Scaling Laws of Associative Memory Retrieval

Sandro Romani
sandro.romani@gmail.com
Weizmann Institute, Department of Neurobiology, Rehovot, Israel 76100;
Center for Theoretical Neuroscience, Columbia University, New York,
NY 10032, U.S.A.; and Donders Centre for Neuroscience, Radboud
University, 6525 Nijmegen, The Netherlands

Itai Pinkoviezky
itai.pinkoviezky@weizmann.ac.il
Alon Rubin
alon.rubin@weizmann.ac.il
Weizmann Institute, Department of Neurobiology, Rehovot, Israel 76100

Misha Tsodyks
misha@weizmann.ac.il
Weizmann Institute, Department of Neurobiology, Rehovot, Israel 76100,
and Center for Theoretical Neuroscience, Columbia University, New York,
NY 10032, U.S.A.

Most people have great difficulty in recalling unrelated items. For example, in free recall experiments, lists of more than a few randomly selected words cannot be accurately repeated. Here we introduce a phenomenological model of memory retrieval inspired by theories of neuronal population coding of information. The model predicts nontrivial scaling behaviors for the mean and standard deviation of the number of recalled words for lists of increasing length. Our results suggest that associative information retrieval is a dominating factor that limits the number of recalled items.

1 Introduction

Human long-term memory contains many thousands of words, pictures, episodes, and other types of information, yet retrieving this information is challenging when no specific cues are available. Real-life examples of this difficulty are abundant (e.g., recalling all the talks attended at the last

Society for Neuroscience meeting). Memory performance assessed with forced-choice recognition was reported to be good even with lists of up to 1000 words (Standing, 1973), suggesting that recall is limited by the retrieval process rather than true forgetting.

In the neural network paradigm of associative memory, the issue of storage capacity was extensively analyzed in the framework of attractor neural networks (Amit, 1989). Memories in these networks are stored by strengthening the connections between specific groups of neurons, resulting in stable activity patterns that encode different items (attractors; Hopfield, 1982). Retrieval of an item from memory is mediated by the reactivation of the corresponding attractor (Gelbard-Sagiv, Mukamel, Harel, Malach, & Fried, 2008). Only retrieval with memory-specific cues was considered, when the network is initialized in the activity state close to an attractor representing the item being retrieved. In this study, we develop a model of associative retrieval to explore the mechanisms of memory recall and its capacity without specific cues. In the psychological literature, it was reported that the number of items retrieved is a power function of the total number of items of this type in memory (Murray, Pye, & Hockley, 1976). Sublinear scaling was observed when lists of words were presented to the subject for subsequent free recall (Binet & Henri, 1894; Roberts, 1972; Standing, 1973; Murray et al., 1976). Surprisingly, similar scaling was observed when subjects were instructed to recall words based on graphemic cues (i.e., all the words that begin with particular letter; Murray, 1975). We therefore considered the retrieval capacity (RC) of the system—the average number of memory items that can be retrieved when the total number of items in memory is large.

We first analyzed an attractor neural network model in which memorized items are stored by the set of attractors that encode different items (Hopfield, 1982; Amit, 1989). To mimic a choice of unrelated words presented in recall experiments, we considered random uncorrelated groups of neurons representing different words. In a network of N neurons, each group on average includes fN neurons, where f is a parameter between 0 and 1 that defines the sparseness of representation. If no specific cues are presented, we assume that the items are retrieved associatively, with each currently retrieved item acting as a trigger for a next one.

In one of the network implementations, transitions between attractors are caused by periodic modulation of inhibition in the network. When inhibition is low, each attractor state is stable. When the inhibition is increased, the number of active neurons drops, and the network converges to a state in which the intersection of two attractors is active. On the subsequent cycle, the new attractor emerges. Neuronal adaptation prevents the network from falling back into the preceding attractor (see Russo, Namboodiri, Treves, & Kropff, 2008).

The main focus of our study is a mathematical abstraction of the network model that illustrates the associative nature of retrieval of information from memory and allows the analytical exploration of its consequences.

2 Materials and Methods

2.1 Network Model. We simulated a binary Hopfield network with sparse coding (Tsodyks & Feigel'Man, 1988) endowed with global inhibition and neuronal adaptation. The network state is characterized by binary neuronal activities $V_i(t) = 0, 1$, with index $i = 1, \ldots, N$ labeling different neurons, each having threshold th_i. Connections between neurons are determined by the synaptic matrix J_{ij} that is defined following (Tsodyks & Feigel'Man, 1988),

$$J_{ij} = \frac{1}{Nf(1-f)} \sum_{\mu=1}^{L} \left(\xi_i^{\mu} - f \right) \left(\xi_j^{\mu} - f \right), \tag{2.1}$$

where the sum is over L randomly chosen binary patterns ξ_i^{μ} with sparseness f, representing different memory items.

The synchronous evolution of binary neuronal activities $V_i(t) = 0, 1$ is described by the following equation,

$$V_i(t+1) = \Theta \left(\sum_{j=1}^{N} J_{ij} V_j(t) - \frac{J_0}{Nf} \sum_{j=1}^{N} V_j(t) - th_i \right), \tag{2.2}$$

where Θ is the Heaviside function and J_0 is the amplitude of global feedback inhibition that controls the overall level of activity. To make the network dynamics robust, we chose thresholds th_i uniformly distributed on the interval $[-T, T]$.

