Long Term Memory and the Densest K-Subgraph Problem

Robert Legenstein[1], Wolfgang Maass[2], Christos H. Papapdimitriou[3], and Santosh S. Vempala[4]

1 **Institute for Theoretical Computer Science, Graz University of Technology, Graz, Austria**
robert.legenstein@igi.tugraz.at
2 **Institute for Theoretical Computer Science, Graz University of Technology, Graz, Austria**
maass@igi.tugraz.at
3 **Computer Science, Columbia University, NY, USA**
christos@cs.berkeley.edu
4 **Computer Science, Georgia Tech, Atlanta, USA**
vempala@gatech.edu

Abstract

In a recent experiment [9], a cell in the human medial temporal lobe (MTL) encoding one sensory stimulus starts to also respond to a second stimulus following a combined experience associating the two. We develop a theoretical model predicting that an assembly of cells with exceptionally high synaptic intraconnectivity can emerge, in response to a particular sensory experience, to encode and abstract that experience. We also show that two such assemblies are modified to increase their intersection after a sensory event that associates the two corresponding stimuli. The main technical tools employed are random graph theory, and Bernoulli approximations. Assembly creation must overcome a computational challenge akin to the Densest K-Subgraph problem, namely selecting, from a large population of randomly and sparsely interconnected cells, a subset with exceptionally high density of interconnections. We identify three mechanisms that help achieve this feat in our model: (1) a simple two-stage randomized algorithm, and (2) the "triangle completion bias" in synaptic connectivity [14] and a "birthday paradox", while (3) the strength of these connections is enhanced through Hebbian plasticity.

1998 ACM Subject Classification F.1.1 Models of Computation; I.2.6 Learning: Connectionism and neural nets

Keywords and phrases Brain computation, long term memory, assemblies, association

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.57

1 Introduction

How do sensory stimuli from entities in the outside world effect the creation of stable memories in the animal cortex, and how are such memories modified by further experience, for example by the introduction of associations between them? A recent experiment [9] provides certain interesting insights into these fundamental questions. They recorded from a total of 613 neurons in the medial temporal lobe (MTL, the brain region near the hippocampus long believed to be implicated in memory) of 14 human subjects. They presented, in a particular rigorous protocol running over several stages, many images of places and people, with repetitions and occasional superpositions. Several neurons were identified that fired consistently at the presentation of a particular place or person. One particular neuron in one

subject may have fired consistently when an image of the Eiffel tower was presented, but failed to fire when other images were presented, such as an image of Barack Obama (example for the present illustration). Then a combined image of Obama in front of the Eiffel tower was presented, and, predictably, the neuron fired (as it always does when the Eiffel tower is seen). Remarkably, after this combined presentation the neuron also fired when an image of Obama was shown: the subject had *learned* the *association* between the two!

In this paper we propose a model for the formation and association of memories, based on random graphs and Hebbian plasticity, which we believe captures in a simplified way the basic mechanisms involved in this phenomenon. Our model predicts that stable memories will be formed in response to stimuli and that two such memories can "nudge closer together" in response to a simultaneous presentation of the two stimuli. Interestingly, as explained in Section 8, our results recall quite vividly the narrative of [5] about a related phenomenon in mouse olfaction.

How is it possible, by monitoring a few dozen neurons of a subject (out of many million in the MTL) and by presenting a hundred or so familiar images (as [9] and similar experiments have done), to identify several neurons consistently responding the images? About the only plausible explanation is that each image shown must excite a great number of neurons, and must do so quite consistently. This and many other experiments (see [3, 13] for reviews) confirm earlier theories and hypotheses going back to Hebb [8] that tokens of cognition (such as the Eiffel tower) are represented by assemblies of many excitatory neurons[1], often called *concept cells* [12]. These assemblies are stable, in the sense that in the short term they fire more-or-less consistently and as a whole with the same stimulus (absent new associations). They are therefore believed to be *densely connected* through many and strong synapses. Every time the corresponding cognitive entity is active in the brain, all these cells (more or less) fire. In fact, the experiment of [9] even suggests that these assemblies are *fluent* in that they can be changed dynamically in response to new experiences and associations.

