in that it aims at describing each patch of, say, the human cortex, a functional voxel of a few mm^3, comprising some 10^5 neurons, as one local network interacting through the B system, whose activity is coarsely subsumed into a Potts unit. A Potts unit has multiple activity states, akin to a *capsule* of the kind recently introduced in deep learning networks [28]. The Potts network, aimed at describing the cortex, or a large part of it, is comprised of N such units, constituting the A-system. We refer to [29] for a detailed analysis of the approximate thermodynamic and dynamic equivalence of the full multi-modular model and the Potts network. We do not dwell on the correspondence here, but use it to discuss correlations in the Potts framework.

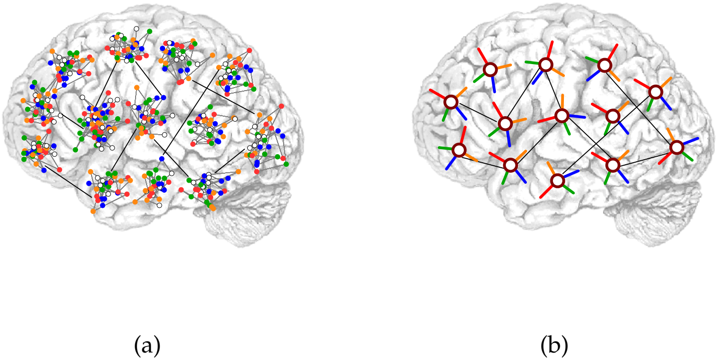

(a) (b)

Figure 3. (**a**) The Potts network, here intended as a model of semantic memory, is a coarse description of the cortex in terms of local patches of dense connectivity, which store activity patterns corresponding to local attractors (**a**). Each patch is a small local network characterized by high connectivity; *diluted* connections are instead present between units of different patches. The configuration of the individual patch is assumed to converge to a local attractor, synthetically captured by a Potts state. Each Potts unit, depicted in (**b**) can be in any of S states, where green, orange, blue and red represent the active states ($S = 4$). The white circle at the center corresponds to the quiescent state, aimed at capturing a situation of no retrieval of the underlying local network.

2. Results

2.1. The Potts network

The Potts neural network is a generalization of Hopfield's binary autoassociative network [30]. A Potts unit can be either in the *quiescent* state or else in one of the S equivalent *active* states. By convention, we label these states with numbers from 0 to S, where $k = 0$ indicates the quiescent state and $k = 1...S$ the active ones, representing the possible local attractors (see Fig. (3)). Due to stochastic fluctuations, a unit can be, with a non-vanishing probability, in several of the $S + 1$ states, so that the *activity* of unit i is a distribution over $k = 0...S$, denoted by σ_i^k, a variable in the interval $[0, 1]$. By network state, or *configuration*, we refer instead to the collection of local states assigned to all units, $\{\xi_i\}$, each of which is an integer from 0 to S, and where $i \in \{1, . . . N\}$, N being the number of units in the network.

Couplings between states of distinct units are defined, which are denoted by J_{ij}: they represent the strength with which connected units influence each other. In the case of the Hopfield network, the couplings J_{ij} are just scalars. In the Potts network, these couplings are matrices J_{ij}^{kl}, which express the strength of the coupling for the pair of units i and j being, respectively, in state k and l.

Of crucial importance in the definition of the network model is the *learning rule*, which prescribes how the couplings in the model depend on a given training data set. In the model that is dealt with here, the training data set consists of a certain number p of network configurations, denoted by $\{\xi_i^\mu\}$. We refer to these configurations as *patterns*.

The way the patterns $\bar{\xi}^\mu$ are generated, i.e. their probability distribution, has effects on the "retrieval properties" of the network, i.e. the ability to retrieve with good accuracy one of the training patterns, if this is partially cued. A quantitative measure of this ability of the network is the *storage capacity*, the number of patterns the network is able to store and retrieve, relative to the number of synaptic connections per unit.

The learning rule according to which the patterns are used to build the synaptic connections between units is a Potts-adapted version of Hebbian learning

$$c_{ij}J_{ij}^{kl} = \frac{c_{ij}}{c_m a(1-\frac{a}{S})} \sum_{\mu=1}^{p} \left(\delta_{\bar{\xi}_i^\mu k} - \frac{a}{S}\right)\left(\delta_{\bar{\xi}_j^\mu l} - \frac{a}{S}\right)(1-\delta_{k0})(1-\delta_{l0}), \tag{1}$$

where the factor c_{ij} denotes the (i,j)-th entry of the adjacency matrix of the connectivity (graph), equal to 1 if an edge exists from j to i and 0 otherwise. The constant c_m is the average degree of this graph, i.e. the average number of connections at a given node, so that $\langle c_{ij}\rangle = c_m/N \equiv \lambda$. The symbol δ here indicates the Kronecker δ-function, which is 1 when the two indices are equal and 0 if they are different. The subtraction of the mean activity by state, a/S, ensures a higher storage capacity, as initially shown for the Hopfield network in [31] and for the Potts neural network in [32].

The fully connected network, in which $c_{ij} = 1$ for all pairs (i,j), is the one which allows for a full-fledged analytic approach, by means of techniques borrowed from spin glass physics [33]. It has been shown, as reviewed in [34], that such connectivity ensures that each of these configurations, if they are not too many, becomes a stable state, or an attractor of the energy function

$$H = -\frac{1}{2}\sum_{i,j\neq i}^{N}\sum_{k,l=1}^{S} J_{ij}^{kl}\sigma_i^k\sigma_j^l + U\sum_{i}^{N}\sum_{k}^{S}\sigma_i^k, \tag{2}$$

where σ_i^k, again, can be interpreted as the probability with which the local network, synthesized into the Potts unit i, finds itself in the attractor k. This probability is given by the Boltzmann distribution with inverse temperature β

$$\sigma_i^k = \frac{e^{\beta h_i^k}}{e^{\beta U} + \sum_{l=1}^{S} e^{\beta h_i^l}}, \tag{3}$$

where h_i^k, referred to as the *field* received by unit i in state k, is determined by the activity of all the Potts units in a way that will specified later. From Eq. (3), it follows that $\sum_{k=0}^{S}\sigma_i^k = 1$ at all times.

A more biologically plausible case is that of *diluted* networks, where the number of connections per unit c_m is less than N. When the connectivity is not full (i.e. $c_{ij} \neq 1$ for some pairs (i,j)), the type of probability distribution assumed for the c_{ij} matters. In this paper we consider *random dilution* (RD), in which

$$P(c_{ij}, c_{ji}) = P(c_{ij})P(c_{ji}), \tag{4}$$

with

$$P(c_{ij}) = \lambda\delta(c_{ij}-1) + (1-\lambda)\delta(c_{ij}). \tag{5}$$

2.2. Generating correlated representations

The initial studies of the capacity of the Potts network [29,32] featured patterns that were uncorrelated. *Uncorrelated patterns* are generated by assigning Potts states to different units in different patterns independently. This means that the p patterns $\{\bar{\xi}^\mu\}$ are generated according to a probability distribution which is factorized into p identical ones, for the individual patterns

$$P(\bar{\xi}^1 \ldots \bar{\xi}^p) = P(\bar{\xi}^1)\cdot\ldots\cdot P(\bar{\xi}^p). \tag{6}$$

In turn, units in each pattern are also independent and identically distributed

$$P(\bar{\xi}^{\mu}) \equiv P(\xi_1^{\mu} \dots \xi_N^{\mu}) = P(\xi_1^{\mu}) \cdot \dots \cdot P(\xi_N^{\mu}) \,. \tag{7}$$

Every unit in each pattern is taken to be in the inactive state with probability $1-a$, with the remaining probability shared uniformly by the S active states.

$$\begin{cases} P\left(\xi_i^{\mu} = 0\right) = 1 - a \\ P\left(\xi_i^{\mu} = k\right) = \tilde{a} \equiv a/S \end{cases} \tag{8}$$

In general, for any two patterns $\mu \neq \nu$, we denote with C_0 the fraction of quiescent units they share, C_{as} the fraction of active units that are in the same state and C_{ad} the fraction of active units which are in different states. Finally C_{0a} is the number of units quiescent in μ and active in ν and conversely C_{a0} is the number of units active in μ and quiescent in ν. As an example

$$C_{as}^{\mu\nu} = \frac{1}{Na} \sum_{i=1}^{N} \delta_{\xi_i^{\mu},\xi_i^{\nu}} (1 - \delta_{\xi_i^{\nu},0}) \,. \tag{9}$$

In this simple uncorrelated scheme, the distributions of these correlation values are straightforward and given by binomial distributions with different success probabilities. We have, for example

$$P(C_{as}) = Na\, B\Big(Na\, C_{as}; N, \frac{a^2}{S}\Big) \,, \tag{10}$$

where $B(k; N, p) \equiv \binom{N}{k} p^k (1-p)^{N-k}$; that is, $\langle C_{as} \rangle = \tilde{a}$.

2.2.1. Single parents and ultrametrically correlated children

The interest in ultrametrically organized patterns was largely due to the discovery of an ultrametric hierarchy of the free energy minima in the formal solution of the Sherrington-Kirkpatrick model of a spin glass [33]. In particular, the Hopfield model of neural networks was extended to allow for the storage and retrieval of hierarchically correlated patterns [12]. In this study [12], a set of random patterns, which we can call "parents", are characterized by independent units, active with probability a

$$P(\xi_i^{\pi}) = a\,\delta(\xi_i^{\pi} - 1) + (1-a)\delta(\xi_i^{\pi}) \,, \tag{11}$$

where ξ_i^{π} denotes the activity of unit i of parent π and $0 < a < 1$ is the sparsity of the parents. In the next step, "child" patterns are drawn from the following distribution

$$P(\xi_i^{\pi\mu}) = \Big\{a + b(\xi_i^{\pi} - a)\Big\}\delta(\xi_i^{\pi\mu} - 1) + \Big\{1 - a - b(\xi_i^{\pi} - a)\Big\}\delta(\xi_i^{\pi\mu}) \,, \tag{12}$$

where $\xi_i^{\pi\mu}$ denotes the activity of unit i of child μ branching from parent π. $0 < b < 1$ parametrizes to what degree children are biased toward their (single) parent. For $b = 0$, child patterns become uncorrelated with no dependence on the parent, while for $b = 1$ the child patterns become identical to their single parent. Given the distributions above, we can compute the average activity of parents and child patterns (since the state of each unit i is drawn identically from the same distribution, in the following we can drop this index)

$$\langle \xi^{\pi} \rangle = a \tag{13}$$

$$\langle \xi^{\pi\mu} \rangle = a \,, \tag{14}$$

as well as child-parent correlations

$$\langle \xi^{\pi\mu} \xi^{\pi'} \rangle = \begin{cases} a^2 + ba - ba^2 & \pi = \pi' \\ a^2 & \pi \neq \pi' \, . \end{cases} \tag{15}$$

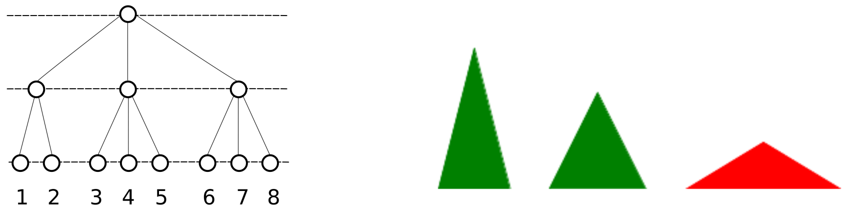
As expected, children of the same branch have higher similarity to their own parent ($\pi = \pi'$), than to a parent of another branch ($\pi \neq \pi'$). We can also compute the correlation between two children of the same parent ($\pi = \pi'$) and that of two children belonging to distinct parents ($\pi \neq \pi'$)

$$\langle \xi^{\pi\mu} \xi^{\pi'\mu'} \rangle = \begin{cases} a^2 + a(1-a)b^2 & \pi = \pi' \\ a^2 & \pi \neq \pi' \, . \end{cases} \tag{16}$$

It trivially follows that

$$\langle \xi^{\pi\mu} \xi^{\pi\mu'} \rangle - \langle \xi^{\pi\mu} \xi^{\pi'\mu'} \rangle = a(1-a)b^2 \, . \tag{17}$$

This is one of the characteristics of this algorithm: it is possible to define a distance d such that three patterns $(x, y, z = \xi^{\pi\mu}, \xi^{\pi\mu'}, \xi^{\pi'\mu'})$ *at the same level of the hierarchy* can be seen to satisfy the *strong triangle inequality*: $d(x,z) \leq max(d(x,y), d(y,z))$. As illustrated in Fig. 4(a), triplets of patterns can only be in one of the two triangle relations: equilateral and isosceles with two long edges, in other words, an ultrametric space has no node intermediate between any two nodes (Fig. 4(b)).



(a)

(b)

Figure 4. (**a**) A tree, reproduced from [33]. 1, 2, 3, 4, 5, and 6 are at the same level of the hierarchy. If we consider the nodes 1, 3 and 6, they are each at a distance of 2 of each other, the distance being defined as the distance to the nearest common branching point. If we consider nodes 3, 4 and 5, then they are each at a distance of 1 of each other, such that we get again an equilateral triangle. If we consider 1, 2 and 3, then $d_{12} = 1$ while $d_{13} = d_{23} = 2$, such that we get an isosceles triangle with two long edges. One alternative, an isosceles triangle with two short edges, is impossible to realize: there are *no intermediate points* between 1 and 3 or 2 and 3, as indicated in red in (**a**).