The stored patterns $V_i = \xi_i^{\mu}$ (Nf neurons active on average) are stable attractors of the network when J_0 falls within the range (see section A.1)

$$T - f < J_0 < 1 - T - f. \tag{2.3}$$

Interestingly, intersections between pairs of patterns $V_i = \xi_i^{\mu} \xi_i^{\nu}$ (Nf^2 neurons active on average) become stable attractors at higher values of J_0 in the range (see section A1)

$$1 - 2f + \frac{T}{f} < J_0 < 2 - 2f - \frac{T}{f}. \tag{2.4}$$

We next generate a sequence of patterns by allowing J_0 to oscillate between these two regimes with period T_{J_0}. To prevent the network from falling back to the previous pattern, we add the activity-dependent threshold adaptation rule:
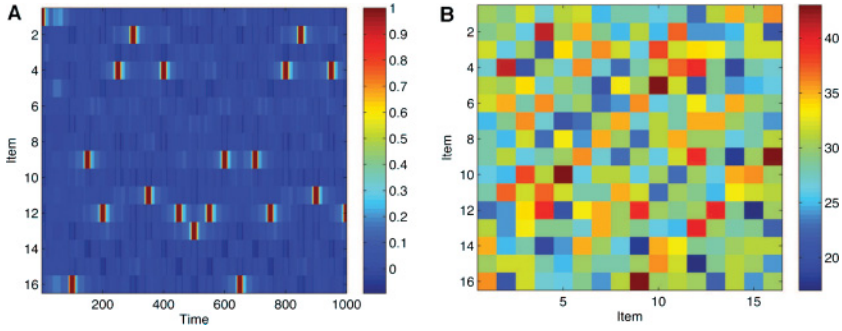
Figure 1: Neural network implementation of associative retrieval. (A) Dynamics of a network storing $L = 16$ items, with periodically modulated feedback inhibition and threshold adaptation. Reactivation strengths (order parameters; see the text) for items stored in the network are color-coded; red denotes network convergence to a specific item. (B) The matrix of overlaps between items. Each entry value (color-coded) is the number of neurons encoding both of the corresponding items. For graphical purposes, all diagonal elements are artificially set at the average value of nondiagonal elements (see the parameters in the text).

$$th_i(t + 1) = th_i(t) - \frac{th_i(t) - th_i(0) - D_{th}V_i(t)}{T_{th}}, \tag{2.5}$$

where D_{th} is the amplitude of the threshold adaptation and T_{th} is the timescale of adaptation.

To characterize the state of the network, we define the following set of order parameters,

$$m^\mu(t) = \frac{1}{Nf(1 - f)} \sum_{i=1}^{N} (\xi_i^\mu - f)V_i(t); \quad \mu = 1, \dots, L, \tag{2.6}$$

where $m^\mu = 0$ means that the network state is uncorrelated with the corresponding pattern and $m^\mu = 1$ means that the network is at the state $V_i = \xi_i^\mu$—that is, that the item μ is retrieved. When the network state is at the intersection of patterns μ and ν, $m^\mu = m^\nu = f$.

The parameters used for the simulation in Figure 1 are $N = 3000$, $L = 16$, $f = 0.1$, $T = 0.015$, $D_{th} = 1.9T$, and $T_{th} = 45$. J_0 oscillates between the values 0.7 and 1.2 with a period of $T_{J0} = 25$.

2.2 Power Law Fit of Mean and Standard Deviation of Number of Recalled Words.

We obtained data from experiment III of Murdock (1960). Lists of words of varying length were presented aurally to human subjects of

either sex at a rate of one word every 2 s. Lists were of $L = 5, 6, 8, 10, 15, 30, 45,$ $60, 75, 100, 150,$ and 200 words in length, and each list was presented to $N =$ $10, 17, 15, 24, 14, 64, 19, 29, 25, 15, 13,$ and 17 subjects, respectively. A detailed description of the experiment can be found in the original work (Murdock, 1960). The estimates of the power law exponents reported below were obtained by numerically minimizing the sum of weighted square errors, separately for the mean and standard deviation of the number of recalled words, using the $a \times L^\alpha$ functional form. For the power law fit of the average number of recalled words, the inverse of their squared standard error was used as weights. For the standard deviation, weights were defined as $\frac{2(N-1)}{s^2}$ (Evans, Hastings, & Peacock, 2000), where N is the number of subjects and s the estimated standard deviation of the number of recalled words. We discarded data corresponding to list lengths of 5, 6, and 8 words in order to minimize the contribution of the fixed-length recency or primacy effects to the number of recalled words (Murdock, 1962). Confidence intervals (CI) for the estimated parameters were obtained as the 2.5th and 97.5th percentiles of bootstrap parameter distributions over 2×10^4 resamplings with replacement of the original data.

3 Results

We simulated a network storing $L = 16$ items, represented by random overlapping groups of neurons. The network repeatedly switches between stored patterns and intersections of two patterns. After several cycles, the dynamics enters a loop in which the same items are retrieved repeatedly. As a result, only 8 of the 16 items are retrieved by the network (see Figure 1A). This partial retrieval is not a consequence of a limited storage capacity, which for this network is much higher than 16 items—all the items can be readily retrieved when cued individually. We then computed the number of neurons in the intersection of patterns representing each pair of items (see Figure 1B). Inspecting Figures 1A and 1B reveals that network transitions typically occur between attractors with the largest intersection. This also holds true when transitions between attractors are induced by short-term synaptic depression alone (Bibitchkov, Herrmann, & Geisel, 2002; Pantic, Torres, Kappen, & Gielen, 2002; results not shown). Hence, for each pair of items, we consider the size of the overlap between the corresponding representations to be a measure of similarity between them. When an item is retrieved, it triggers the subsequent retrieval of the item with which it has the largest similarity. Different realizations of the groups of neurons coding memory items will result in a different number of items retrieved. The aim of our analysis is therefore to characterize the statistical properties of retrieval, such as the distribution of the number of items retrieved. To this end, we consider an ensemble of lists of L items, each endowed with a particular realization of an $L \times L$ matrix of similarities S_{ij} between items.
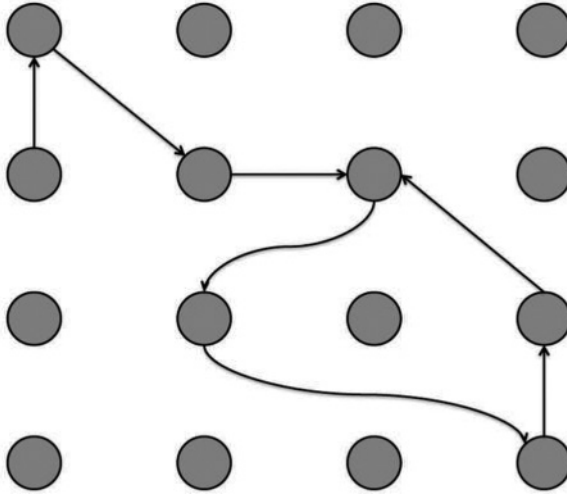
Figure 2: Graph representation of retrieval process. Gray circles denote items. Edges connect items with maximal similarity. Only a subset of edges is shown for clarity.