Despite the emerging consensus that concept cell assemblies in the MTL are an important piece of the puzzle of memory and cognition, and simulation results verifying that assemblies can indeed emerge (see [11, 16], and [10] for related work on assembly *binding*) we are not aware[2] of theoretical models predicting the creation of cell assemblies, much less their association. Here we present such a model for the formation of assemblies in a recurrent network of *memory neurons,* in response to the spiking of a separate cell population of *sensory neurons* representing the sensory experience — we call such spiking *the presentation of the stimulus.*

We model the memory neurons as a directed $G_{n,p}$ graph, and the projection from the sensory neurons as a *bipartite, one-way* $G_{n,p}$ graph. Our model assumes that firing of neurons happens in discrete steps and synchronously. One key simplifying assumption of our model is that, at each step, *exactly* K *memory neurons fire*, namely the ones receiving at that instant the largest synaptic input. This is of course a strong simplifying assumption; it is intended to capture the way in which the firing thresholds of the excitatory memory neurons are regulated by inhibitory neurons (not modeled explicitly here), resulting in an equilibrium in which a relatively stable number of excitatory neurons end up firing.

We assume that the neural network is fixed; the only possible modification comes through *Hebbian plasticity:* if there is a synapse (directed edge in the graph) from i to j, and i and j fire in two consecutive steps in this order, then the *strength* of this synapse (and thus the

[1] Naturally, two such sets can overlap — and a simple calculation suggests that they are likely to do so.
[2] Valiant's important and relevant theory [**?**] is discussed extensively in the sequel.

strength of the firing signal it can transmit theretofore) increases. Our model, and especially our way of modeling inhibition, was inspired and informed by the discussion in [5] of a related phenomenon in the mouse brain, see the description in Section 8.

We prove formally that, in this model, when a stimulus is presented through the repeated firing of a set of sensory neurons representing the stimulus, a corresponding stable assembly of neurons can indeed be formed, with high probability. We also show that two such assemblies, once both formed in response to the presentations of two different stimuli at different times, can subsequently be modified by increasing their intersection in response to the simultaneous presentation of the corresponding stimuli (as happens in the "Obama at Eiffel" example). We first analyze a *linearized* version of the model, in which thresholds and saturation are ignored; we arrive at a dynamical system (Eq. 3.1), which we were able to solve through an equilibrium equation *in closed form.* We establish convergence under minimal assumptions, see Theorem 3.1. The analytical solution (see the statement of Theorem 3.1) recalls vividly the description of the related phenomenon in mouse olfaction [5], see Section 8.

We then proceed to analyze the strongly nonlinear dynamics of the full model. We prove that, here too, the description of [5] prevails: already after two steps of stimulus spiking, a set of cells has been selected comprised of two kinds, quite balanced in cardinality: cells that have strong projection from the stimulus population, and cells to which *those* project strongly (see Theorem 4.1(1)). Subsequent steps modify this assembly in rather limited ways (Theorem 4.1(2)).

Furthermore, simulations show that, if two stimuli A and B are presented in the order A, then B, then A + B (both stimuli spike), then A, then B, association happens: the assemblies responding to A and B change slightly so they intersect a little more. We prove an analytical result (Theorem 5.1) establishing that some fraction of the two assemblies will indeed migrate towards each other.

Synaptic Density of Assemblies and Valiant's Model

Ever since Hebb, assemblies were conjectured to be dense in synaptic connections. In fact, several of our proofs take advantage of the fact that synaptic density within the assembly being formed is markedly higher than random. The synaptic density of the formed assemblies is further enhanced in a more sophisticated random graph model that we call $G^{++}_{n,p}$, capturing experimental observations [14, 7] that the distribution of synaptic connections is biased towards *triangle completion* (see Section 6); in this model a combinatorial *birthday paradox* argument establishes that, for the parameter range of interest, assemblies are far more intraconnected than one would expect.

High synaptic density of assemblies is a major advantage when it comes to the maintenance of stability and consistency, but of course is a severe design burden at creation time[3]:

> *In a random graph, how do you select an induced subgraph that is much more dense than average?*

This looks and feels like the intractable DENSEST K-SUBGRAPH problem [4]; in Section 6 we briefly discuss how our proposed mechanism can be abstracted as a rather apt algorithm for solving approximately this computational problem in a G_{np} graph, and how it compares with other known algorithms for it.

[3] According to Les Valiant (private communication to CHP, 2017) dense assemblies are "infinitely harder" to create than the *items* of Valiant's theory, discussed next.