From the point of view of semantics, this is an implausible situation: if one considers superordinate categories as the single archetypal parent from which all concepts descend, it becomes clear that such an ultrametric structure is unsuitable in describing all the semantic relations in which the ultrametric inequality is not satisfied: for example when a concept finds itself "in between" two other concepts. On the other hand, the very meaning of a concept can be thought of as the set of features that are associated to it. It may then be more sensible to consider the features characterizing a concept as its building blocks, hence its parents. In the following, we describe an algorithm, first sketched in [35], in which each child pattern (concept) is generated from multiple parents (features), a random subset of the total group of parents relevant to it.

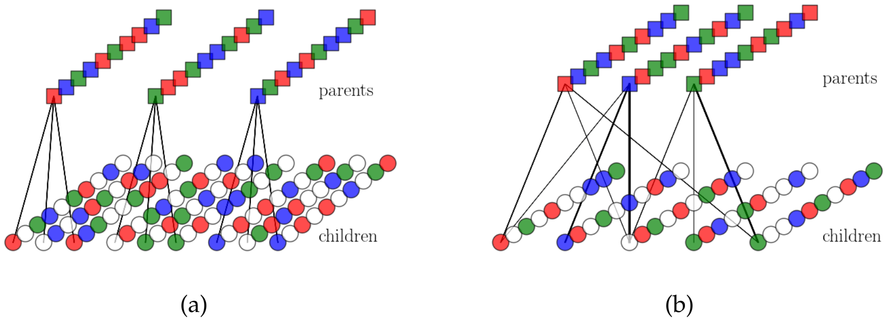


Figure 5. (**a**) The workings of a hierarchical algorithm with 3 parents and 3 child patterns per parent. Colors correspond to active Potts states while white denotes the quiescent states. $S = 3$. (**b**) The workings of the multi-parent algorithm with $\Pi = 3$ parents and $p_{par} = 3$ child patterns per parent and 5 total child patterns. Black arrows and their thickness denote strength of input. The main difference with the hierarchical algorithm is that each child pattern can receive input from multiple parents. If each parent is to represent a feature and each child a concept, the algorithm entails the generation of a concept from multiple features.

2.2.2. Multiple parents and non-trivially organized children

How can we incorporate a plausible featural description into our model of semantic memory? One may consider features as the parents from which child concepts are derived. We can then map quantities such as the number of features, their sharedness, and their dominance to appropriate parameters in our model.

Our simple version of the multi-parent pattern generation algorithm works in three stages. In the first stage, a set of Π random patterns are generated to act as parents. In the second stage, each of the Π parents are assigned to p_{par} randomly chosen children. Then, each "child" pattern is generated: each pattern, receiving the influence (or input) of its parents, aligns itself, unit by unit, in the direction of the largest input. In the third and final step, the fraction a of the units with the largest inputs is set as active in each child pattern. A schematic representation can be seen in Fig. 5(b).

2.2.3. The algorithm operating on simple binary units

Each parent is assigned p_{par} children out of a total of p. The probability distribution that a given child has n_p parents, out of a total pool of Π is given by a binomial, with the *prolificity* $f = p_{par}/p$

$$P(n_p) = \binom{\Pi}{n_p} f^{n_p} (1-f)^{\Pi - n_p} \tag{18}$$

The algorithm draws, for the input $x_i^{\pi \to \mu}$ from unit i of parent π to unit i of pattern μ, a uniformly distributed random number in the interval $(0, 1]$ with probability a_p and zero with probability $1 - a_p$ such that we can write

$$P(x_i^{\pi \to \mu}) = a_p \, U_{(0,1]}(x_i^{\pi}) + (1 - a_p)\delta(x_i^{\pi}) \,, \tag{19}$$

where a_p, which we can call the *extent* of the input from one parent, is analogous to the a parameter in Eq. (11); indeed, if $a_p \sim 0$, a child pattern is very unlikely to receive, on a particular unit, the contribution from one of its parents. On the other hand, if $a_p \sim 1$ then all parents influencing a child contribute to its input, whichever the unit. $U_{(0,1]}$ denotes the uniform distribution, such that input from parents is graded, contrary to the previous section.

Here, we have made the choice of non-sparse parents, but sparse input from parents, aimed at decorrelating units, while conserving correlations between patterns. This choice will prove to be crucial in Sect. 2.3.1, where statistical independence between units will lead to a vanishing mean noise, using only a simple covariance rule. For $S = 1$, this means that the patterns generated by the algorithm are uncorrelated, but the importance of having non-sparse parents with sparse input from them becomes important when dealing with more than one Potts state. Nevertheless, in this section, we consider $S = 1$, before treating genuine Potts units.

The main difference with respect to the single-parent algorithm is that now, one must compute the total field h_i^μ that a unit i of pattern μ receives from all parents

$$h_i^\mu = \sum_{\pi=1}^{\Pi} x_i^{\pi\to\mu}\, \mathbb{I}_{\Omega_\mu}(\pi) + \epsilon\,, \tag{20}$$

where Ω_μ is the set of all parents acting on pattern μ and where we have that $|\Omega_\mu| = n_p(\mu)$. $\mathbb{I}_{\Omega_\mu}(\pi)$ is the indicator function that is 1 if parent π is assigned to pattern μ and 0 otherwise. ϵ is a small random input ($\epsilon \ll 1$) allowing for some input, even when $a_p \ll 1$. The fields of all units of all patterns have the same distribution. In App. A, the full derivation of the probability distribution for the field h_i^μ is reported. Such a distribution has a non-trivial expression and, to our knowledge, it can only be evaluated numerically. However, a simple analytic expression can be given for the moments of the distribution of h_i^μ

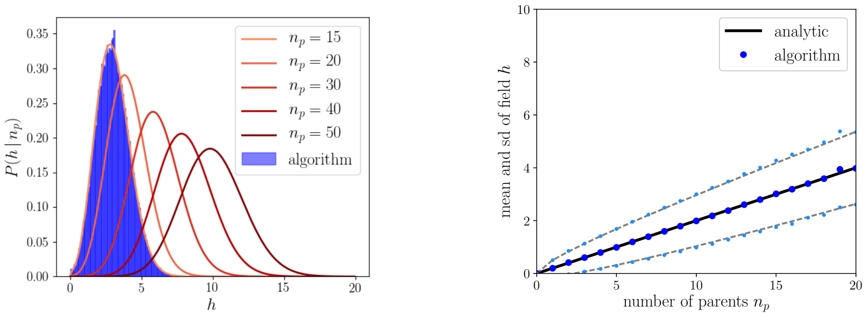
$$\langle h \rangle = n_p\, \frac{a_p}{2}\,, \tag{21}$$

$$\sigma_h = \sqrt{n_p\, a_p \left(\frac{1}{3} - \frac{a_p}{4}\right)}\,, \tag{22}$$

as shown in Fig. 6(b). In Fig. 6(a), we see that these analytical results match tightly those from implementation of the algorithm.

As a last step, a fraction a of the units within a given pattern having fields above a threshold h_m are set to become active. The threshold h_m is then implicitly given in terms of the cumulative distribution function

$$P(h' < h_m \,|\, n_p) = 1 - a\,. \tag{23}$$



(a) (b)

Figure 6. (**a**) Solid lines correspond to the analytical distributions of the field, Eq. (A8), in blue is the distribution of the fields produced by a simulation of the algorithm for $n_p = 15$. The parameters are $N = 2000$, $S = 1$, $a_p = 0.4$, $n_p = 15 \dots 50$ and $\Pi = 100$. (**a**) The mean and standard deviation of the field as a function of the number of parents.

For any given child pattern μ with number of parents n_p, we can now define the probability that it will be activated, given the field that it receives

$$P(\xi_i^\mu = 1 \,|\, h_i^\mu) = \Theta(h_i^\mu - h_m)\,. \tag{24}$$

2.2.4. The algorithm operating on genuine Potts units

With genuine Potts states, the main difference with respect to the previous case is that the input from a parent π to the field of its child patterns can be, on a given unit, to any one of S states, with equal probability. This means that only a subset $\Omega_{i,k}$ of the total parents will contribute to state k of unit i. We denote the number of parents in the subset as $|\Omega_{i,k}| = n_i^k$. The joint distribution of number of parents by state is

$$P(n_i^1, ..., n_i^S) = \frac{n_p!}{S^{n_p} \prod_{k=1}^{S} n_i^k!} \,, \tag{25}$$

such that the constraint $\sum_{k=1}^{S} n_i^k = n_p$ is satisfied. We can then write the field of unit i in state k of pattern μ

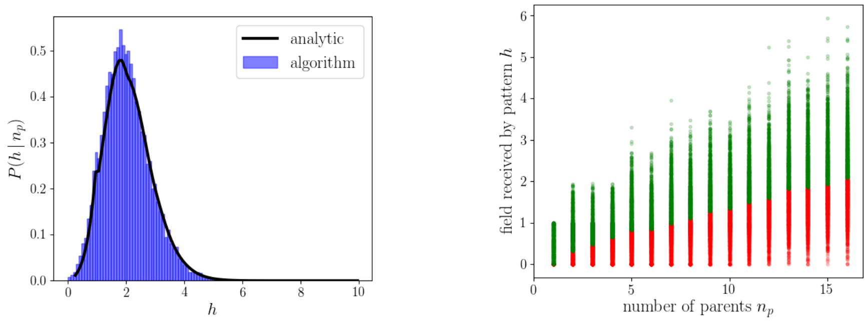
$$h_{i,k}^{\mu} = \sum_{\pi=1}^{\Pi} x_{i,k}^{\pi\to\mu} \mathbb{I}_{\Omega_{i,k}^{\mu}}(\pi) + \epsilon \,. \tag{26}$$

Then, the algorithm is such that it selects, unit by unit, the state receiving the maximal input. Following some calculations shown in App. B, we can compute the distribution of the fields for those states having received maximal input H (Fig. 7(a)). We can then compute, exactly as before, the threshold above which the unit becomes activated

$$P(H' < H_m \,|\, n_p) = \int_{-\infty}^{H_m} P(H' \,|\, n_p)\, dh' = 1 - a \,. \tag{27}$$

Having obtained the minimal field H_m required to activate a unit (Fig. 7(b)), we now need only the distribution of the field given the number of parents in that state $P(h^k|n^k)$, which is none other than Eq. (A8) (replacing n_p with n^k). We finally get to the distribution of activity across units and states, given the field received

$$P(\xi_{ik}^{\mu} = 1 \,|\, h_{ik}^{\mu}) = \Theta(h_{ik}^{\mu} - H_m) \,. \tag{28}$$



(a)

(b)

Figure 7. (**a**) Distribution of the maximal fields for $S = 2$ and $n_p = 30$. In blue is the distribution of the fields produced by the algorithm and the black line is Eq. (B7). (**b**) The x-axis orders patterns with different number of parents and the y-axis the fields of the units in that pattern. Red points correspond to units that are set to quiescent and green to those that are activated. The boundary between the green and the red corresponds to h_m, the minimum field required for a unit to be set to active. Parameters are $N = 2000$, $S = 2$, $a_p = 0.4$ and $\Pi = 100$.

Given the algorithm just described, the main mechanism determining the state of a unit in a given pattern is how many of the parents affecting a child are in the same state. If parents are all aligned, this makes the unit receive a higher field in a single state, making it more probable to become activated. On the other hand, lower alignment between parents results in the field received by a child unit to be spread among the different states, and make it less probable for the child unit to find itself among those with maximal fields, as given by Eq. (B7).

We have described the mechanism through which individual child patterns are generated. At this level, in order to determine whether or not a unit of a pattern will become activated, the only relevant parameter is the number of parents, we well as their degree of alignment in Potts space. From the point of view of an individual child pattern then, all parents are equivalent and can be considered as identical and independently distributed, a property exploited above. In the next section, we turn to the correlations between patterns. Are they dominated by the number of parents that a pair of child patterns have in common? Is this a plausible model for semantic memory?

2.2.5. Resulting patterns and their correlations

In Fig. 8(a) and Fig. 8(b) we can see sample patterns generated randomly and with the algorithm from a common set of Π parents, respectively. Patterns generated by the algorithm sample different active states uniformly, such that Eq. (8) still holds, though the joint distribution $P(\bar{\xi}^1 \dots \bar{\xi}^p)$ is not factorizable anymore, as it was in Eq. (6).

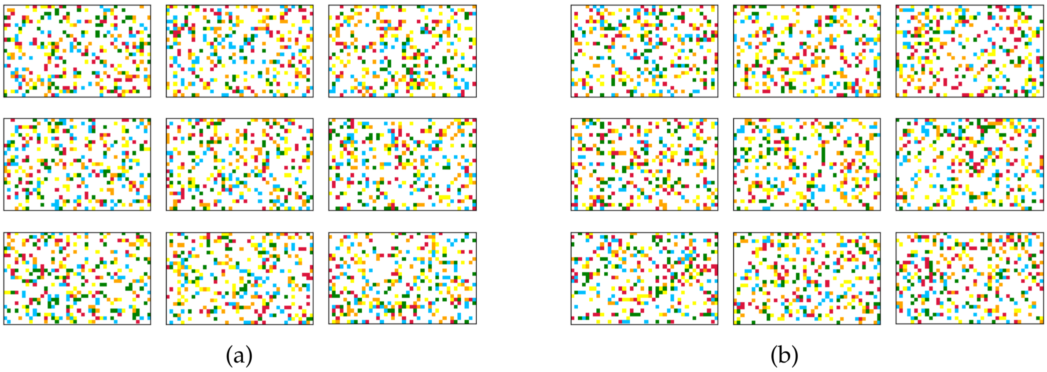


Figure 8. (**a**) Nine sample patterns generated independently following Eq. (8). Each subplot corresponds to one pattern. Each colored square to one unit, the colors indicating active states and white indicating the quiescent state. (**b**) Same as (**a**) but generated from one run of the multi-parent algorithm. The uniform sampling of the different Potts states is such that Eq. (8) holds. By design, also Eq. (7) holds (see text) while Eq. (6) does not hold anymore. Correlation parameters are $a_p = 1$, $f = 0.05$ and $\Pi = 150$ parents. Parameters are $S = 5$ and $a = 0.3$.