3.1 Random Asymmetric Similarity Matrix. Retrieval in the model depends on the properties of the similarity matrix. We first considered a simplified but very instructive version of the model with a random similarity matrix S in which elements are independent and identically distributed. Retrieval can be visualized by a directed graph, in which each node represents an item stored in memory, and each outgoing edge points to the item with maximal similarity (see Figure 2). Our task is to characterize the distribution of the number of nodes reached by following the edges starting from a random node, computed over the ensemble of graphs corresponding to different realizations of a similarity matrix.

Once an item is reached the second time, the retrieval process enters a loop, and no new items will be retrieved (see Figure 2). For an ensemble of graphs of size L, the transition probabilities between any two items are identical and equal $p_0 = \frac{1}{L-1}$. When m items are retrieved, the probability of returning to one of the previously visited $m-1$ items is $(m-1)p_0$. The probability that exactly k items will be retrieved is therefore

$$P(k; L) = (1 - p_0)(1 - 2p_0)\ldots\left(1 - (k-2)p_0\right)(k-1)p_0. \qquad (3.1)$$

Using this distribution, the average number of words recalled (RC) asymptotes to a power law function for large L:

$$RC = \langle k \rangle \approx \sqrt{\pi/2}L^{1/2}. \qquad (3.2)$$

This scaling law is known in the theory of random graphs (Flajolet & Odlyzko, 1990). Here we present a simple derivation of the scaling law of equation 3.2. Considering the limit $L \gg k \gg 1$, we rewrite the probability distribution of equation 3.1 using $1 - kp_0 \approx \exp(-kp_0)$:

$$P(k; L) \approx kp_0 \exp\left(-\frac{k^2 p_0}{2}\right) \approx \frac{k}{L} \exp\left(-\frac{k^2}{2L}\right). \tag{3.3}$$

Designating $x = k\sqrt{p_0} = \frac{k}{\sqrt{L}}$, x has the normalized probability distribution $p(x) = x\exp(-x^2/2)$. Notably, this distribution is independent of L, resulting in the scaling relationship between RC and L expressed in equation 3.2.

The standard deviation of the recall over different lists can also be computed as a function of L. A naive approach would be to consider the recall as a random process in which each word is recalled independently with equal probability $q = \langle k \rangle / L \approx L^{-(1-\alpha)}$. For large L, this assumption results in the standard deviation of recall scaling with L as the square root of the mean (law of large numbers): $std(k) \sim L^{\frac{\alpha}{2}}$. Using equation 3.3 to compute the standard deviation, however, yields the prediction (see section A.2)

$$std(k) \approx \sqrt{(2 - \pi/2)L}. \tag{3.4}$$

The standard deviation scales as the mean, indicating that the statistics of recall does not match the independence assumption.

3.2 Similarity as Overlap Between Neuronal Populations. The associative retrieval model with random asymmetric matrix S is compatible with the power law scaling of RC observed in recall experiments (Murray et al., 1976; see Figure 4). This model does not have free parameters, because any distribution of similarities will result in equivalent retrieval statistics, which is determined by the relative ranking of similarities, not their absolute values. This universality could be viewed as too restrictive, however, as it predicts that all subjects exhibit identical recall performance and all memory items have the same probability of being recalled. We therefore considered a more realistic model that reflects our underlying population coding hypothesis, with similarities between stored items defined as overlap between the corresponding neuronal groups:

$$S_{\mu\nu} = \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu, \tag{3.5}$$

where $\xi_i^\mu = 0, 1$ indicates whether neuron i encodes the item μ. As opposed to the previous case, the similarity matrix is symmetric ($S_{\mu\nu} \equiv S_{\nu\mu}$)

by construction, and elements in the same row have a positive correlation coefficient ($\rho(S_{\mu\nu}, S_{\mu\delta}) = \frac{f}{1+f}$), with the correlations growing with sparseness parameter f. Due to the symmetry, the retrieval process quickly enters into a two-items loop, since if $S_{\mu\nu}$ is the maximal element in row μ, with probability $\sim 1/2$, it is also the maximal element in row ν. We therefore prevent the return to the just-retrieved item, which could reflect, for example, the activity-dependent neuronal adaptation of our network implementation. When the most similar item is the one just retrieved, the second most similar item is chosen instead. In the graphical representation of the model, we mark these transitions with dashed arrows (see Figure 3A for a sample sequence of retrieved items).

The new algorithm has strong implications for the dynamics of retrieval. When an item is visited for the second time (see, e.g., item 5 in Figure 3A), there is a finite probability (p_1) that retrieval will not enter into a loop but rather will continue along the already visited trajectory in the opposite direction. This will happen if at the first retrieval, this item was followed by the item with second-maximal similarity (dashed arrow from item 5 to item 6 in Figure 3A). Subsequently, at each "old" item, the retrieval can either continue backward or turn to a new item ($3 \rightarrow 8$ in Figure 3A), starting another exploration of the graph. As opposed to the simplified matrix case considered in the previous section, dynamics enter a loop when a previously retrieved sequence of two consecutive items is retrieved again.

The retrieval process is characterized by three probabilities: p_0, the probability of returning from a new item to any one of previously visited items; p_1, the probability that retrieval proceeds along the previous trajectory in the opposite direction; and p_2, the probability of retrieving a new item after an old one. p_0 largely determines the RC of the model, as in the asymmetric case (RC $\approx \sqrt{1/p_0}$), p_1 determines how many loops on average there are during retrieval ($1/(1-p_1)$), and p_2 is related to the number of old items visited before retrieval of new items begins ($1/p_2$). In the sparse limit ($f \rightarrow 0$), correlations between the elements of the similarity matrix can be neglected, and these probabilities become (see section A.3)

$$p_0 = \frac{1}{2L} \quad p_1 = \frac{1}{3} \quad p_2 \approx 0.3. \tag{3.6}$$

This indicates that the model has the same scaling of the RC as the asymmetric model ($\langle k \rangle \approx \sqrt{\pi} L^{1/2}$) but with a larger prefactor, only 1.5 loops on average, and roughly three old items visited before starting a new retrieval trajectory.