The high synaptic density of assemblies is one point of stark contrast of our theory of assemblies against L. G. Valiant's theory of *items*. More than two decades ago, Valiant articulated his important computational theory of cortex. He proposed an elegant model of cortex consisting of neurons connected through random synaptic connections and equipped with an automaton-like *vicinal* programming language. He proposed that tokens of human cognition, such as "Eiffel" and "Obama", are represented in cortex by sets of neurons called *items*, which can be combined, through vicinal algorithms, in ways akin to logical gates to form new items, and associations between such. Unlike our current theory of assemblies, an item in Valiant's theory is an arbitrary set of neurons with no particularly high connectivity. Perhaps the most critical difference between Valiant's theory and our current discussion of assemblies, and our main technical contribution, is this: Valiant's vicinal programming model allows a generous repertoire of elementary instructions (modifications of the parameters of neurons and synapses, such as threshold and synaptic strength as an arbitrary function of local state), whereas our model is far more minimalistic, relying only on the simple, and rather standard and broadly accepted, rule of Hebbian plasticity explained next, and a simplified rigorous treatment of inhibition.

2 Our Model

- There is a *memory area* M consisting of n neurons randomly connected through synapses according to the directed $G_{n,p}$ model (for every $i \neq j \in M$, the probability that there is a synapse (i, j) is p).
- There is a *sensory area* S, whose neurons project through synapses to the neurons in M according to the one-way bipartite $G_{n,p}$ model (for every $i \in S, j \in M$, the probability that there is a synapse (i, j) is p). A *stimulus* is a set of L neurons in S.
- *Firing of neurons* happens in discrete steps $1, 2, \ldots, t, \ldots$ and in synchrony. The *presentation of a stimulus* is the firing of the corresponding S neurons for a large number of subsequent steps (such repetitive firing is called *spiking*).
- Each synapse (i, j) (within M and from S to M) has a *strength* w_{ij}, initially 1.
- We denote the set of neurons in M that fire at time t by $F^t \cap M$ (the set F^t includes also neurons in S). This set is defined as follows: We first calculate for each neuron $i \in M$ its *synaptic input* $I_i^t = \sum_{j \in F^{t-1}} w_j i$, the sum of all synaptic weights of the synapses to i coming from neurons $j \in M \cup S$ that fired at the previous step. Then F^t is the set of K neurons in M with the largest I_i^t.

  Justification: In a simple model of the cortex, besides the *excitatory* neurons considered here there are also *inhibitory* neurons, whose role is, roughly speaking, to make sure that the number of firing excitatory neurons is not excessive. There are random synaptic connections between the two populations: Excitatory neurons project positively to inhibitory neurons, while inhibitory neurons project *negatively* on excitatory ones, increasing their firing threshold. Here we assume that, at the equilibrium of this random process, exactly K of the (excitatory) neurons in M will fire. This is obviously a strongly simplifying assumption, inspired by the narrative about inhibition in [5]. We are confident that a more detailed random graph model of the process described above would also result in a number of firing neurons that is strongly concentrated, by the law of large numbers, near a value K; this would be an interesting extension of our work, which we intend to pursue.
- *Hebbian plasticity.* If there is a synapse (i, j) and it so happens that $i \in F^t$ and $j \in F^{t+1}$,

then the weight of this synapse is increased by a small amount $\beta > 0$, say[4].

Indicative ranges of these parameters for the human MTL are these: $n = 10^7, p = 10^{-3}, K = L = 10^4$, and $\beta = 0.1$. In our simulations we use values such as $n = 10^3 - 10^4, p = 10^{-2}, K = L = 10^2$, and $\beta = 0.1$.

3 The Linearized System

We start with a simplified model that ignores the nonlinearity of thresholds (a useful and familiar mathematical simplification from the theory of artificial neural networks). The assembly creation process is then captured by the following dynamical system, where z_j is the stimulus projection strength at neuron j, $x_j(t)$ the activation probability of neuron j at time t and $W_{ij}(t)$ is the strength of the synapse ij at time t:

$$\begin{aligned} x_j(t+1) &= z_j + (W^T x(t))_j \\ W_{ij}(t+1) &= W_{ij}(t) + \beta x_i(t) x_j(t+1) \end{aligned} \tag{1}$$

In addition, the pre-synaptic weights at each neuron are normalized to sum to 1 after each weight update. We assume that initially W is a random adjacency matrix in $G_{n,p}$, and that at time 0, the activations $x(0)$ are set to 1 for a random subset of K neurons and 0 for the rest.

▶ **Theorem 1.** *With high probability over W and $x(0)$, the dynamics (1) converge linearly to the following equilibrium ($ij \in E$ denotes a synapse from i to j):*

$$x_j^* = z_j + \frac{\sum_{i:ij\in E} (x_i^*)^2}{\sum_{i:ij\in E} x_i^*}.$$

This equation captures and verifies in a rather striking way the description of a similar phenomenon in [5], see Section 8: the probability that a neuron joins the assembly has two components, the first being the size of the stimulus projection on the neuron, and the second a function of the corresponding probabilities of (recursively) its presynaptic neurons — a function that is monotonically increasing in the region of interest (most neurons have near-zero probabilities, while the rest have comparable probabilities).