In Sect. 2.2.2 we discussed how the activity of different units is still approximately uncorrelated. We can see this by computing, analogously to the correlation between patterns, Eq. (9), the correlation between units as the fraction of patterns in which two units are co-active and in the same state

$$C_{ij} = \frac{1}{pa} \sum_{\mu}^{p} \delta_{\bar{\xi}_i^\mu,k} \delta_{\bar{\xi}_j^\mu,k} (1 - \delta_{k,0}) \,. \tag{29}$$

In Fig. 9 we can see the distributions of $C_{\mu\nu}$ and C_{ij} for nine different combinations of the extent a_p and prolificity f parameters. The distributions are very sensitive to the specific values of the parameters. For low values of a_p and f, pairs of Potts units have uncorrelated activity when averaged across patterns, in the sense that the distribution C_{ij} has zero covariance. Pairs of patterns, instead, are correlated with a distribution $C_{\mu\nu}$ of non-zero covariance, that is positively skewed. Low values of a_p and high values of f result in both distributions becoming more and more normal, while high values of a_p and low values of f result in a normally distributed correlation between units and a highly skewed multi-modal distribution between patterns.

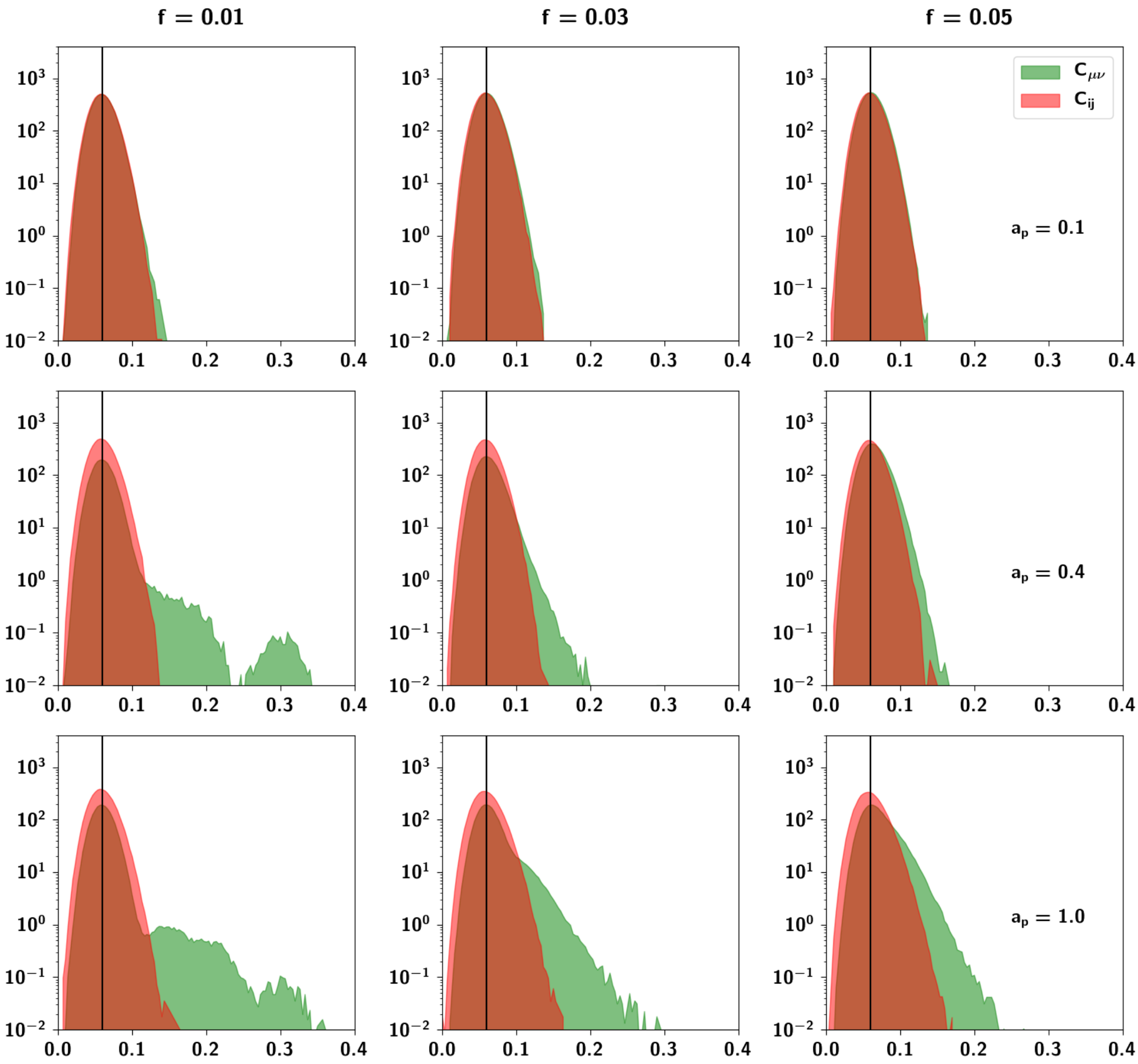
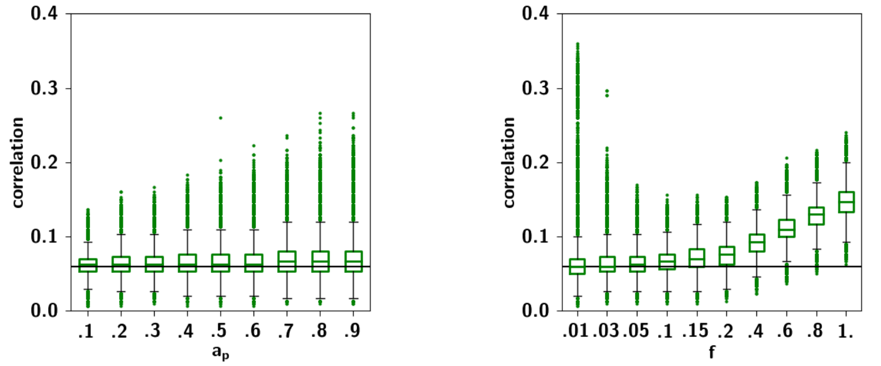
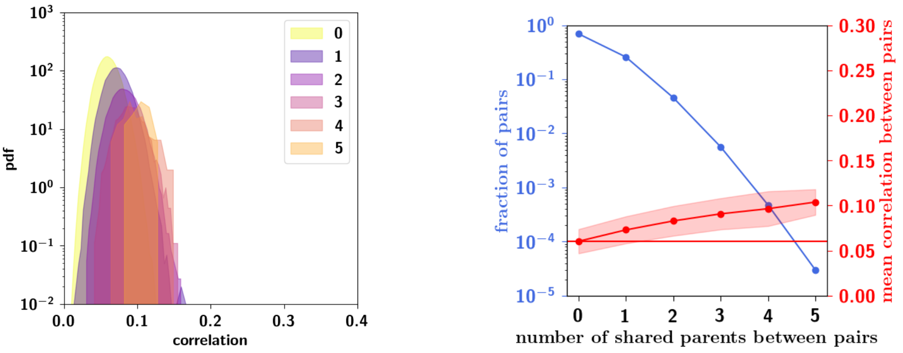


Figure 9. Probability density function of correlations between units (in red) and between patterns (in green) for three different values of both a_p and f, the latter yielding in this case an average of 1.5, 4.5 and 7.5 parents per pattern. The black vertical line corresponds to the average correlation with uncorrelated patterns distributed independently according to Eq. (8). The parameters are $S = 5$, $a = 0.3$, and $\Pi = 150$. The algorithm produces correlations between patterns with high variability relative to the correlation between units, in line with ideas about semantic memory. Note that the algorithm is sensitive to the parameters and their values strongly affect the correlation between patterns.



(a) (b)

Figure 10. (**a**) Boxplots of $C^{\mu\nu}$ for different values of a_p, with $f = 0.05$ fixed. (**b**) Boxplots of $C^{\mu\nu}$ for different values of f with $a_p = 0.4$ fixed. The parameters a_p and f play different roles in generating the correlations. Increasing the extent a_p of the input they receive from each parent increases the overall similarity of those children having shared parents, as evidenced by the increasing skewness of the distributions. In contrast, increasing the prolificity f, leads to an increase in the mean number of shared parents, such that all children are more correlated, as shown by the shift in the overall distribution. The black horizontal line corresponds to the average correlation with uncorrelated patterns distributed according to Eq. (8). Other parameters are $a = 0.3$, $S = 5$ and $\Pi = 150$.



(a) (b)

Figure 11. (a) Another visualization of the correlation distribution of Fig. 9, with $f = 0.05$, $a_p = 0.4$ and $\Pi = 150$, decomposed into the distribution for each number of shared parents. **(b)** Fraction of pairs of patterns (left y-axis, note the logarithmic scale) and mean correlation between those pairs (right y-axis, linear scale) as a function of number of shared parents. The red horizontal line corresponds to the average correlation with uncorrelated patterns distributed according to Eq. (8). Pairs of patterns having more shared parents are markedly fewer, although on average more correlated, so they do not affect much the overall mean correlation.

To assess these observations more systematically, in Fig. 10(a) we can see boxplots of the $C_{\mu\nu}$ distributions for different values of a_p keeping $f = 0.05$ fixed. While the mean correlation is unaffected by increasing a_p, the standard deviation and the skewness increase. In 10(b), conversely, we can see boxplots of $C_{\mu\nu}$ distributions for different values of f keeping $a_p = 0.4$ fixed. It can be seen that increasing f increases the mean correlation between patterns. The effects observed can be understood intuitively because of the different roles that these parameters play in the algorithm. The extent a_p is the parameter that increases the probability that a child unit receives input from a parent, increasing the overall similarity of a child to its parents. This means that those children that have a larger number of shared parents will be more similar and more strongly correlated, giving rise to the larger values in the distribution. The prolificity f, on the other hand, is the ratio of the pool of children affected by one parent to the total number of children. Increasing this ratio leads to an increase in $\langle n_p \rangle$, the mean number of parents, such that children tend to share more parents. It can be seen in Fig. 11(b), in which pairs of patterns are decomposed into different distributions sharing an increasing number of parents ($0-5$ shared parents), that for a pair of patterns, a higher number of shared parents leads to a higher mean correlation. The number of such pairs is markedly fewer, as can be seen in the left axis of Fig. 11(a) (plotted in a logarithmic scale), but if f is high enough, this effect is enough to increase the overall mean correlation between all patterns. The two parameters a_p and f therefore play different roles in generating the correlations.

2.2.6. The ultrametric limit

It is interesting to note a limit case of the algorithm. For low prolificity, if e.g. $\langle n_p \rangle_\mu = \Pi f \sim 1$ as in Fig. 9 (left column, i.e., $f = 0.01$, $\Pi = 150$), on average most children will have a single parent, which effectively produces ultrametric patterns. Indeed, for these parameters, since the number of total parents $\Pi = 150$ is smaller than the total number of children generated, $p = 1000$, several children share a given single parent. The mean value of their correlation with all other children, however, at a/S, is the same as the mean correlation between uncorrelated patterns, as stated in Eq. (16). Note that the distribution is multimodal. The values forming the second mode of the distribution express the correlation between children belonging to the same (single) parent.

2.2.7. The random limit

Another limit is the random or limited-parent-influence limit, in which $a_p \ll 1$ (effectively, the top row in Fig. 9). In this case, most units will not receive input from their respective parents, regardless of how many they are, and the unit will align itself in the direction of a random Potts state given by the input ϵ. In this way, it is possible to parametrically generate patterns ranging from independent ($a_p \ll 1$) to ultrametric ($a_p = 1$, $f\Pi = 1$), from the top row to the left column of Fig. 9, but also to enter the area of complexity to the bottom right, where correlations might begin to resemble plausible semantic relations.

2.2.8. Semantic dominance

Returning to the correlation observed among the nouns we considered in Sect. 1.1 as our toy example, how important is, there, each individual feature? We can quantify it through a simple measure of semantic "dominance", by simply summing the feature weights of all nouns $s_j = \sum_i^N w_{ij}$.

In Fig. 12, we report the summed weights of the $M = 50$ features across all the nouns considered, sorted and plotted on a semi-logarithmic scale. Remarkably, given the very small dataset used, it can be approximated to a good extent by an exponential law. The suboptimal fit may conceivably be the result of limited and unbalanced sampling. Indeed the words, the nouns or the verbs were not chosen with comparable frequency. This measure is therefore only approximate, as an aggregate measure of dominance. Our measure is related to the measure called "semantic relevance" used by Sartori and colleagues [36] as well as to the "semantic differential" used by Osgood [37]. The difference with the latter measure, however, is that ours is cumulative across all of the nouns and derived from co-occurrence statistics in a corpus, while the semantic differential refers to a scale in which individuals rate the connotative meaning of objects, events, and concepts.