The model with finite sparseness f is harder to analyze due to correlations between different elements of the similarity matrix. Correlations arise because items that are encoded by larger-than-average groups of neurons tend to have larger-than-average overlaps (similarities) with all the other
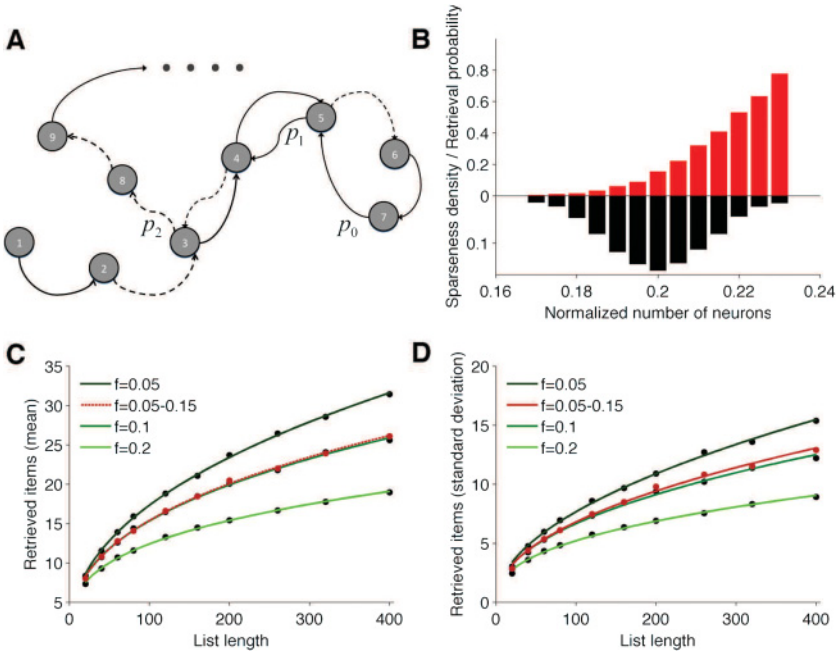
Figure 3: Associative retrieval with similarity matrix computed as overlaps between neuronal representations of memory items. (A) A sample sequence of retrieved items. Numbers inside the gray circles indicate the order of retrieval of the corresponding items. Solid (dashed) arrows indicate transitions to items with maximal (second-maximal) similarities. p_0: probability of returning to a previously retrieved item; p_1: probability that the retrieval proceeds along the previous trajectory; and p_2: probability of retrieving a new item after an old one. Retrieval enters a loop once two previously visited items are retrieved in the same order. (B) In red, the probability of retrieving an item versus the normalized size of its neuronal representation (number of neurons representing an item/total number of neurons). Parameters: $f = 0.2$, $N = 1000$, $L = 40$. In black: total probability distribution of normalized representation sizes. (C) Number of retrieved items for different list lengths, averaged over 5000 realizations of the similarity matrix with $N = 20,000$. Green: fixed values of sparseness parameter. Red: 20 values of f equally spaced in the interval [0.05, 0.15]. Filled circles: simulation results; lines: corresponding power law fits. Least square estimates of the exponents are 0.43, 0.38, 0.38, 0.31 corresponding respectively to $f = 0.05$, $f \in [0.05, 0.15]$, $f = 0.1$ and $f = 0.2$. The corresponding estimates of the prefactors are 2.35, 2.63, 2.71, 2.97. (D) Standard deviation of the number of retrieved items. Exponent estimates: 0.51, 0.47, 0.45, 0.40; prefactors: 0.74, 0.78, 0.82, 0.83.

items. Such patterns have a higher overall chance of being retrieved (see Figure 3B) and tend to be retrieved early. Subsequently, they have a higher chance of being retrieved again, thus reducing the total number of items that can be retrieved, the RC. The probability of returning to an old item scales with the list length as $p_0 \sim L^{-\beta}$ with $\beta = \frac{1-f}{1+f}$, increasing with f (see section A.4). As in the asymmetric similarity matrix model, the scaling of RC with L can then be computed as

$$RC = \langle k \rangle \sim L^\alpha, \quad \alpha = \frac{\beta}{2} = \frac{1}{2}\frac{1-f}{1+f}, \tag{3.7}$$

that is, increasing the average number of neurons per item reduces the RC.

To confirm that the asymptotic results apply to retrieval with a moderate number of items, we simulated the model with different values of f and plotted the average and standard deviation of the number of retrieved items versus the total number of items (see Figures 3C and 3D). As expected, the retrieval capacity decreases with f, and the corresponding exponents are similar to equation 3.7.

3.3 Power Laws in Free Recall Data. Several psychological studies of recall with lists of various lengths produced results that look like the power law relationship $\langle k \rangle = aL^\alpha$, with α in the range of $\alpha = 0.3 - 0.5$ (Murray et al., 1976). A representative example from the experiments of Murdock (1960) is shown in Figure 4A (data courtesy of the author). The number of words recalled, averaged over subjects, is well fit by a power law relationship with exponent $\alpha = 0.41$ (95% bootstrap confidence interval [0.37–0.44]). Similar scaling was reported in a different experimental paradigm in which subjects were asked to produce words beginning with a certain letter (Murray, 1975). As in free recall, the average number of words followed the power law function with respect to the total number of corresponding words in the vocabulary, with $\alpha \approx 1/2$. The standard deviation of recall across subjects in the data obtained by B. Murdock can also be fit with a power law, with $\alpha = 0.56$ (but with a relatively large 95% bootstrap confidence interval: 0.42–0.68; see Figure 4B and section 2, data courtesy of B. Murdock). These estimates are broadly compatible with our model if we assume a sparseness parameter in the range of 10% (see Figures 3C and 3D). With this parameter choice, the prefactors (a) of the power law functions for the average and standard deviation, obtained from simulation, are 2.7 and 0.8, respectively. The corresponding values and confidence intervals (CIs) estimated from the data are for the average number of recalled words 2.57 with CI [2.24–2.93], and for the standard deviation 0.35 with CI [0.22–0.53].