Proof. An equilibrium activation x^* must satisfy

$$x^* = z + W^T x^* \text{ and so } x^* = (I - W^T)^+ z$$

where A^+ is the pseudo-inverse of A. The equilibrium weight matrix satisfies:

$$W_{ij}^* = \frac{W_{ij}^* + \beta x_i^* x_j^*}{\sum_{l:lj\in E} W_{lj}^* + \beta \sum_{l:lj\in E} x_l^* x_j^*}.$$

Therefore, using the fact that we normalize the incoming synaptic weights of each node:

$$W_{ij}^*(1 + \beta(\sum_{l:lj\in E} x_l^*)x_j^*) = W_{ij}^* + \beta x_i^* x_j^*$$

[4] Or instead multiplied by $1 + \beta$, or in either case up to a saturation level B. Our results are robust with respect to these popular variants of Hebbian plasticity.

which implies

$$W^*_{ij} = \frac{x^*_i}{\sum_{l:lj\in E} x^*_l}$$

and therefore

$$x^*_j = z_j + \frac{\sum_{i:ij\in E}(x^*_i)^2}{\sum_{i:ij\in E} x^*_i}.$$

The mathematically demanding part is the proof of convergence; this follows from Lemma 2 below (whose proof can be found in the Appendix): the first part shows progress in each step at a rate depending on the spectral gap of $W(t)$, and the second part shows that weight updates cannot slow down the convergence. In addition, convergence implies *stability*: If at a later time the same input signal is presented, the same probabilities of formation will be effected. The "high probability" clause of the theorem refers to the fact that the following events are highly probable for random W and $x(0)$: (a) W is irreducible, and (b) at each time t, the matrix $W(t)$ has non-negligible spectral gap, and therefore the lemma applies. ◀

▶ **Lemma 2.** *Let W be an $n\times n$ nonnegative, irreducible matrix and $z \in \mathbf{R}^n_+$ be a nonnegative vector.*

1. *The iteration $x(t+1) = z + W^T x(t)$ with random $x(0)$ satisfies*

$$\frac{\|W^T x(t+1)\|^2}{\|x(t+1)\|_2^2} > \frac{\|W^T x(t)\|^2}{\|x(t)\|_2^2}.$$

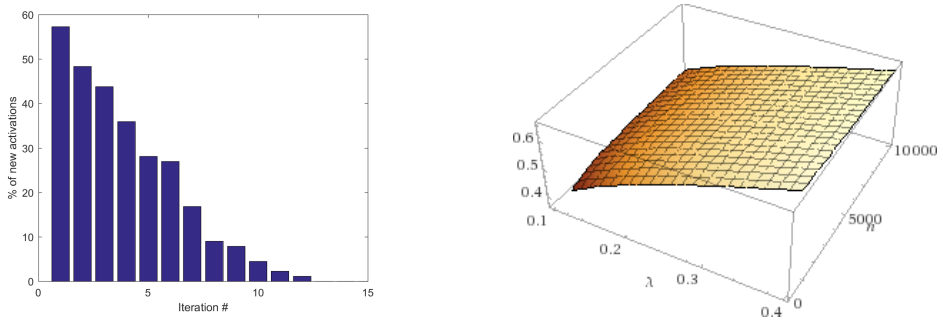
2. *Let the weight update rule (1) be applied repeatedly to synapse weights W for some $\beta > 0$ and current activation vector x. Then for each cell j, the vector w of incoming synaptic weights converges to a vector $\tilde{w}$ which satisfies*

$$\frac{\tilde{w}\cdot x}{\|\tilde{w}\|_2} \geq \frac{w\cdot x}{\|w\|_2}.$$

4 The Nonlinear System

Continuing to the full nonlinear model, our quantitative narrative of assembly formation (see Theorem 4.1 below) also recalls the key features in the description in [5]. Let $A(1), A(2), \ldots, A(t), \ldots$ denote the sets of K cells in the memory area firing at each discrete time step $t > 0$ during stimulus presentation. It is clear that $A(1)$ consists of the cells with the largest projection from the stimulus — we intuitively think of these cells as "born rich". At the next step, the theorem states that $A(2)$ contains a balanced mix of $A(1)$ cells, and cells that have strong *combined* projection from the stimulus *and* from $A(1)$. In experiments, the quantity λ capturing this balance is indeed reasonably far from 0 and 1 for the range of interest (see Figure 1b). Moreover, assuming powerful enough synaptic plasticity, subsequent sets $A(t)$ will converge rapidly to the final assembly A. Thus the theorem reasserts the interpretation of [5]. In the statement of the theorem, asymptotics are in terms of n, assuming that K is bounded from above and below by powers of n (e.g., $K = \sqrt{n}$).