To take into account this observation, we consider a more refined model in which the parents in our algorithm (the features), ranked from 1 to Π, have the strength of their inputs damped exponentially with a dominance rate ζ, such that Eq. (26) is revised in the following way

$$h^\mu_{i,k} = \sum_{\pi=1}^{\Pi} x^{\pi\to\mu}_{i,k} \mathbb{I}_{\Omega^\mu_{i,k}}(\pi) \exp(-\zeta\pi) + \epsilon\,, \tag{30}$$

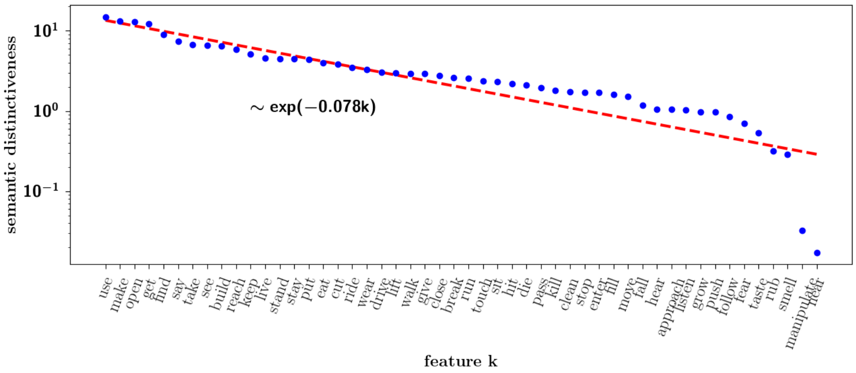


Figure 12. The x-axis lists all of the features used to compute the correlation between the nouns in the toy example of Sect. 1.1, sorted according to their summed weights across all nouns (reported on a semi-logarithmic y-axis). The exponent of the fit, $\zeta = 0.078$, indicates that the semantics of this particular set of nouns is effectively dominated by a set of order $1/\zeta \simeq 10$ features.

where $x_{i,k}^{\pi \to \mu}$ is the input from parent π to child pattern μ, $\Omega_{i,k}^{\mu}$ is the set of all parents acting on pattern μ and $\mathbb{I}_{\Omega_{i,k}^{\mu}}(\pi)$ is the indicator function that is 1 if parent π is assigned to pattern μ and 0 otherwise. The limit $\zeta \to 0$ corresponds to the algorithm described in the previous sections, such that we recover Eq. (26). In this way, we introduce a parameter, ζ, which can be related to the slope seen in dominance distributions observed in real data, such as the one in Fig. 12.

In Fig. 13 we can see a schematic representation of this new algorithm. In contrast to the extent a_p, the parameter ζ, though also affecting the strength of input, plays a different role, as it affects the global strength with which each parent affects its children, leading to variability of input across patterns. A high value of ζ contributes to highly unbalanced input from parents influencing a child pattern, such that units tend to align each with the most powerful parent, or the most dominant feature.

How are the correlations affected by the dominance ζ? In Fig. 14 we report the distributions for three different values of the dominance ζ and prolificity f. While for low values of ζ, i.e. parents homogeneous in their strengths, the correlation between patterns is unaffected (see Fig. 9), increasing ζ we see the emergence of a tail of highly correlated patterns. For small f, this has the effect of smearing the bi-modal distribution, while for larger f, the already existing tail becomes fatter. This effect can also be seen more summarily in Fig. 15. Parents' dominance reinforces children correlations.

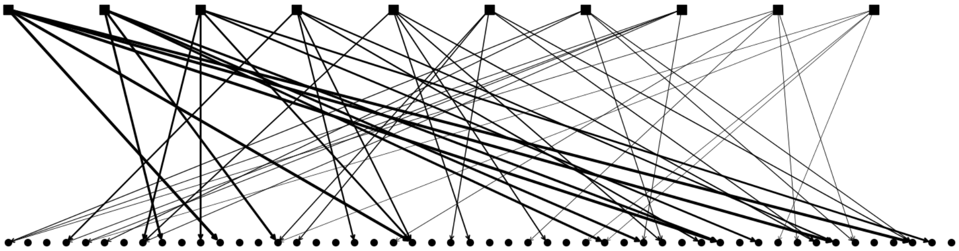


Figure 13. One sample representation of parent-child relations. The squares on the top row represent parents, while the circles at the bottom row represent children. Black lines represent input from the parents to the children. The strength with which each parent affects its children is proportional to $\exp(-\zeta\pi)$, where π indexes the parents, as explained in the text. For illustration, there are $\Pi = 10$ parents, $p_{par} = 5$ children per parent and $p = 50$ total children.

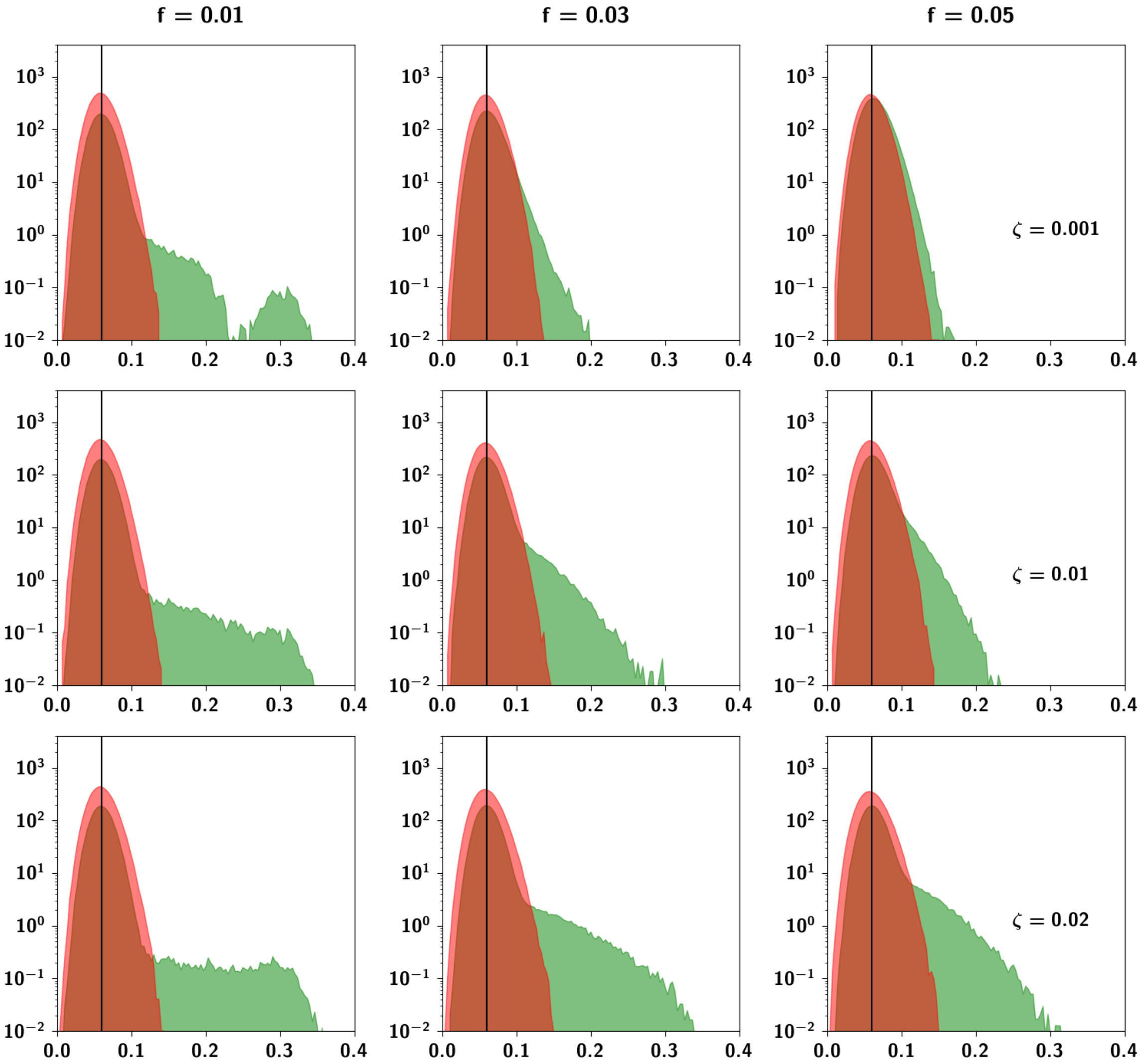


Figure 14. Probability density function of correlations between units (in red) and between patterns (in green) for three different values of the dominance rate ζ and prolificity f, keeping $a_p = 0.4$ constant. For the low value of $\zeta = 0.001$, this figure reproduces the middle panel of Fig. 9. For higher values of ζ, where the parents become highly heterogeneous, we see the emergence of large correlations.

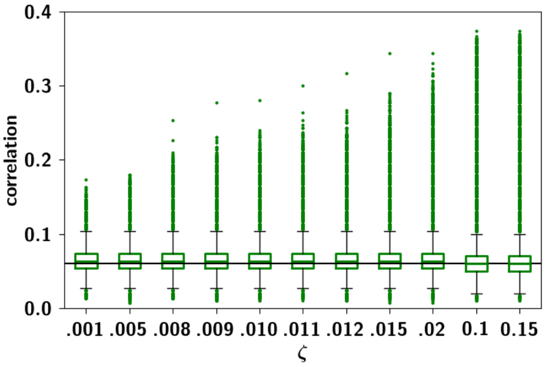


Figure 15. $C^{\mu\nu}$ for different values of ζ, with $a_p = 0.4$ and $f = 0.05$ fixed. Other parameters are $a = 0.3$, $S = 5$ and $\Pi = 150$.

2.3. Storage capacity of the Potts network with correlated patterns

Having defined an algorithm which generates correlated patterns, we can turn to study the storage capacity and how it is affected by the correlations. We have carried out numerical simulations for Potts networks [32] with the learning rule in Eq. (1), and have observed that the storage capacity is diminished in the case of correlated patterns, a result that has been obtained analytically by others [38,39], albeit for different sources of correlations.

2.3.1. Self-consistent signal to noise analysis

In [29], we have discussed the application of the self-consistent signal to noise analysis (SCSNA) to the Potts network with uncorrelated patterns. In the following section we extend this analysis to the case of correlated patterns encoded by a Potts network with diluted random connectivity (Eq. (5)) and obtain estimates of the storage capacity accounting for correlations. The local field of unit i in state k writes

$$h_i^k = \sum_j \sum_l c_{ij} J_{ij}^{kl} \sigma_j^l - \tilde{U}\left(1 - \delta_{k,0}\right) , \tag{31}$$

where the coupling strength between two states of two different units J_{ij}^{kl} is given by Eq. (1). We make the assumption that the field can be written simply as the sum of two terms, signal and noise. While the signal, the contribution from the condensed pattern (that we label as $\mu = 1$ in Eq. (1)), is what pushes the activity of the unit such that the network configuration converges to an attractor, the noise, or the crosstalk from all of the other patterns, is what deflects the network away from the cued memory pattern. The noise term writes

$$n_i^k \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} \sum_l v_{\xi_i^\mu k} v_{\xi_j^\mu l} \, \sigma_j^l \, , \tag{32}$$

that is, the contribution to the weights J_{ij}^{kl} by all non-condensed patterns ($\mu > 1$). By virtue of the subtraction of the mean activity in each state $\tilde{a}$, the noise has vanishing average:

$$\langle n_i^k \rangle_{P(\xi)} \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} \sum_l \langle v_{\xi_i^\mu,k} \rangle \langle v_{\xi_j^\mu,l} \sigma_j^l \rangle = 0 \, . \tag{33}$$

The variance of the noise can be approximately written in the following way:

$$\langle (n_i^k)^2 \rangle \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)=1}^{N} \sum_l \sum_{\mu'>1}^{p} \sum_{j'(\neq i)=1}^{N} \sum_{l'} c_{ij} c_{ij'} \langle v_{\xi_i^\mu,k} \, v_{\xi_i^{\mu'},k} \rangle \langle v_{\xi_j^\mu,l} \, v_{\xi_{j'}^{\mu'},l'} \, \sigma_j^l \sigma_{j'}^{l'} \rangle \, , \tag{34}$$

where statistical independence between units is implicitly used. While in the case of uncorrelated patterns, all terms but $\mu = \mu', j = j'$ and $l = l'$ vanish, with correlated patterns this is not the case. Now, the additional terms $\mu \neq \mu', j = j'$ and $l = l'$ must be considered. Given the statistical independence of units, however, all other terms are zero. Having identified the non-zero terms, we can proceed with the capacity analysis. We can express the field, Eq. (31) using the overlap parameter

$$h_i^k = v_{\xi_i^1 k} m_i^1 + \sum_{\mu>1} v_{\xi_i^\mu k} m_i^\mu - \tilde{U}(1 - \delta_{k0}) \, , \tag{35}$$

where we define the local overlap m_i^ν as

$$m_i^\nu = \frac{1}{c_m a(1 - \tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^\nu l} \sigma_j \, . \tag{36}$$