The standard deviation obtained from the data measures across-subject variability in the number of recalled words. To assess the influence of
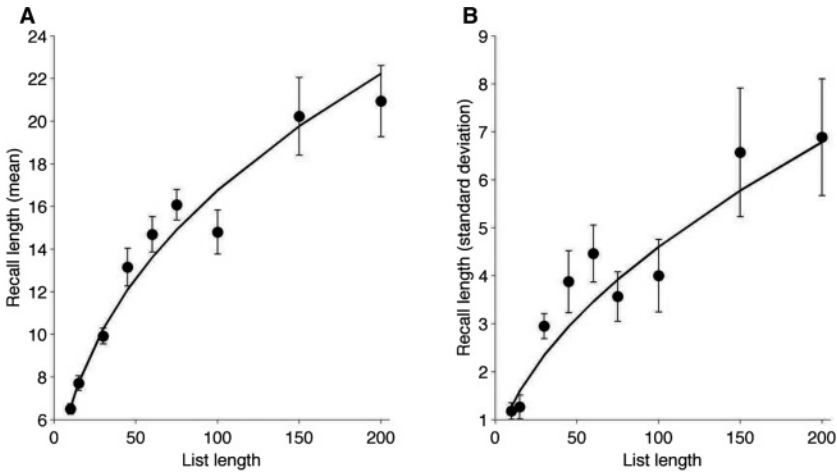
Figure 4: Power laws in free recall experiments. (A) Filled circles: number of recalled words versus list lengths, averaged over subjects. Bars represent standard error. Solid line: power law fit. (B) Same as panel A for the standard deviation of recalled words. Details of the experiments in the original paper (Murdock, 1960); data courtesy of the author.

across-subject variability on the power law behavior, we simulated the model with f uniformly distributed on the interval [0.05, 0.15]. The resulting average and standard deviation are still well fit by power law functions, and the corresponding curves are similar to those obtained with $f = 0.1$ (see Figures 3C and 3D).

4 Discussion

We conclude that power law scaling of retrieval capacity is a generic property of memory systems characterized by sequential retrieval of items based on similarity between them, independent of the particular mechanism generating the transitions between the items. This behavior can be mapped to a family of random graph models with the statistics of connections determined by the underlying encoding features of memory items.

Our model is characterized by a single parameter, the average sparseness of memory representations in the network. Representation sparseness strongly affects the retrieval capacity, which could account for the different recall abilities of different subjects. The retrieval probability of an item depends on the size of its neural representation, which could underlie variability in ease of item recall. To emphasize the universal scaling laws of retrieval capacity, the model makes several simplifications. In particular,

we assumed that the retrieval process is deterministic and that items stored in memory are randomly encoded and do not form hierarchical structures characterized by classes and subclasses. The model can readily take into account the hierarchically organized data if one assumes that items within the same class are encoded as patterns with a common core of neurons representing this class (Tsodyks, 1990). Increasing inhibition will first cause transitions between the attractors of the same class and then transitions between classes. Adding weak stochasticity to retrieval results in an interesting behavior where the network can spend a long time in a loop and then initiate a new retrieval sequence (results not shown). Our preliminary simulations show that power law scaling of retrieval capacity is still observed, but with the exponent that very slowly increases with the time available for recall.

Retrieval in the model is determined by the similarity between the memory items, without any reference to how the information was acquired by the subject. In free recall experiments, words are acquired by subjects as a list, and retrieval exhibits certain temporal regularities with respect to the order of exposure: words that are presented at the beginning and the end of the list have a higher probability of being recalled ("primacy" and "recency"), and neighboring words have an increased probability of being recalled in proximity ("contiguity") (Murdock, 1962; Howard & Kahana, 1999). These temporal regularities have been studied in the context of detailed phenomenological models of free recall (Raaijmakers & Shiffrin, 1980; Polyn, Norman, & Kahana, 2009; Howard & Kahana, 2002; Hasselmo & Wyble, 1997). The influence of primacy and recency on retrieval is reduced when a delay between acquisition and recall is introduced (Howard & Kahana, 1999). Both effects were present in the data we analyzed (Murdock, 1960) and could potentially affect our estimate of the experimental retrieval capacity. However, primacy and recency have been shown to result in a small fixed number of additionally retrieved items that is independent of list length (Murdock, 1962). Hence, the confounding effect of primacy and recency becomes less relevant as longer list lengths are considered.

Contiguity is less sensitive to the manipulation of the experimental protocol (Howard & Kahana, 1999). The model could account for contiguity if off-diagonal terms are added to the similarity matrix, which could reflect additional associations temporarily formed between words during acquisition. Our preliminary simulations show that when contiguity strength is comparable to an experimentally observed one, resulting retrieval capacity is characterized by very similar power law behavior (results not shown).

We predict that recall terminates when the same items are repeatedly retrieved, a hypothesis already proposed in the psychology literature (Laming, 2009). In free recall experiments, subjects generally report retrieved items once, but when they erroneously report the same item for the second time, there is a higher probability that no new items will be recalled

(Miller, Weidemann, & Kahana, 2012). This observation is compatible with the model prediction. The phenomenological model we propose could be implemented in neural networks with different mechanisms of transitions between memory representations. In this study, we considered one particular mechanism, involving periodic modulation of inhibition. We are not aware of experimental evidence in favor of this mechanism; however, an increase in oscillatory activity in the theta range was reported during memory retrieval (Lega, Jacobs, & Kahana, 2011). Our model provides an impetus for the study of neural oscillations during memory recall.

In conclusion, we presented the first analytical study of associative memory retrieval in the absence of cues. The model predicts a power law scaling dependence of retrieval capacity on the size of the memory. We suggest that power law scaling of cueless retrieval is a general feature of associative memory. Retrieval capacity is intricately linked to the way memories are encoded in the network; in particular, more items can be retrieved from memory with sparser representations.

Appendix: The Model: Analysis and Solutions

A.1 Stability of Memory Items and Intersection Between Them in the Network Model.

If the network state is described by a binary pattern V_i, the input received by the ith neuron is given by

$$h_i = \frac{1}{Nf(1-f)} \sum_{j=1}^{N} \sum_{\mu=1}^{L} (\xi_i^\mu - f)(\xi_j^\mu - f)V_j - \frac{J_0}{Nf} \sum_{j=1}^{N} V_j(t). \qquad (A.1)$$

To determine the stability of a memory pattern $V_i = \xi_i^\mu$, we compute the corresponding input:

$$h_i^\mu = (\xi_i^\mu - f) - J_0. \qquad (A.2)$$

This pattern is stable if the input is above threshold for all neurons encoding this memory item ($\xi_i^\mu = 1$) and below threshold for all other neurons. This requirement results in the following condition for J_0,

$$-th_i - f < J_0 < 1 - th_i - f, \qquad (A.3)$$

which has to be satisfied for all i. Since thresholds are distributed in the interval $[-T, T]$, this requires

$$T - f < J_0 < 1 - T - f, \qquad (A.4)$$

which is equation 2.3 in the main text. Similar analysis for the intersection of two memory items, $V_i = \xi_i^{\mu} \xi_i^{\nu}$, results in the condition

$$1 - 2f + \frac{T}{f} < J_0 < 2 - 2f - \frac{T}{f}, \tag{A.5}$$

which is equation 2.4 in the main text.