▶ **Theorem 3.** *The following hold with high probability over random initial synaptic connections:*



(a) Fraction of new activations. **(b)** Solutions to the β, λ equation.

Figure 1 Illustration of Theorem 3. The first plot is the relative size of the newly activated cells $A(t) \setminus A(t-1)$ in each iteration, with $n = 2000, K = L = 2\sqrt{n} = 89$.

1. *For all $t \geq 2$, we have*

$$|A(t) \cap A(1)| = (\lambda + o(1))K$$

for a constant $\lambda \in (0,1)$ depending on the synapse probability p, plasticity factor β, plasticity ceiling B and the ratio of assembly size to input size K/L.
2. *For large enough β and B, there is a $\bar{\lambda} < 1$ s.t. for all $t \geq 2$ we have*

$$|A(t) \setminus A(t-1)| \leq \left((1-\bar{\lambda})^t + o(1)\right) K.$$

Thus, the sequence of activated sets stabilizes rapidly, with the change to the previous set decaying geometrically with the number of steps. This can be seen in simulation, even for modest parameter values (see Figure 1a).

▶ **Lemma 4.** *Let $X \sim Bin(n,p)$. Then for $t > np$,*

$$\Pr(X \geq t) \leq \exp\left(-nH\left(p, \frac{t}{n}\right)\right)$$

where $H(p,q) = q\log(q/p) + (1-q)\log((1-q)/(1-p))$ is the entropy function.

For $p < 0.5$ and $np < t < 2np$, the above bound is at most $\exp(-(t-np)^2/np)$. We will use this in the proof of the main theorem.

Proof. (of Thm. 3.) The set $A(1)$ consists of the K cells that receive the maximum total signal from the L stimulus cells. Since we model synaptic structure as a random graph with synapse probability p, each cell j receives a signal $y_j = \sum_{i \in S} W_{ij}$ where S is the set of stimulus cells. This is the Bernoulli distribution $B(L,p)$, the sum of L independent Bernoulli random variables each with expectation p. The y_j's for different cells j are independent and thus the set $A(1)$ is exactly the K-cap of (K largest samples from) n independent copies of $B(L,p)$. This is the tail of the the Bernoulli $B(L,p)$ of probability K/n. A simple calculation using the Binomial tail bound (Lemma 4) gives us that the threshold for the K-cap is (very close) to

$$t_1 = pL + \sqrt{pL\ln(n/K)}$$

where n is the total number of cells in the MTL, and each cell in $A(1)$ receiving at least this much signal from the stimulus.

For the second step, the distribution of the signal to a cell depends on whether it is in $A(1)$ or not. A cell j *not* in $A(1)$, receives the signal of the input stimulus as well as the signal from cells in $A(1)$. We approximate this distribution by the Binomial $B(K+L,p)$, which ignores the conditioning that such a cell j was *not* in the K-cap of the initial Binomial; the latter conditioning can only reduce the probability that a cell not in $A(1)$ is a winner in the next round. For a cell j in $A(1)$, the signal from the external stimulus cells is fixed by the first step but amplified by a factor $(1+\beta)$ due to plasticity; the signal from the K cells in $A(1)$ is random. So their signal comes from the distribution $(1+\beta)t_1 + B(K,p)$. The threshold for the K-cap of this joint distribution is then close to

$$t_2 = (1+\beta)t_1 + pK + \sqrt{pK \ln \frac{K}{\lambda K}} = p(K+L) + \sqrt{p(K+L) \ln \frac{n}{(1-\lambda)K}}$$

where λ is the fraction of $A(2)$ that is also in $A(1)$. The equation above can be solved numerically for λ. For our range of interest, with $K = L = 2\sqrt{n}$, for λ in $[0.1, 0.4]$, the plasticity parameter β is in $[0.3, 0.6]$, and gets slightly smaller for larger graph size (see Figure 1b).