At the root of the SCSNA [21,40,41] is the assumption that the noise term itself can be expressed as the sum of two terms, one proportional to the activity of unit i and the other a Gaussian random variable,

$$\sum_{\mu>1} v_{\xi_i^\mu,k}\, m_i^\mu = \gamma_i^k \sigma_i^k + \sum_{n=1}^{S} v_{n,k}\, \rho_i^n z_i^n\,; \tag{37}$$

z_i^n are standard Gaussian variables, and γ_i^k and ρ_i^n are positive constants to be determined self-consistently. The first term, proportional to σ_i^k, represents the noise resulting from the activity of unit i on itself, after having reverberated in the loops of the network; the second term contains the noise which propagates from units other than i. The activation function writes

$$\sigma_i^k = \frac{e^{\beta h_i^k}}{\sum\limits_l e^{\beta h_i^l}} \equiv F^k\Big(\{y_i^l + \gamma_i^l \sigma_i^l\}_l\Big)\,, \tag{38}$$

where $y_i^l = v_{\xi_i^1,l}\, m_i^1 + \sum_n v_{n,l}\rho_i^n z_i^n - U(1-\delta_{l,0})$. The activity σ_i^k is then determined self-consistently as the solution of Eq. (38)

$$\sigma_i^k = G^k\Big(\{y_i^l\}_l\Big)\,, \tag{39}$$

where G^k are functions solving Eq. (38) for σ_i^k. However, Eq. (38) cannot be solved explicitly. Instead we make the assumption that $\{\sigma_i^l\}$ enters the fields $\{h_i^l\}$ only through their mean value $\langle\sigma_i^l\rangle$, so that we write

$$G^k\Big(\{y_i^l\}_l\Big) \simeq F^k\Big(\{y_i^l + \gamma_i^l \langle\sigma_i^l\rangle\}_l\Big)\,. \tag{40}$$

The coefficients in the SCSNA ansatz, Eq. (38), $\gamma_i^k = \gamma$ and $\rho_i^k = \rho^k$ are found to be

$$\gamma = \frac{\alpha\lambda}{2S}\frac{\Omega}{1-\Omega}\,, \tag{41}$$

$$(\rho^n)^2 = \frac{\alpha P_n}{S(1-\tilde a)}\, q\left\{1 + \frac{p\,\overline{C_{as}}}{a(1-\tilde a)}\Big(\overline{C_{as}} - \tilde a\Big)\right\}\left\{1 + 2\lambda\Psi + \lambda\Psi^2\right\}. \tag{42}$$

where $\alpha = p/c_m$ as before, and where C_{as}, defined in Sect. 2.2.5, is the fraction of units that are in the same Potts state in two different patterns, normalized by a. Note the second term in the first curly brackets that scales with p^2/c_m and is proportional to $\overline{C_{as}} - \tilde a$, the covariance between patterns. This term originates from the additional non-zero terms in the sum in Eq. (34) due to correlations between patterns. When uncorrelated patterns are considered, such that $\overline{C_{as}} = \tilde a$, it becomes zero. In this calculation, we assume that correlations between the v operators of order higher than the second are negligible. As a consequence, the only quantities involved are their covariances. This approximation corresponds to the assumption that the v operators are normally distributed. Following the same procedure reported in [29], Ω, q and Ψ are found to be

$$\Omega = \left\langle \frac{1}{NS}\sum_j \sum_l \frac{\partial G_j^l}{\partial y^l} \right\rangle, \tag{43}$$

$$q = \left\langle \frac{1}{Na}\sum_j \sum_l (G_j^l)^2 \right\rangle, \tag{44}$$

$$\Psi = \frac{\Omega}{1-\Omega}\,, \tag{45}$$

where $\langle\cdot\rangle$ indicates the average over all patterns. The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k} m + \frac{\alpha}{2S}\lambda\Psi(1-\delta_{k,0}) - U(1-\delta_{k,0}) + \sum_{n=0}^{S} v_{n,k} z^n \sqrt{\frac{\alpha P_n}{S(1-\tilde{a})} q\left\{1+2\lambda\Psi+\lambda\Psi^2\right\}\left\{1+\frac{p\,\overline{C_{as}}}{a(1-\tilde{a})}\left(\overline{C_{as}}-\tilde{a}\right)\right\}}. \tag{46}$$

Taking the average over the non-condensed patterns (the average over the Gaussian noise z), followed by the average over the condensed pattern $\mu = 1$ (denoted by $\langle\cdot\rangle_\xi$), in the limit $\beta \to \infty$, we get the self-consistent equations satisfied by the order parameters

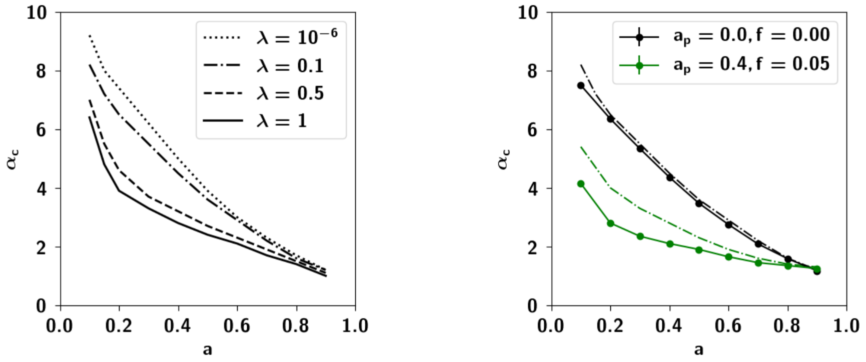
$$m = \frac{1}{a(1-\tilde{a})}\left\langle \int D^S z \sum_{l(\neq 0)} v_{\xi,l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\right\rangle_\xi , \tag{47}$$

$$q = \frac{1}{a}\left\langle \int D^S z \sum_{l(\neq 0)} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\right\rangle_\xi , \tag{48}$$

$$\Omega = \frac{1}{\tilde{a}\sqrt{\alpha q\left\{1+2\lambda\Psi+\lambda\Psi^2\right\}\left\{1+\frac{p\,\overline{C_{as}}}{a(1-\tilde{a})}\left(\overline{C_{as}}-\tilde{a}\right)\right\}}} \cdot \left\langle \int D^S z \sum_{l(\neq 0)} \sum_{k} \sqrt{\frac{P_k}{S(1-\tilde{a})}}\, v_{kl}\, z^k \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\right\rangle_\xi \tag{49}$$

The averaging in Eqs. (47)-(49) can be performed analytically and we refer to [32] and [29] for their expressions. The storage capacity α_c is defined as the maximal α that solves the set of equations Eqs. (47)-(49).

2.3.2. Numerical solutions of mean-field equations and simulations



(a) (b)

Figure 16. (a) Storage capacity as a function of the sparsity a for different values of the dilution parameter λ, for uncorrelated patterns ($C_{as} = \tilde{a}$). **(b)** Storage capacity for uncorrelated patterns (in black) and for correlated patterns (in green). Dots correspond to simulations while the dashed lines to solutions of the mean-field equations. It is apparent that the mean-field treatment yields better results for uncorrelated patterns; for correlations, it over-estimates the storage capacity. Parameters are $N = 2000$, $c_m = 200$, $S = 5$, $U = 0.5$ and $\beta = 200$.

In Fig. 16(a) we solve the set of self-consistent equations Eqs. 47-49 for different values of the sparsity a and λ for the simpler case of uncorrelated patterns. In Fig. 16(b) we can see the agreement of the former solutions for $\lambda = 0.1$ with simulations. We can also see, in the same figure, the mean-field solutions as well as simulations for correlated patterns, with the values of $\overline{C_{as}}$ obtained from simulations of the algorithm. For lower values of the sparsity, the solution to the mean-field equations over-estimates the capacity compared to what we obtain through the simulations, possibly because the mean-field treatment does not account for the fluctuations in the correlations obtained through the algorithm, but only the increase in the excess mean correlation. For higher values of the sparsity, the agreement is better presumably because the correlations produced by the algorithm become dominated by the mean.

In Fig. 17 we show the storage capacity for correlated patterns, for different values of the correlation parameters a_p and f. As can be seen in Fig. 17(a), increasing either extent of influence or prolificity, whatever the sparsity, is detrimental to the capacity. In Fig. 17(b), instead, we can see the capacity as a function of the number of Potts states S. For $S = 1$, as the algorithm produces uncorrelated patterns, the capacity remains the same, regardless of the correlation parameters. For higher values of S, on the other hand, the capacity decreases with increasing values of the correlation parameters f and a_p. The behavior of the capacity as a function of c_m, shown in the simulations of Fig. 17(c) which have been carried out with random dilution of the connectivity (see Eq. 5) shows the same strong dependence on correlations. The decrease in capacity brought on by the correlations is due to the increased variance of the noise, discussed in the previous section.

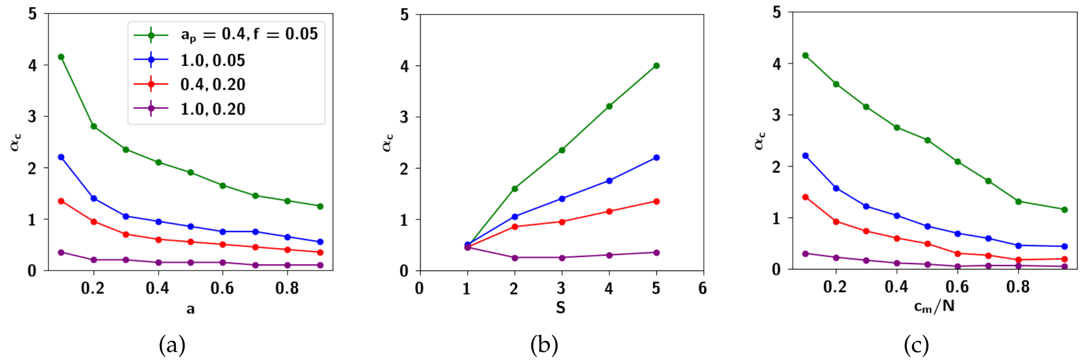


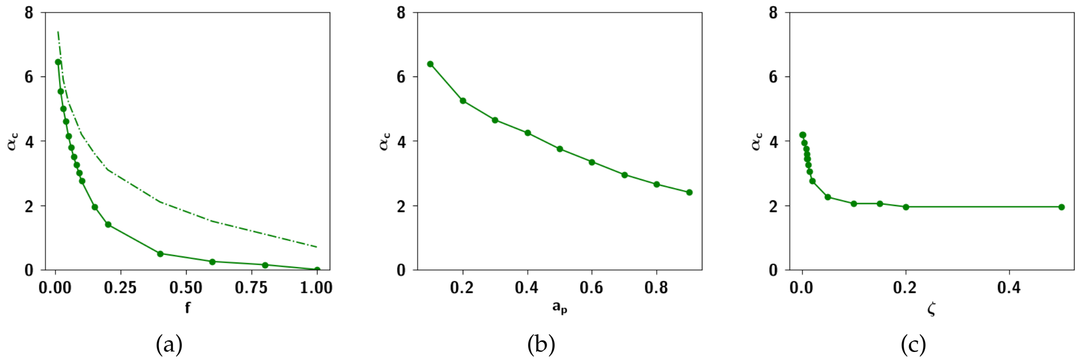
Figure 17. (a) Storage capacity α_c as a function of the sparsity a for different values of the correlation parameters a_p and f. The storage capacity is defined as the critical storage at which half of all cued patterns are retrieved with overlap of 0.7 and above. Increasing a_p and f are both generally detrimental to the capacity. **(b)** α_c as a function of the number of Potts states S. **(c)** α_c as a function of the connectivity c_m for random dilution (see Eq. 5). The capacity decreases as a function of increasing connectivity. This can be contrasted to uncorrelated patterns, where for this model it is found that the capacity remains quasi-constant with increasing c_m, at least for the parameters in a certain parameter range. When not explicitly varied, parameters are $N = 2000$, $c_m = 200$, $a = 0.1$, $S = 5$, $U = 0.5$, $\beta = 200$, $\zeta = 10^{-6}$ and $\Pi = 150$.

2.3.3. The effect of correlation parameters f, a_p and ζ

In Fig. 18 we see the storage capacity as a function of the three different correlation parameters f, a_p and ζ. We can see that increasing each of these parameters decreases capacity, albeit in very different manners. The dependence of α_c on the prolificity f can be seen in Fig. 18(a): α_c decreases dramatically with increasing f, and goes to zero for very high values of f, in which children are each affected by a large number of parents. This result makes sense in light of the fact that f affects the mean correlation between children, as shown in Fig. 10(b).

In contrast, α_c decreases almost linearly with increasing the extent of parent input a_p, as shown in Fig. 18(b), but does not go to zero for the highest possible value of $a_p = 1$. As we saw in the Sect. 2.2.5, a_p affects the degree to which children are similar to each of their individual parents. Increasing this parameter increases the similarity between those children receiving input from the same parents, increasing their overall similarity and therefore decreasing their discriminability. In terms of the effect on the correlation distribution, in Fig. 10(a) it can be seen that with increasing a_p, there is an increase in the fluctuations in the correlations, making them more positively skewed.

Finally, α_c initially decreases dramatically with increasing rate of dominance ζ. High values of ζ correspond to only a handful of parents out of the total dominating the activity of the children. For very high ζ, the strongest parents dominate the activity to an extent that those children affected by the strongest parents tend to become increasingly similar. The capacity however, does not go to zero, and stabilizes to a constant value, as the activity corresponding to the strongest parents will be recovered during retrieval dynamics, to the detriment of the activity corresponding to the relatively weaker parents.