A.2 Similarity Matrix with Independent Elements. In this model, each active item triggers a deterministic transition to its most similar item. Hence, when any item is reached for the second time, the system will indefinitely loop over the already retrieved items and no new items can be retrieved. Following the random graphs literature, we call the trajectory leading to a loop together with the loop itself a "rho" (Menezes, Van Oorschot, & Vanstone, 1996). For the class of similarity matrices with independent and identically distributed elements, transitions between any two items are equally probable. As explained in the main text, the probability that k out of L items will be retrieved before reaching an already visited item is

$$P(k; L) = \left(1 - \frac{1}{L-1}\right)\left(1 - \frac{2}{L-1}\right)\cdots\left(1 - \frac{k-2}{L-1}\right)\frac{k-1}{L-1}, \tag{A.6}$$

which for $L \gg k \gg 1$ can be approximated by

$$P(k; L) \approx e^{-\frac{1}{L}\sum_{i=1}^{k-2} i}\frac{k}{L} \approx e^{-\frac{k^2}{2L}}\frac{k}{L}. \tag{A.7}$$

With this distribution, the statistics of retrieval asymptotic in L can be completely specified. For an arbitrary function $f(k)$, its average over the probability distribution in equation A.7 can be computed as

$$\langle f(k)\rangle = \sum_{k=2}^{L} e^{-\frac{k^2}{2L}}\frac{k}{L} f(k). \tag{A.8}$$

We can approximate this sum by an integral with the following substitution:

$$k = \sqrt{L}x \quad \frac{1}{\sqrt{L}} \to dx,$$

$$\langle f(k)\rangle = \int_0^{\infty} f(\sqrt{L}x)e^{-\frac{x^2}{2}}x\,dx. \tag{A.9}$$

Choosing $f(k) = k$, we obtain the average number of retrieved items, or retrieval capacity (RC),

$$\langle k \rangle = \sqrt{\frac{\pi L}{2}}, \tag{A.10}$$

which is related to the famous "birthday paradox" (Flajolet, Grabner, Kirschenhofer, & Prodinger, 1995). Choosing $f(k) = k^2$, we obtain

$$\langle k^2 \rangle = 2L$$

and thus

$$std(k) = \sqrt{\left(2 - \frac{\pi}{2}\right) L}. \tag{A.11}$$

A.3 Symmetric Similarity Matrix with Independent Elements. In this model, the similarity matrix is symmetric but otherwise has independent elements. It fits the definition of the similarities as the number of neurons in the intersections between corresponding neuronal representations but still neglects the correlations between different elements of the matrix. If retrieval proceeds from an item to its most similar, as in the asymmetric case, the dynamics will quickly converge to a two-items loop. The reason is that if item B is most similar to item A, then item A will be most similar to item B with a probability of approximately 0.5. We hence let the system choose the second most similar item if the most similar one has just been retrieved, as explained in the main text. Two types of transitions can therefore be distinguished: strong transitions (to most similar items) and weak transitions (to second-most similar items).

The modified rule is history dependent, which significantly influences the retrieval dynamics. Most important, when reaching an already visited item, retrieval can either repeat the original trajectory (resulting in a loop) or continue backward along the already visited items and then open a new rho (see Figure 3A). As explained in the main text, the statistics of the retrieval can be characterized by three probabilities: p_0 the probability of returning from a new item to any one of already visited items; p_1, the probability that the retrieval proceeds along the previous trajectory in the opposite direction; and p_2, the probability of retrieving a new item after an old one (see Figure 3A). Here we present a way to estimate these probabilities in the limit of large number of items L.

In order to return back from item k to item n ($k = 7$, $n = 5$ in Figure 3A), the nth element of the kth row of the similarity matrix, S_{kn}, has to be the largest of the remaining $L - 2$ elements in the kth row (excluding the diagonal and the element corresponding to the item visited just before the kth one). The probability for this would be $\frac{1}{L-2}$ for an asymmetric matrix.

For a symmetric matrix ($S_{nk} = S_{kn}$), we have an additional constraint: that the element S_{kn} is not the largest in the nth row of S, since we require that the kth item was not retrieved after the first retrieval of the nth one. The probability p_0 can thus be computed as

$$p_0 = P\big(S_{kn} = \max \vec{S}_k | S_{kn} < \max \vec{S}_n\big) = \frac{P\big(S_{kn} = \max \vec{S}_k, \, S_{kn} < \max \vec{S}_n\big)}{P\big(S_{kn} < \max \vec{S}_n\big)}$$

$$= \frac{P\big(S_{nk} = \max \vec{S}_k\big) P\big(\max \vec{S}_k < \max \vec{S}_n\big)}{P\big(S_{kn} < \max \vec{S}_n\big)} \approx \frac{1}{L}\frac{1}{2} = \frac{1}{2L} \qquad (A.12)$$

where $\vec{S}_k$ denotes the vector of relevant elements in the kth row of matrix S. Hence, p_0 has the same scaling with L as in the model with asymmetric similarity matrix but an extra prefactor of $1/2$. The probability that k items will be retrieved in the first trajectory is

$$P(k; L) \approx e^{-\frac{k^2}{4L}} \frac{k}{2L}. \qquad (A.13)$$

The RC of the first trajectory is

$$\langle k \rangle = \sqrt{\pi L}. \qquad (A.14)$$

We now present the calculation of the probability p_1 for the retrieval to turn toward already visited items after first retrieving an old item n. This will happen if the original transition from this old item ($5 \to 6$ in Figure 3A) was a weak one, so that the item closest to it is the preceding one. The unconstrained probability for a weak transition is $1/2$, but we must impose the constraint that the kth item was not retrieved after the first retrieval of the nth one. If the item preceding n is $j = n - 1$ ($j = 4$ in Figure 3A), the corresponding probability is given by

$$p_1 = P\big(\max \vec{S}_n < \max \vec{S}_j | \max \vec{S}_n > \max \vec{S}_k\big) \approx \frac{1}{3}, \qquad (A.15)$$

which follows from the observation that any ordering for the maximal elements of three vectors of equal size is equally probable, and given the condition, there are three possible orderings, out of which only one ($\max \vec{S}_k < \max \vec{S}_n < \max \vec{S}_j$) is compatible with the requirement that transition from nth item to the next one is the weak one. From this result, we conclude that the probability of having l rhos during the retrieval process decays exponentially as $P_\rho(l) = \frac{2}{3}(\frac{1}{3})^{l-1}$ and the average number of rhos is therefore $\frac{3}{2}$.