This establishes the intersection between $A(1)$ and $A(2)$ is at least a λ fraction with high probability. The first part of the theorem says that almost all of this intersection remains activated for all future time steps. To see this, note that after step 2, the weights from the input as well as from the all of $A(1)$ to cells in the intersection are increased again by a factor $1+\beta$. These cells, which were already ahead, and are now further ahead. The rest of $A(1), A(2)$ are strictly inferior and the cells outside $A(1) \cup A(2)$ have gained no advantage at all, even ignoring the effect of the cap. The advantage of the intersection gets magnified with each iteration.

The general proof for both parts proceeds by induction on t. We claim inductively that, with high probability, any cell that is activated for a second time remains activated for all future steps. To see the inductive step, clearly such cells have an advantage over all other cells of factor of at least $(1+\beta)$ for the signal coming from the external input cells, and for the signal coming from all such cells in the previous iteration (which are an increasing fraction of K, by the hypothesis). When such a cell is activated for the second time, it was already ahead of all other cells not in the activated set; this advantage is magnified by a factor of $1+\beta$ for the signal from the input and from all such cells. Among the remaining cells, some will be activated for the second time and some for the first time. The relative fraction is bounded by $\bar{\lambda}$ via a calculation similar to the base case above — at each step the competition is between cells that have just been activated and received a $(1+\beta)$ boost for the first time on part of their input signal (a diminishing fraction of K of such cells), and most of the n other cells that have not felt any plasticity yet. ◀

This result and its proof, as well as the equilibrium result for the linearized model, imply *consistency and stability* of the assemblies formed: If the same stimulus (or even a fairly similar stimulus, in some appropriate metric) is presented at a later time (even after certain limited changes in the weights of the circuit) then with high probability the same (more-or-less) cells will fire.

5 Association

The experiment with the sequence of presentations A, B, A + B, A, B, described in the introduction shows that, after the joint presentation, the assemblies that respond to A and B "creep closer together," increasing their intersection to reflect the association between the

two stimuli (see Figure 5.1). To understand and illustrate the underlying mechanism, we shall consider a stylized special case of assembly creation. In particular, we assume that plasticity is so intense that the assembly corresponding to stimulus A consists of only the cells receiving the K largest signal from stimulus A, and similarly for B. The advantage of this assumption is that it allows us to study the association phenomenon divorced from the subtleties of the proof of Theorem 4.1. Consider now the presentation of A + B; we can show the following (we assume for algebraic convenience that $K = L$):

▶ **Theorem 5.** *There is a $\mu > 0$ such that, with high probability, at least a μ fraction of the cells of the assembly for B will respond to the next presentation of A, and vice-versa.*

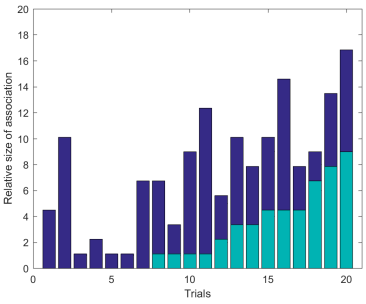
Proof. It is easy to see that, upon the presentation of A + B, the K cells that will fire consist of a fraction of the cells of assembly A and a fraction of the cells of assembly B, namely those that maximize the combined signal from A +B. The signal the cells of assembly A receives from stimulus A is very close to the K-cap $Lp(1+\sqrt{\frac{2\ln(n/K)}{Lp}})$, while from assembly B they only receive a binomial distribution with mean Lp (since those cells have not been selected for B). By symmetry, with high probability the number of cells from each of the two assemblies that fire will be very close to $\frac{K}{2}$. Let us call these sets $\frac{1}{2}$A and $\frac{1}{2}$B, respectively. We claim that all cells in $\frac{1}{2}$A receive a signal from stimulus A equal to $Lp(1+\sqrt{\frac{2\ln(n/K)}{Lp}})$ as before, plus a signal from stimulus B equal to $Lp(1+\sqrt{\frac{2\ln(n/K)}{Lp}}))$, since they were selected to be the cells in assembly A that are above the median with respect to received signal from B. Now notice that there are about pK^2 synapses between $\frac{1}{2}$A and $\frac{1}{2}$B. After the presentation of A + B, these synapses, as well as the synapses from stimulus A to $\frac{1}{2}$B, are boosted to plasticity saturation (since their endpoint cells spiked together for long enough).