(a) (b) (c)

Figure 18. Storage capacity curves as a function of different correlation parameters. **(a)** α_c as a function of f. The full lines correspond to simulations while the dashed line corresponds to solutions of the mean-field equations. It can be seen that similar to Fig. 17(a), the over-estimation of the capacity through the SCSNA approach holds also for other values of f. **(b)** α_c as a function of a_p. **(c)** α_c as a function of ζ. When not explicitly varied, the correlation parameters are $a_p = 0.4$, $f = 0.05$, $\zeta = 10^{-6}$ and $\Pi = 150$. Network parameters are $N = 2000$, $c_m = 200$, $a = 0.1$, $S = 5$.

2.3.4. Residual information: memory beyond capacity

In the previous section we saw that correlations decrease the storage capacity of the network. In particular, in terms of the dominance parameter ζ, what is the configuration that the network settles into? Do these configurations correspond to the activity of the strongest parents? We carried out simulations with correlated patterns for different values of ζ and computed the mutual information between the pattern cued c and the configuration in which the network settles r

$$I(c,r) = \sum_{k,l=0}^{S} C^{kl}(c,r) \log_2 \left(\frac{C^{kl}(c,r)}{C^k(c) C^l(r)} \right), \tag{50}$$

where

$$C^{kl}(c,r) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i^c,k} \sigma_i^l, \tag{51}$$

$$C^k(c) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i^c,k}, \tag{52}$$

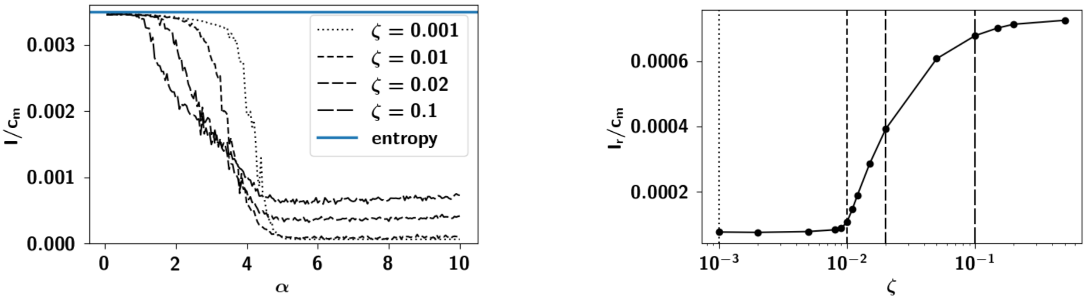
$$C^l(r) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^l\,. \tag{53}$$

The maximum value of this quantity is attained when the cued pattern is also the one retrieved: $c = r$. In this case the mutual information can reach up to

$$I(c) = \sum_{k=0}^{S} C^k(c)\log_2\left(\frac{1}{C^k(c)}\right) = \Big\{-(1-a)\log_2(1-a) + a\log_2(S/a)\Big\}\,, \tag{54}$$

that we recognize to be the entropy of the cued pattern. In Fig. 19(a) we can see the mutual information as a function of the loading α for different values of the parameter ζ, averaged across cued retrieval of many patterns. The mutual information has a sharp fall-off upon increasing α, which is more and more abrupt as ζ decreases. For small values of α, the mutual information does not depend on ζ: its value at this plateau corresponds to the entropy.

The most interesting observation in Fig. 19(a), however, is the *residual information*, its remaining roughly constant value, after capacity collapse. In Fig. 19(b), this residual information is plotted as a function of the dominance rate ζ, and it can be seen that it increases sharply between $\zeta_c \simeq 0.01$ and $\zeta \simeq 0.1$ before saturating at a value still well below the entropy of the stored memory (the initial plateau). This effect is reminiscent of a phase transition with control parameter ζ, where the information plays the role of the order parameter. Below the critical value ζ_c, once the capacity is exceeded, there is no more retrievable information at all. Above ζ_c, however, the network retrieves some information about the cued pattern. In Fig. 20 we plot, as a phase diagram, the residual information as a function of ζ in the x-axis and f in the y-axis. One sees that a non-vanishing residual information requires, essentially, sufficiently large values of both ζ and f. In terms of either parameter, the complete transition between the two regimes spans about an order of magnitude (note the logarithmic scales).



(a) (b)

Figure 19. **(a)** Mutual information per connection as a function of the storage load α, for different values of the dominance ζ. For low values of ζ, the information falls abruptly at a value of the storage load α, while for larger values of ζ, we observe a more gradual decay, starting at lower values of the storage load. For high enough values of ζ however, the information does not go to zero, but rather saturates at a certain value. We call this *residual information*. In **(b)**, we plot this residual information as a function of ζ. The sharp increase of this residual information from $\zeta_c \sim 0.01$ suggests a phase transition separating two regimes. In the first, with $\zeta < \zeta_c$, the residual information is approximately zero. In the second, $\zeta > \zeta_c$, this information is non-zero and indicates that the network, though not able to retrieve the *fine* structure of the memory cued, still manages to retrieve the *gross* structure. Network parameters are $N = 2000$, $c_m = 200$, $S = 5$, $a = 0.1$, $U = 0.5$, $\beta = 200$. Correlation parameters are $a_p = 0.4$, $f = 0.05$ and $\Pi = 150$.

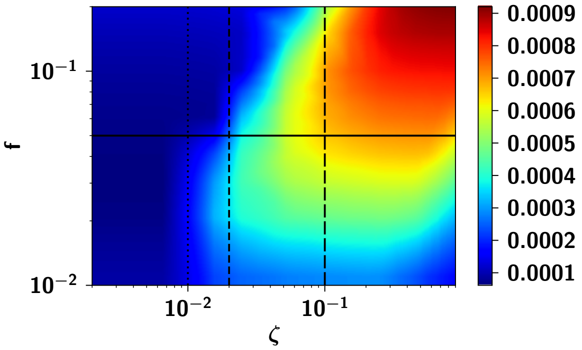


Figure 20. Phase diagram of the residual mutual information between cued memory and configuration retrieved, after capacity collapse, as a function of the dominance ζ in the x-axis and prolificity f in the y-axis, giving a fuller picture of the phase transition of Fig. 19(b). Note that the transition occurs at higher values of ζ with increasing f. The black dashed lines, plotted for clarity, correspond to salient locations of the curve shown in Fig. 19(b).

2.3.5. Residual memory interpreted through cluster analysis

Although it was argued in the introduction that the presence of clusters is only one component of the metric relations embedding semantic memories, it is instructive to interpret the residual memory phase expressed by our model in terms of cluster analysis. Fig. 21 shows the outcome of applying a standard clustering algorithm to patterns generated at salient locations of the phase diagram in Fig. 20. For $f = 0.05$ and $\zeta = 0.001$ (low dominance), the algorithm is essentially unable to identify clear clusters, so that pairs of patterns that end up together in one of the weak clusters forcibly assigned are not markedly more correlated than pairs that end up in different clusters. For higher dominance ($\zeta = 0.02, 0.1$) the clustering structure becomes more real, as indicated by the expanding white area to the left, because a few parents dominate the rest. In the high dominance region, increasing prolificity (to $f = 0.2$) makes the extracted clusters larger and the residual information, while remaining non-zero, decreases, possibly because of the concomitant decrease in the number of clusters. We can conclude, therefore, that the residual information largely expresses resilient memory associated with the distinction between clusters, whereas within-cluster distinctions are lost above the capacity limit. This interpretation remains a qualitative description, in that our pattern generation algorithm does not produce well-defined clusters, but a more complex set of metric relations among patterns, where clusters emerge as one component if a few parents dominate.

2.3.6. Residual memory rides on fine differences in ultrametric content

A complementary perspective is that afforded by our measure of ultrametric content, which is derived from a measure of distance between patterns, applied to all triplets of patterns. A suitable distance measure, for Potts patterns, can be

$$D_{\mu\nu} = C_{a0} + C_{0a} + 2C_{ad}\,, \tag{55}$$

where the quantities in the r.h.s. have been defined in Sect. 2.2. In the dominance-prolificity region we have been considering, this distance measure yields the values seen in Table 1 for the ultrametric content index (see App. C).

Peer-reviewed version available at *Entropy* **2018**, *20*, 824; doi:10.3390/e20110824

Table 1. Ultrametric content computed for distances of triplets of patterns generated by the algorithm, for six different parameter values of the prolificity and the dominance. An increased ultrametric content reflects an increased clustering in the correlations between patterns. For $f = 0.2$ and $\zeta = 0.1$, the patterns yield values of the ultrametric content index close to that obtained from the nouns (~ 0.5). The corresponding clustering structure of the patterns can be seen in Fig. 21(d).

		ζ		
		0.001	0.02	0.1
	0.05	0.395	0.429	0.435
f	0.2	0.389	0.404	0.507

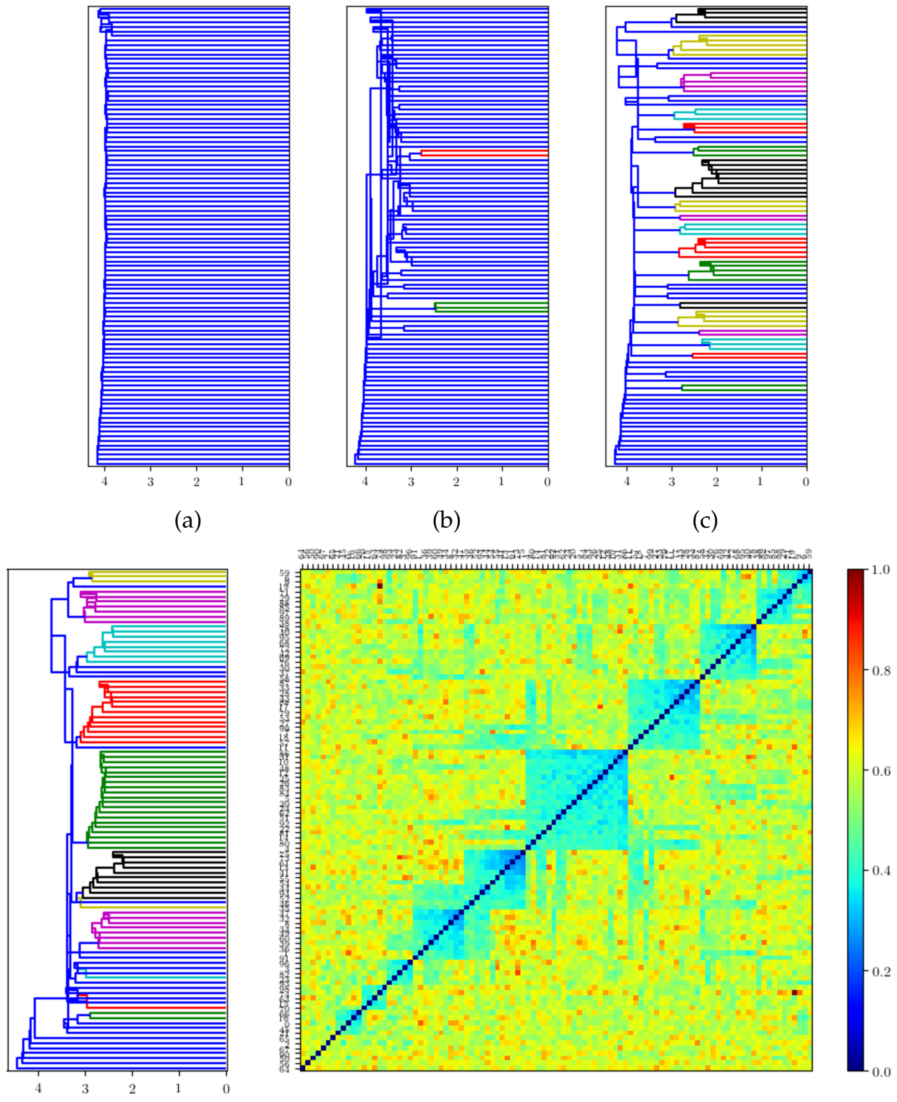
Interestingly, a completely different measure of distance, similar instead to the one extracted from the feature-based norms in the toy example reported in Sect. 1.1, yields values very close to these, within a few percent, when applied to patterns generated with our algorithm. We can see, therefore, that the emergence of residual memory does correlate with increased ultra-metric content, but not in a simple one-to-one correspondence; and the putative phase transition occurs in the midst of a relatively minor increase in the values of the ultrametric content index.

A tentative conclusion is that semantic resilience, at least as crudely modelled by a Potts network, requires a degree of clustering or ultrametric structure, which in the pattern-generation model reflects sufficient values of prolificity and dominance, but is still an emergent phenomenon. Quantitative differences in the parameters produce what tends to be an all-or-none difference in semantic resilience. Yet another form, possibly, of analog-to-digital transform produced by a neural circuit.

3. Discussion: a new model for the extraction of semantic structure

In recent years, the Potts network has been proposed as an effective model of a cortical network organized with distinct local and global connections. Several aspects of the model have been studied in quantitative detail, under many simplifying assumptions, including that of uncorrelated patterns. However, it can be argued that such patterns are irrelevant for the study of semantic memory. The various feats of "mind-reading", achieved with fMRI studies (e.g. [7,42,43]), reflect that correlations between memories are not simply a nuisance that degrades memory capacity, but express the core ability of the cortex as a machine for encoding structured information. We have attempted to make theoretical progress in this direction by designing a plausible algorithm to generate patterns. In our algorithm, the patterns are generated by the contribution of multiple factors, which can be considered as semantic category generators (except that categories overlap and have loose boundaries), or else features in a somewhat wider sense, that carry information on the statistical co-occurrence of attributes. Through competition, those attributes concur, each with its relative strength (parametrized by a_p, f and ζ) to construct the statistical structure of the memories.