Once the retrieval process turns toward already visited items in the backward direction, at each item it can either continue in the same direction or turn to new items and initiate a new rho. The probability of the latter depends on how many old items were already retrieved, as we now illustrate. For the first item j, the new rho will be opened if there is a new item more similar to it than the preceding item $i = j - 1$. Since in the first pass through j the retrieval proceeded to n, the similarity between j and this new item is the second maximum of the vector $\vec{S}_j$, which we denote $\max_2 \vec{S}_j$. Conditions we have to take into account are that we went back to j from n ($\max \vec{S}_j > \max \vec{S}_n$) and that n did not trigger k in the first pass ($\max \vec{S}_n > \max \vec{S}_k$). The corresponding probability is therefore given by

$$p_2 = P(\max_2 \vec{S}_j > \max \vec{S}_i | \max \vec{S}_j > \max \vec{S}_n, \max \vec{S}_n > \max \vec{S}_k). \quad \text{(A.16)}$$

After some algebra, we obtain $p_2 = \frac{3}{8}$. For older visited items (i and so on), we have to take into account more and more conditions in order to estimate p_2, whose value we found to converge quickly to ~ 0.3. This means that the average number of old items visited before a new rho is opened up is approximately 3.

If there are several rhos during the retrieval process, rhos after the first one are on average shorter since there are more old items to come back to. For example, if the first rho length is k_1, the probability for the second one to be of length k_2 is

$$P(k_2 | k_1) = \left(1 - \frac{k_1 - 1}{2L}\right)\left(1 - \frac{k_1}{2L}\right)\left(1 - \frac{k_1 + 1}{2L}\right)\cdots$$
$$\left(1 - \frac{k_1 + k_2 - 3}{2L}\right)\frac{k_1 + k_2 - 2}{2L}$$
$$\approx e^{-\frac{k_1 k_2}{2L} - \frac{k_2^2}{4L}}\left(\frac{k_1 + k_2}{2L}\right). \quad \text{(A.17)}$$

Combining it with equation A.13 for $P(k_1)$, and introducing scaled variables $x = \frac{k_2}{\sqrt{2L}}, y = \frac{k_1}{\sqrt{2L}}$, one can obtain the average length of the second rho as

$$\langle k_2 \rangle = \sum_{k_1, k_2} k_2 P(k_1) P(k_2 | k_1) \approx \sqrt{2L} \int_0^\infty dx \int_x^\infty dy\, xy(x + y)e^{-\frac{(x+y)^2}{2}}$$
$$= \frac{1}{2}\sqrt{\pi L}, \quad \text{(A.18)}$$

that is, the second rho is on average two times shorter than the first one. Analogous calculations for the following rhos result in the sequence of lengths (relative to the first rho) of $\frac{3}{8}, \frac{5}{16}, \frac{35}{128}, \cdots$

A.4 Neuronal Representation of Memory Items. In this model, we assume that each stored item is represented by randomly selected groups of neurons. On average, each group contains fN neurons out of N neurons involved in coding, where f is the sparseness of representation, and we assume that N is very large. The similarity between two items is then defined as proportional to the number of neurons representing both of them (the size of the overlap):

$$S^{km} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^k \xi_i^m, \tag{A.19}$$

where index i labels neurons and $\bar{\xi}^k = 0, 1$ is a binary vector of size N indicating neurons that represent item k. For mathematical convenience, we normalized the overlaps by the size of the network, compared to equation 3.5, without affecting the retrieval process. The statistics of S up to second order are

$$\langle S^{km} \rangle = f^2, \quad k \neq m,$$

$$\text{var}(S^{km}) = \frac{f^2(1 - f^2)}{N}, \quad k \neq m,$$

$$\text{cov}(S^{km}, S^{kn}) = \frac{f^3(1 - f)}{N}, \quad k \neq m, n; m \neq n. \tag{A.20}$$

The third equation indicates that different elements of similarity matrix are not independent; in particular, elements of the same row have positive covariance. The reason for this is that items that are encoded by larger-than-average groups of neurons will tend to have larger-than-average overlaps with all the other items. This correlation structure can be captured by considering a simplified matrix with the same first- and second-order statistics:

$$\tilde{S}^{km} = M^k M^m Z^{km}, \tag{A.21}$$

where M^k is the normalized size of the kth item representation and Z^{km} is a symmetric matrix of independent gaussian variables with properly chosen mean and variance:

$$M^k = \frac{1}{N} \sum_{i=1}^{N} \xi_i^k \sim N\left(f, \sqrt{\frac{f(1-f)}{N}}\right),$$

$$Z^{km} \sim N\left(1, \frac{1-f}{f\sqrt{N}}\right). \tag{A.22}$$

We used here a gaussian approximation for M^k, which is valid for large N. In order to estimate the RC of this model, we should consider the probability of returning to an old item, p_0. We estimate the scaling behavior of p_0 with list length L in the limit of large L.