If stimulus A is presented next, all cells of assembly A will fire at the first step. At the second step, however, they will have new competition (which was missing at assembly creation time) from the cells in $\frac{1}{2}$B which have, during the presentation of A + B, acquired strong synapses from stimulus A and assembly A. As in our analysis in the proof of Theorem 4.1, the threshold equation for μ becomes, after simplification $\frac{1}{2}+\sqrt{2\ln(n/k)}+\sqrt{2\ln(\frac{1}{1-\mu})} = \sqrt{\frac{2}{3}\ln(\frac{1}{2\mu})}$. The third term of the lhs is negligible for small μ, and the rhs can take over the other two terms by selecting μ appropriately small. ◀

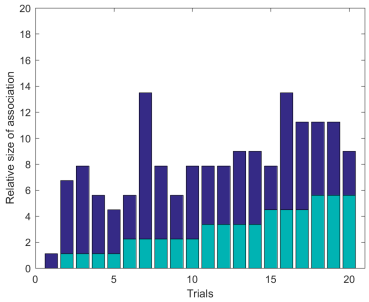
This argument gives $\mu < 1\%$, which is conservative: the fractions of the two assemblies that intersect after association seem to amount to several percent (see Figure 2a). A similar statement can also be shown in the absence of the high plasticity assumption, by focusing on $A(1) \cap A(2)$ (which we know is a constant fraction of K), instead of the whole assembly.

6 The Model $G^{++}_{n,p}$ and the Birthday Paradox

As we have seen, the assemblies created by the process analyzed here have higher density than random sets of cells, essentially due to the second step of the stimulus presentation (Theorem 4.1(1)). This effect is enhanced considerably if we adopt a more sophisticated random graph model. Experiments in [14, 7] and elsewhere suggest that synaptic connectivity in brain areas including parts of the MTL is not uniformly random, but *biased* towards *reciprocity* and *triangle completion*. Reciprocity means that, even though the overall density of edges remains p, conditioned on synapse ij being present, the probability of synapse ji is larger than p, perhaps between 3 and 5 times larger. Triangle completion means that, conditioned on the existence of synapses ij and ik, the probability of synapse jk is similarly larger.



(a) Association fraction in the standard model $G_{n,p}$.



(b) And in the extended model $G^{++}_{n,p}$.

Figure 2 Illustration of association *à la* Ison et al [9].The intersection of the assemblies for stimuli A and B, before and after a joint A+B presentation, over 20 trials with $n = 2000$ cells and assembly/stimulus sizes $K, L = \lceil 2\sqrt{n} \rceil = 89$. Trials sorted by size of initial overlap.

Here we shall ignore reciprocity bias, and adopt a limited form of triangle completion bias: Again, the overall density of edges is p, except that, conditioned on the synapses ij and ik being present, *where i is a sensory cell and j, k are memory cells,* the probability of synapse jk is γp for some $\gamma > 1$. The reason we ignore reciprocity bias is because it seems to have only a small effect on our present focus; the reason we restrict triangle completion bias in the bipartite graph between the sensory and memory areas is because this part of triangle completion matters most, and also because there are formal difficulties involved in the precise definition of a generative model of random non-bipartite directed graphs with triangle completion bias. We call the resulting random graph model $G^{++}_{n,p}$.

Within this model, and for some reasonably broad range of parameters, we can show that there is substantial enhancement of the density of the assemblies. The underlying mathematical reason is the *birthday paradox*: Upon presentation of a stimulus, memory cells receive a signal of strength Lp on average — that is, on the average they each have Lp presynaptic stimulus cells. Consider two cells i and j in the memory area, and call them *siblings* if they have a common presynaptic cell in the stimulus; the chance that this is the case is clearly p^2L. Suppose however that we have identified a subset S of memory cells whose signal is known to be of strength at least αLp for some $\alpha > 1$; for example, within the initial cells A_1 of the assembly (recall Theorem 4.1 (1)), $\alpha = 1 + \sqrt{\frac{2\ln(n/K)}{pK}}$. Then the chance that two cells in S are siblings is increased to $\alpha^2 p^2 L$. For parameter values $n = 10^7, L = K = 10^3, p = 10^{-2}$ (all very much within the range of interest), the probability of two cells of S being siblings is increased from .1 to about .8. Since in $G^{++}_{n,p}$ two siblings have enhanced probability of synaptic connection, say $\gamma = 4$, it follows that the synaptic density within the region A_1 of an assembly in $G^{++}_{n,p}$ will be more than 3 times its value in $G_{n,p}$. The rest of the assembly will also have increased interconnections, because of a similar birthday paradox argument, but in addition for the reasons obtaining in the proof of Theorem 4.1(1). Adding plasticity to the picture, we conclude that *assembly creation selects a set of cells with denser and stronger synaptic connections than random,* and achieves this in three ways: Through the second step of the creation process (Theorem 4.1(1)); through plasticity; and through triangle completion and the birthday paradox in the $G^{++}_{n,p}$ model.