We have further studied the storage capacity of the network as a function of both network parameters and correlation parameters through the SCSNA analysis as well as extensive simulations, and we find that with a Hebbian rule for the storage of patterns, the network can store and retrieve *fewer* correlated patterns, though still of order CS^2/a, yielding $\sim 10^7$ with human cortical parameters [29]. Other prescriptions for learning, enhancing capacity, may be explored and studied, and we leave such studies for future investigations. Of the correlation parameters, the effect of the dominance ζ is particularly interesting. $\zeta \sim 0$ corresponds to a situation in which all parents are on equal footing, while the opposite limit corresponds to only a handful dominating the rest. For large enough values of ζ, we observe *correlated retrieval*, in that with the decrease in successful retrieval, the fraction of trials in which another, correlated pattern is retrieved, increases. In terms of the mutual information between the cued pattern and the final configuration of the network, after retrieval dynamics, we observe that it does not always go to zero: it can stabilize at a roughly constant value, after the capacity limit has been reached. We call it the *residual information*. The residual information displays a nontrivial dependence



(a)

(b)

(c)

(d)

Figure 21. Cluster analysis applied to patterns generated by the algorithm, for parameter values where we do not observe semantic resilience ($\zeta = 0.001$) and where we do observe it, $\zeta = 0.02, 0.1$ (see also Fig. 19(b), Fig. 20). The top row has been obtained with $f = 0.05$, and clusters are seen to be smaller and numerous, while their average correlation increases with ζ. The bottom row instead shows fewer and larger clusters of the patterns obtained with $\zeta = 0.1$, $f = 0.2$, for which we observe semantic resilience, as well as the corresponding distance matrix, obtained through the correlation matrix C_{as}, as explained in App. C.

on the dominance ζ: for $\zeta < \zeta_c$, this residual information is approximately zero. At ζ_c, there is a sharp increase in the residual information, which saturates at large values of ζ, when only a handful of parents become relevant. This sharp increase is reminiscent of a phase transition, in which the residual information is the order parameter and ζ is the control parameter. The residual information has an interesting interpretation: it can be thought of as the information pertaining to the *gross*, core semantic component of the memories, after the *fine* details have been compromised. Note that $1/\zeta$ is a measure of the number of parents/factors/attributes that effectively dominate semantic space.

Taken at face value, the diminished capacity of the Potts network accompanied by the emergence of residual information suggests that this ability for generalization comes at the cost of losing the resolution with which we can retrieve the individual memories. However, this result as such is incomplete, and must be considered also in relation to the differential role of other memory structures and in particular the hippocampus in retrieval. For example, in humans, it has been shown that the ventral hippocampus projects directly to the medial prefrontal cortex, providing an immediate route for representations to reach the prefrontal cortex, suggesting a model of bidirectional hippocampus/prefrontal cortex interactions that support context-dependent memory retrieval [44]. Several studies have attempted to dissociate between the contributions of the hippocampus and the cortex in human memory retrieval. In particular, in [4] and [45] it was found that putatively different access modes to information stored in long-term memory in a remember/know paradigm lead to different distributions of classification errors of different groups with memory disorders. An information derived measure, the metric content, quantifying the concentration of errors was computed: high levels of metric content are indicative of a strong dependence on perceived relations among the set of stimuli, and therefore of a relatively preferred semantic access mode, while low levels (and similar correct performance), suggest a preferential episodic access mode. It was found that compared with normal controls, the metric content index was increased in patients with Alzheimer's disease, decreased in patients with herpes encephalitis, and unvaried in patients with damage to the prefrontal cortex. Moreover, a significant correlation between the metric content and measures quantifying episodic and semantic retrieval mode in the remember/know paradigm introduced by Tulving [46] was found. If we think of the access modes, to a first approximation, as reflecting a stronger reliance on specific memory structures, the distribution of errors may then be a window into understanding their relative contributions. Within this larger picture, a cortical impairment as modelled for example by a Potts network with reduced connectivity may be somewhat mitigated by a complementary episodic mode of access, supported by other structures.

Our finding of semantic resilience, as characterized by the residual information has an interesting interpretation also in relation to the findings in the neuropsychology of semantic dementia. A typical finding is that the finer-grained or "subordinate" aspects of such patients' knowledge are more susceptible to damage than the more "ordinate" aspects, for example in naming tasks. Moreover, the naming errors that such patients make tend to change in time from "circumlocutions to category coordinates to superordinate labels" [47]. It has been argued that such a finding is in favor of tree-like models of semantic knowledge, in which the mental representation of a concept occupies nodes in a branching tree, where the origin of the tree corresponds to its most general and the periphery to its most selective designation. Subsequent research however, pointed to findings that could not be explained through a tree model, such as verification latency [48] or typicality effects [49], or others which question where in the tree to store concepts that belong to more than one category [50]. In our account, instead, semantic resilience is an outcome that emerges naturally through higher values of the dominance parameter ζ, in which finer-grained or subordinate features of the concepts are overtaken by ordinate features, which then become the only retrievable ones, as shown through the cluster analysis in Fig. 21. Crucially though, such clusters emerge only when our dominance parameter becomes large enough, and they are neither well-defined nor designed to have strict boundaries. The non-trivial behavior of the residual information, i.e. its phase transition with the dominance parameter

ζ, cannot be predicted from a qualitative model encoding uncorrelated patterns. Our model offers one plausible way in which such resilience emerges.

Finally, our account may have implications for the question of how the cortex extracts and encodes the general statistical structure of the ensemble of stimuli that it receives. There is a well-established view of the cortex as a slow memory system that uses overlapping distributed representations to represent the general statistical structure of the environment. It has been suggested that the interaction between the hippocampus and the cortex is a crucial element in the consolidation of memories. The general idea is that memories are first stored in the hippocampal system via synaptic changes and that these support the reinstatement of recent memories in the neocortex. Neocortical synapses are slightly modified on each reinstatement and the gradual, neocortical changes accumulating over time encode remote memory. This division of labor would allow the hippocampus to rapidly encode new episodic items without disrupting semantic memories, and the cortex to slowly integrate them in a structured fashion into such memories.

This view is consistent with evidence that damage to the hippocampal system results in recent memory disruption but leaves remote memory intact, but it does not really specify what makes the consolidation process gradual or slow [51]. Early modelling attempts typically resorted to backpropagation to account for the structured learning of the cortex [52]: consolidation would then be slow, because backpropagation is effective with low learning rates. However, backpropagation has been widely criticized on the basis that it lacks a plausible biological mechanism. While hippocampal learning in these accounts was taken to fit the framework of learning unrelated patterns of activity, it had remained unclear how to model neocortical learning. Our account offers an alternative framework for neocortical learning, in which semantic structure is extracted progressively from the statistics of generating features and encoded in the cortex via Hebbian learning and is resilient, i.e. it is preserved when the storage capacity for "episodic details" is exceeded.

Author Contributions: VB and AT conceived and designed the study, which were performed primarily by VB and AT, with contributions by RB. VB and AT wrote the paper, with input from RB.

Funding: This research was supported by Human Frontier Science Program grant RGP0057/2016.

Acknowledgments: The authors would like to thank Alberto Pezzotta for very fruitful suggestions. Discussions with Michelangelo Naim and Chol-Jun Kang are gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Calculation of the probability distribution of the field for $S = 1$

In this section we outline the main steps in deriving Eq. (24). The distribution for h_μ can be computed by making use of the probability generating function

$$G(s) \equiv \mathcal{L}\{P(x = \xi_i^\pi)\} = \int_0^\infty dx\, P(x)\, e^{-sx} = \frac{a_p}{s}(1 - e^{-s}) + (1 - a_p)\,. \tag{A1}$$

Since the ξ_i^π are identically and independently distributed for all π, we can use the following property

$$P(h_\mu \,|\, n_p) = \mathcal{L}^{-1}\{G(s)^{n_p}\}\,. \tag{A2}$$

The number of parents as well as the total field received by a pattern is i.i.d, so we drop the index μ for h_μ. We can compute the conditional distribution of the field received by a unit, given the number of parents, as

$$P(h \,|\, n_p) = \lim_{\gamma\to\infty} \frac{1}{2\pi i} \int_{c-i\gamma}^{c+i\gamma} ds \{1 - a_p + \frac{a_p}{s}(1 - e^{-s})\}^{n_p}\, e^{sh}\,. \tag{A3}$$

Using the binomial theorem

$$\{1 - a_p + \frac{a_p}{s}(1 - e^{-s})\}^{n_p} = \sum_{k=0}^{n_p} \binom{n_p}{k} (1 - a_p)^{n_p - k} \frac{a_p^k}{s^k} \sum_{j=0}^{k} (-1)^j e^{-sj} , \tag{A4}$$

$$P(h \,|\, n_p)(h) = \lim_{\gamma \to \infty} \frac{1}{2\pi i} \int_{c-i\gamma}^{c+i\gamma} \sum_{k=0}^{n_p} \sum_{j=0}^{k} (-1)^j \binom{n_p}{k} \binom{k}{j} (1 - a_p)^{n_p - k} \frac{a_p^k}{s^k} e^{s(h-j)} ds \tag{A5}$$

We can carry out the integral to find

$$I(k = 0) = \delta(h) , \tag{A6}$$

$$I(k \geq 1) = \frac{(h - j)^{k-1}}{(k - 1)!} . \tag{A7}$$

The distribution of the field h for a given number of parents n_p is then

$$P(h \,|\, n_p) = (1 - a_p)^{n_p} \delta(h) \, + \sum_{k=1}^{n_p} \sum_{j=0}^{k} \frac{(-1)^j \, n_p! \, a_p^k \, (1 - a_p)^{n_p - k}}{(n_p - k)!(k - j)! j! (k - 1)!} \, (h - j)^{k-1} \, \Theta(h - j) . \tag{A8}$$

The first term in this equation expresses the fact that the only way to get zero field is if all n_p parents contribute zero field and this occurs with probability $(1 - a_p)^{n_p}$. For a given pattern μ, with n_p parents, the field of each unit is distributed according to Fig. 6. The cumulative distribution function writes

$$P(h' < h \,|\, n_p) = \int_{-\infty}^{h} dh' \, P(h' \,|\, n_p) \tag{A9}$$

$$= (1 - a_p)^{n_p} \Theta(h) + \sum_{k=1}^{n_p} \sum_{j=0}^{k} \frac{(-1)^j \, n_p! \, a_p^k \, (1 - a_p)^{n_p - k}}{(n_p - k)!(k - j)! j! k!} \, (h - j)^k \, \Theta(h - j) . \tag{A10}$$

The minimal threshold h_m is implicitly given by the cumulative probability

$$P(h' < h_m \,|\, n_p) = 1 - a . \tag{A11}$$

540 **Appendix B. Calculation of the probability distribution of the field for $S = 2$**

To derive Eq. (28), we start with the joint distribution of number of parents by state

$$P(\hat{n}^1 = n^1, ..., \hat{n}^S = n^S) = \frac{n_p!}{S^{n_p} \prod_{k=1}^{S} n^k!} . \tag{B1}$$

Note that we define the field to be identically distributed across states. The probability that the fields of all states are below that of the first is given by

$$P(h = h^1) = \int_0^h P(h^1, ..., h^S) \prod_{k=2}^{S} dh^k . \tag{B2}$$

The probability distribution of the maximal field is given by S times the one above

$$P(h_{max}) = S \int_0^{h^1} P(h^1, ..., h^S) \prod_{k=2}^{S} dh^k . \tag{B3}$$

The joint distribution of the fields across states writes

$$P(h^1, ..., h^S | n_p) = \frac{n_p!}{S^{n_p}} \prod_{k=1}^{S} \sum_{n^k=1}^{n_p} \frac{P(h^k | n^k)}{n^k!} \, \delta_{n_p, \sum\limits_{k=0}^{S} n^k} \,, \tag{B4}$$

where the constraint $n_p = \sum\limits_{k=0}^{S} n^k$ has been included in the last line. $P(h^k|n^k)$ is given by Eq. (A8), replacing n_p with n^k. We then have

$$P(h^1, ..., h^S | n_p) = \frac{n_p!}{S^{n_p}} \prod_{k=1}^{S} \sum_{n^k=1}^{n_p} \left\{ \frac{(1-a_p)^{n^k}}{n^k!} \, \delta(h^k) + \right. \tag{B5}$$

$$\left. + \sum_{i=1}^{n^k} \sum_{j=0}^{i} (-1)^j \frac{(a_p)^i \, (1-a_p)^{n^k - i}}{(n^k - i)!(i-j)!j!} \, \frac{(h^k - j)^{i-1}}{(i-1)!} \Theta(h^k - j) \right\} \delta_{n_p, \sum\limits_{k=0}^{S} n^k} \,. \tag{B6}$$