If an old item n is retrieved after item k (see Figure 3A), it means that $n = \mathrm{argmax}_l(M^l Z^{kl})$. Since n was already retrieved in the past, after, say, item $j = n - 1$, it satisfies another condition that we have to take into account: $n = \mathrm{argmax}_l(M^l Z^{jl})$. These two conditions are not independent due to the common factor M^l. We therefore have

$$p_0 = P(\mathrm{argmax}_l(M^l Z^{kl}) = \mathrm{argmax}_l(M^l Z^{jl})), k \neq j. \tag{A.23}$$

To simplify the analysis, we rewrite the product $M^l Z^{jl}$ in terms of zero average random variables and note that in the large N limit,

$$
\begin{aligned}
M^l Z^{jl} &= (f + \delta M^l)(1 + \delta Z^{jl}) \\
&= \left(f + \sqrt{\frac{f(1-f)}{N}} m \right) \left(1 + \frac{1-f}{f\sqrt{N}} z \right) \\
&\approx f + \sqrt{\frac{f(1-f)}{N}} m + \frac{1-f}{\sqrt{N}} z \\
&= f + \sqrt{\frac{f(1-f)}{N}} \left(m + \sqrt{\frac{1-f}{f}} z \right),
\end{aligned}
\tag{A.24}
$$

where m and z are independent gaussian variables with zero mean and unit variance. We can observe that in the sparse representation limit $f \to 0$, z-terms dominate the similarities, and correlations between different similarities due to M-factors can be neglected, so the model reduces to the one of the previous section, with the scaling of $p_0 \sim L^{-1}$.

By means of equations A.23 and A.24, the problem has been reduced to the estimation of the probability that two vectors of length L, X and Y, with elements of the form $x = m + az_1$ and $y = m + az_2$ ($a = \sqrt{\frac{1-f}{f}}$; $m, z_1, z_2 \sim N(0,1)$), have maximal elements at the same vector component. The problem is equivalent to considering vectors with elements x and $y = x + \gamma z$ ($\gamma = \frac{\sqrt{1-f^2}}{f}$; $x, z \sim N(0,1)$) instead. Going back to equation A.23, we then obtain

$$
\begin{aligned}
p_0 &= P(\mathrm{argmin}(X) = \mathrm{argmin}(Y)) = LP(x_1 = \min(X), y_1 = \min(Y)) \\
&= L \int Dx_1 Dz_1 \left[\prod_{i=2}^{L} \int_{x_i > x_1} Dx_i \int_{z_i > \frac{y_1 - x_i}{\gamma}} Dz_i \right] \\
&= L \int Dx_1 Dz_1 \left[\int_{x_i > x_1} Dx_i \int_{z_i > \frac{y_1 - x_i}{\gamma}} Dz_i \right]^{L-1},
\end{aligned}
\tag{A.25}
$$

with $Dx = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$. The term in brackets, when taken to a power $L \gg 1$, will be negligible unless x_1 is approaching $-\infty$. We can therefore estimate this term as

$$\int_{x_i > x_1} Dx_i \int_{z_i > \frac{y_1 - x_i}{\gamma}} Dz_i = \int_{x > x_1} Dx \left(1 - \varphi\left(\frac{y_1 - x}{\gamma}\right)\right)$$

$$= 1 - \varphi(x_1) - \int_{x_1}^{\infty} \frac{dx}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \varphi\left(\frac{y_1 - x}{\gamma}\right)$$

$$\approx 1 - \varphi(x_1) - \varphi\left(\frac{y_1}{\sqrt{1 + \gamma^2}}\right), \tag{A.26}$$

where $\varphi(x) = \int_{-\infty}^{x} Dz$ is the gaussian CDF. Equation A.25 then becomes

$$p_0 = L \int Dx_1 Dz_1 e^{-L\left(\varphi(x_1) + \varphi\left(\frac{y_1}{\sqrt{1+\gamma^2}}\right)\right)}. \tag{A.27}$$

In the large L limit, using saddle point approximation for this integral results in the following scaling of p_0 with L:

$$p_0 \sim \frac{1}{L^{\frac{1-f}{1+f}}}. \tag{A.28}$$

Therefore, the scaling of the RC$\sim L^\alpha$ for large L will be

$$\alpha = \frac{1}{2} \frac{1 - f}{1 + f}. \tag{A.29}$$

Acknowledgments

References

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks.* Cambridge: Cambridge University Press.

Bibitchkov, D., Herrmann, J. M., & Geisel, T. (2002). Pattern storage and processing in attractor networks with short-time synaptic dynamics. *Network*, *13*(1), 115–129.

Binet, A., & Henri, V. (1894). La memoire des mots. *L'annee psycholog., Bd. I*, *1*, 1–23.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions*. New York: Wiley-Interscience.

Flajolet, P., Grabner, P. J., Kirschenhofer, P., & Prodinger, H. (1995). On Ramanujan's Q-function. *Journal of Computational and Applied Mathematics*, *58*(1), 103–116.

Flajolet, P., & Odlyzko, A. (1990). Random mapping statistics. In *Advances in cryptology EUROCRYPT'89*, 737 (pp. 329–354). New York: Springer.

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., & Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, *322*(5898), 96–101.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1–2), 1–34.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–941.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Laming, D. (2009). Failure to recall. *Psychological Review*, *116*(1), 157–186.

Lega, B. C., Jacobs, J., & Kahana, M. (2011). Human hippocampal theta oscillations and the formation of episodic memories. *Hippocampus*, *761*, 748–761.

Menezes, A. J., Van Oorschot, P. C., & Vanstone, S. A. (1996). *Handbook of applied cryptography*. Boca Raton, FL: CRC.

Miller, J. F., Weidemann, C. T., & Kahana, M. J. (2012). Recall termination in free recall. *Memory and Cognition*, *40*, 540–550.

Murdock, B. B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology*, *60*(4), 222–234.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Murray, D. J. (1975). Graphemically cued retrieval of words from long-term memory. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(1), 65–70.

Murray, D. J., Pye, C., & Hockley, W. E. (1976). Standing's power function in long-term memory. *Psychological Research*, *38*(4), 319–331.

Pantic, L., Torres, J. J., Kappen, H. J., & Gielen, S.C.A.M. (2002). Associative memory with dynamic synapses. *Neural Computation*, *14*(12), 2903–2923.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychol. Rev.*, *116*(1), 129–156.

Raaijmakers, J.G.W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. *The Psychology of Learning and Motivation: Advances in Research and Theory*, *14*, 207–262.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*(3), 365–372.

Russo, E., Namboodiri, V. M. K., Treves, A., & Kropff, E. (2008). Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, *10*(1), 015008.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Tsodyks, M. V. (1990). Hierarchical associative memory in neural networks with low activity level. *Mod. Phys. Lett. B.*, *4*(4), 259–265.

Tsodyks, M., & Feigel'Man, M. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, *6*, 101.