For $S = 1$ all contributions go to a single state, so we automatically have $n^1 = n_p$, then the first sum disappears and we fall back onto Eq. (A8). For $S = 2$ we have, denoting the state receiving the maximal field by H,

$$P(H \,|\, n_p) = \frac{n_p!}{2^{n_p - 1}} \sum_{n^1=1}^{n_p} \left\{ \frac{(1-a_p)^{n^1}}{n^1!} \delta(H) + \sum_{i=1}^{n^1} \sum_{j=0}^{i} \frac{(-1)^j (a_p)^i \, (1-a_p)^{n^1 - i}}{(n^1 - i)!(i-j)!j!(i-1)!} \, (H-j)^{i-1} \, \Theta(H-j) \right\}$$

$$\left\{ \frac{(1-a_p)^{n_p - n^1}}{(n_p - n^1)!} \Theta(H) + \sum_{i'=1}^{n_p - n^1} \sum_{j'=0}^{i'} \frac{(-1)^{j'} (a_p)^{i'} \, (1-a_p)^{n_p - n^1 - i'}}{(n_p - n^1 - i')!(i'-j')!j'!i'!} \, (H - j')^{i'} \, \Theta(H - j') \right\} , \tag{B7}$$

where we drop the indices denoting the units (they are drawn from the same distribution). Note that the state does not appear in this expression because it is the distribution for the state that receives maximal input, regardless of which one it is. The μ dependence is through $n_p = n_p(\mu)$. We then get the minimal threshold for activation H_m implicitly in terms of the cumulative distribution

$$P(H' < H_m \,|\, n_p) = \int_{-\infty}^{H_m} P(H' \,|\, n_p) \, dH' = 1 - a \,. \tag{B8}$$

We can compute it to find

$$P(H' < H \,|\, n_p) = \frac{n_p!}{2^{n_p - 1}} \sum_{n^1=1}^{n_p} \left\{ \frac{(1-a_p)^{n_p}}{n^1!(n_p - n^1)!2} \right.$$

$$+ \frac{(1-a_p)^{n_p - n^1}}{(n_p - n^1)!} \sum_{i=1}^{n^1} \sum_{j=0}^{i} \frac{(-1)^j (a_p)^i \, (1-a_p)^{n^1 - i}}{(n^1 - i)!(i-j)!j!i!} (H-j)^i \, \Theta(H-j) \tag{B9}$$

$$\left. + \sum_{i'=1}^{n_p - n^1} \sum_{j'=0}^{i'} \frac{(-1)^{j'} (a_p)^{i'} \, (1-a_p)^{n_p - n^1 - i'}}{(n_p - n^1 - i')!(i'-j')!j'!i'!} \sum_{i=1}^{n^1} \sum_{j=0}^{i} \frac{(-1)^j (a_p)^i \, (1-a_p)^{n^1 - i}}{(n^1 - i)!(i-j)!j!(i-1)!} \, I(H, i, i', j, j') \right\} .$$

where $max\{j, j'\} = j^*$

$$
\begin{aligned}
I(H, i, i', j = j') &= \frac{(H - j)^{i+i'}}{i + i'} \Theta(H - j) \\
I(H, i, i', j \neq j') &= \\
= &\begin{cases} i'!(i-1)! \left[\sum\limits_{q=0}^{i'} (-1)^q \frac{(H-j)^{i+q}(H-j')^{i'-q}}{(i+q)!(i'-q)!} \Theta(H - j^*\}) + (-1)^{i'+1} \frac{(j^* - j)^{i+i'}}{(i+i')!} \right] & i - 1 \geq i' \\ i'!(i-1)! \left[\sum\limits_{q=0}^{i-1} (-1)^q \frac{(H-j)^{i-q-1}(H-j')^{i'+q+1}}{(i-q-1)!(i'+q+1)!} \Theta(H - j^*) + (-1)^{i} \frac{(j^* - j)^{i+i'}}{(i+i')!} \right] & i - 1 < i' \end{cases}
\end{aligned}
\tag{B10}
$$

Appendix C. Ultrametric content

A possible characterization of the correlations between the memory patterns is in terms of a distance. A quasi-distance measure can be derived from the correlation following the same procedure as in [10]. We first define a so-called "confusion" matrix

$$P(\mu|\nu) = \frac{C_{\mu\nu}}{\sum\limits_{\mu=0}^{p} C_{\mu\nu}} . \tag{C1}$$

where $C_{\mu\nu}$ is an element of the correlation matrix and where P, the confusion matrix, is obtained by normalizing each element of the correlation matrix appropriately. Next, we symmetrize the above function to obtain

$$d(\mu, \nu) = -\log \left(\frac{P(\nu|\mu)P(\mu|\nu)}{P(\mu|\mu)P(\nu|\nu)} \right) , \tag{C2}$$

a quasi-distance, in the sense that it satisfies only the reflective and symmetric properties, $d(\mu, \mu) = 0$ and $d(\mu, \nu) = d(\nu, \mu)$. The triangular inequality $d(\mu, \nu) + d(\mu, \rho) \leq d(\mu, \rho)$ does not necessarily hold. It can be made to hold by raising d to a sufficiently small power $d \to d^{1/p}$, called the "trivialization" of d, as explained in detail in [10]. Using this procedure, distances between triplets of patterns $\{\mu, \nu, \rho\}$ can be computed. If we note by d_{min} the edge of minimal length, d_{max} the edge of maximal length and d_{med} the edge of intermediate length, then we can plot, in a two-dimensional graph, the ratios $\delta_1 = d_{min}/d_{max}$ and $\delta_2 = d_{med}/d_{max}$.

Triplets that satisfy the triangular inequality lie above the line $\delta_1 = 1 - \delta_2$, while triplets that satisfy the ultrametric inequality lie on the vertical line where $\delta_2 = 1$. Among these, triplets that are equilateral triangles lie at the point $\delta_1 = \delta_2 = 1$. To measure the overall closeness of the cloud of triplets to the fully ultrametric limit one can define the *ultrametric content*

$$\lambda_{um} = \left\langle \frac{\log \delta_1 - \log \delta_2}{\log \delta_1 + \log \delta_2} \right\rangle \tag{C3}$$

where $\langle \cdot \rangle$ denotes the mean over all triplets. This quantity does not depend on the trivialization of d and it ranges from 0 (for triplets forming isosceles triangles with two short sides) to 1 (for a fully ultrametric set: equilateral triangles and isosceles triangles with two long sides).

1. Yonelinas, A.P. The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language* **2002**, *46*, 441–517.
2. Treves, A.; Rolls, E.T. Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus* **1992**, *2*, 189–199.

3. Alme, C.B.; Miao, C.; Jezek, K.; Treves, A.; Moser, E.I.; Moser, M.B. Place cells in the hippocampus: eleven maps for eleven rooms. *Proceedings of the National Academy of Sciences* **2014**, *111*, 18428–18435.
4. Lauro-Grotto, R.; Ciaramelli, E.; Piccini, C.; Treves, A. Differential impact of brain damage on the access mode to memory representations: an information theoretic approach. *European Journal of Neuroscience* **2007**, *26*, 2702–2712.
5. Huth, A.G.; Nishimoto, S.; Vu, A.T.; Gallant, J.L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **2012**, *76*, 1210–1224.
6. Huth, A.G.; de Heer, W.A.; Griffiths, T.L.; Theunissen, F.E.; Gallant, J.L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **2016**, *532*, 453–458.
7. Mitchell, T.M.; Shinkareva, S.V.; Carlson, A.; Chang, K.M.; Malave, V.L.; Mason, R.A.; Just, M.A. Predicting human brain activity associated with the meanings of nouns. *Science* **2008**, *320*, 1191–1195.
8. Collins, A.M.; Quillian, M.R. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior* **1969**, *8*, 240–247.
9. Warrington, E.K. The selective impairment of semantic memory. *The Quarterly journal of experimental psychology* **1975**, *27*, 635–657.
10. Treves, A. On the perceptual structure of face space. *BioSystems* **1997**, *40*, 189–196.
11. Parga, N.; Virasoro, M.A. The ultrametric organization of memories in a neural network. In *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*; World Scientific, 1987; pp. 436–443.
12. Gutfreund, H. Neural networks with hierarchically correlated patterns. *Physical Review A* **1988**, *37*, 570–577.
13. Franz, S.; Amit, D.J.; Virasoro, M.A. Prosopagnosia in high capacity neural networks storing uncorrelated classes. *Journal de Physique* **1990**, *51*, 387–408.
14. Virasoro, M.A. Categorization in neural networks and prosopagnosia. *Physics Reports* **1989**, *184*, 301–306.
15. Shallice, T.; Cooper, R. *The organisation of mind*; Oxford University Press, 2011.
16. Rumelhart, D.E.; Hinton, G.E.; McClelland, J.L.; others. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition* **1986**, *1*, 45–76.
17. Farah, M.J.; McClelland, J.L. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of experimental psychology: General* **1991**, *120*, 339.
18. Plaut, D.C. Semantic and associative priming in a distributed attractor network. Proceedings of the 17th annual conference of the cognitive science society, 1995, Vol. 17, pp. 37–42.
19. Rogers, T.T.; Lambon Ralph, M.A.; Garrard, P.; Bozeat, S.; McClelland, J.L.; Hodges, J.R.; Patterson, K. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review* **2004**, *111*, 205.
20. Hellwig, B. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological cybernetics* **2000**, *82*, 111–121.
21. Roudi, Y.; Treves, A. An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment* **2004**, *2004*, P07010.
22. Pucak, M.L.; Levitt, J.B.; Lund, J.S.; Lewis, D.A. Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *Journal of Comparative Neurology* **1996**, *376*, 614–630.
23. Braitenberg, V.; Schüz, A. *Anatomy of the cortex: statistics and geometry*; Vol. 18, Springer Science & Business Media, 1991.
24. O'Kane, D.; Treves, A. Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General* **1992**, *25*, 5055.
25. O'Kane, D.; Treves, A. Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Computation in Neural Systems* **1992**, *3*, 379–384.
26. Mari, C.F.; Treves, A. Modeling neocortical areas with a modular neural network. *Biosystems* **1998**, *48*, 47–55.
27. Dubreuil, A.M.; Brunel, N. Storing structured sparse memories in a multi-modular cortical network model. *Journal of computational neuroscience* **2016**, *40*, 157–175.
28. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. Advances in Neural Information Processing Systems, 2017, pp. 3859–3869.
29. Naim, M.; Boboeva, V.; Kang, C.J.; Treves, A. Reducing a cortical network to a Potts model yields storage capacity estimates. *Journal of Statistical Mechanics: Theory and Experiment* **2018**, *2018*, 043304.

30. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* **1982**, *79*, 2554–2558.
31. Tsodyks, M.V.; Feigel'Man, M.V. The enhanced storage capacity in neural networks with low activity level. *EPL (Europhysics Letters)* **1988**, *6*, 101.
32. Kropff, E.; Treves, A. The storage capacity of Potts models for semantic memory retrieval. *Journal of Statistical Mechanics: Theory and Experiment* **2005**, *2005*, P08010.
33. Mézard, M.; Parisi, G.; Virasoro, M.A. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*; Vol. 9, World Scientific Publishing Co Inc, 1987.
34. Amit, D.J. *Modeling brain function: The world of attractor neural networks*; Cambridge University Press, 1992.
35. Treves, A. Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive Neuropsychology* **2005**, *22*, 276–291.
36. Sartori, G.; Lombardi, L. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience* **2004**, *16*, 439–452.
37. Osgood, C.E. Semantic differential technique in the comparative study of cultures. *American Anthropologist* **1964**, *66*, 171–200.
38. Löwe, M. On the storage capacity of Hopfield models with correlated patterns. *The Annals of Applied Probability* **1998**, *8*, 1216–1250.
39. Engel, A. Storage capacity for hierarchically correlated patterns. *Journal of Physics A: Mathematical and General* **1990**, *23*, L285.
40. Shiino, M.; Fukai, T. Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Physical Review E* **1993**, *48*, 867.
41. Kropff, E. Full solution for the storage of correlated memories in an autoassociative memory. *Computational Modelling in Behavioural Neuroscience: Closing the Gap Between Neurophysiology and Behaviour* **2009**, *2*, 225.
42. Haxby, J.V.; Gobbini, M.I.; Furey, M.L.; Ishai, A.; Schouten, J.L.; Pietrini, P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **2001**, *293*, 2425–2430.
43. Norman, K.A.; Polyn, S.M.; Detre, G.J.; Haxby, J.V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences* **2006**, *10*, 424–430.
44. Preston, A.R.; Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology* **2013**, *23*, R764–R773.
45. Ciaramelli, E.; Lauro-Grotto, R.; Treves, A. Dissociating episodic from semantic access mode by mutual information measures: evidence from aging and Alzheimer's disease. *Journal of Physiology-Paris* **2006**, *100*, 142–153.
46. Tulving, E. Episodic memory: from mind to brain. *Annual review of psychology* **2002**, *53*, 1–25.
47. Garrard, P.; Perry, R.; Hodges, J.R. Disorders of semantic memory. *Journal of Neurology, Neurosurgery, and Psychiatry* **1997**, *62*, 431.
48. Conrad, C. Cognitive economy in semantic memory. **1972**.
49. Rosch, E.; Mervis, C.B. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology* **1975**, *7*, 573–605.
50. Spivey, M.; Joanisse, M.; McRae, K. *The Cambridge handbook of psycholinguistics*; Cambridge University Press, 2012.
51. Kitamura, T.; Ogawa, S.K.; Roy, D.S.; Okuyama, T.; Morrissey, M.D.; Smith, L.M.; Redondo, R.L.; Tonegawa, S. Engrams and circuits crucial for systems consolidation of a memory. *Science* **2017**, *356*, 73–78.
52. McClelland, J.L.; McNaughton, B.L.; O'Reilly, R.C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* **1995**, *102*, 419.