Our experiments show that in $G^{++}_{n,p}$ assembly formation converges faster than in $G_{n,p}$, while the association effect of the previous section does not change much.

7 The Densest K-Subgraph Problem

Finding a set of K nodes with maximum density is an intractable problem to solve exactly or approximately in general graphs. The mechanism for assembly creation proposed here can be abstracted as an approximation heuristic for $G_{n,p}$ graphs:

```
1 Select a set S of λK nodes at random
2 Let T be the (1 − λ)K nodes with highest number of edges from S
3 Return S ∪ T; optimize λ ∈ (0, 1)
```

It does fairly well: The expected density of the result is about $p(1+\sqrt{\frac{\ln \frac{n}{K}}{2Kp}})$, compared to density p of a random set. The expected maximum density of a K-node subgraph of $G_{n,p}$ (achievable through exponential exhaustive search) turns out to be — after a calculation paralleling that in [2] (see also [6]) — the solution to this equation $Kx \ln p = Kx \ln x - 2 \ln n$, which turns out to be $d_{\max} = \frac{2 \log n}{KW(\frac{2 \log n}{Kp})}$, where $W(\cdot)$ is the Lambert W function[5].

A competing algorithm is the *Cliques* algorithm: Let $c = \frac{\log n}{\log \frac{1}{p}}$, the maximum size of a clique that we know how to produce in $G_{n,p}$:

`Repeat` $\frac{K}{c}$ `times: create a clique of size` c.

The resulting density is $p + \frac{c}{K}$, which does not compare well with the present algorithm.

Another competitor is the *Greedy* algorithm:

`Repeat` $n - K$ `times: delete the lowest degree node.`

Greedy is hopelessly sequential (and thus irrelevant to our concerns here), and it is not known how to estimate its performance in $G_{n,p}$, but in experiments it does perform better than AssemblyCreation. Naturally, AssemblyCreation performs much better in $G^{++}_{n,p}$, arguably a more realistic model of synaptic connectivity.

8 A Distant Mirror: The Mouse Piriform Cortex

The memories in the human MTL discussed here are often termed "abstract," and not without justification. During sensory processing, the stimulus is coded, over several stages e.g. in the visual cortex, in a distributed way. This coding spatially reflects the perceived reality, in that features of the perceived reality (such as edges, frequencies, color, motion) are processed and coded by neural systems specializing in those features. After the conclusion of sensory processing, a process may be initiated, possibly mediated and supervised by some attention mechanism, that creates a new, sparse representation of the perceived item in the MTL, in which any links to the perceived world have been severed through random projection; this is the sense of abstraction imputed above.

A simple instance of this phenomenon has been identified recently in a rather unexpected place, the piriform cortex of the mouse [5]. Odorant molecules excite olfactory receptors in the animal's nose specializing in that molecule, and the axons of those excite in turn a small area (glomerulus) in the olfactory bulb; here again, each glomerulus specializes in one odor out of many hundreds. Next, the odorant's glomerulus projects strongly to the piriform cortex, creating a seemingly uniformly random — "abstract," disconnected from the outside world — sparse representation of the odorant. Here is the prescient interpretation in [5] of their experimental findings about this latter phenomenon:

1. *An odorant may [cause] a small subset of [...] neurons [in the piriform cortex to fire].*

[5] Many thanks to Cris Moore for help in this calculation

2. *This small fraction of [...] cells would then generate sufficient recurrent excitation to recruit a larger population of neurons.*
3. *The strong feedback inhibition resulting from activation of this larger population of neurons would then suppress further spiking.*
4. *In the extreme, some cells could receive enough recurrent input to fire [...] without receiving [initial] input.*

This narrative was an important inspiration for the present work, and the mathematical analysis of our model (Theorems 3.1 and 4.1) recalls it with rather striking fidelity.

9 Discussion and Further Research

We provide rigorous proof that, in a strongly simplified mathematical model of the brain, an assembly can emerge in response to spiking stimulus cells, will be exceptionally dense in synapses (a nontrivial algorithmic feat in a random network), and will fire consistently on future presentations of the stimulus; furthermore, upon a joint representation of two established stimuli, the assemblies will adapt by increasing their intersection (as has been observed in experiments). Despite the restrictions of our model, our probabilistic approach is quite robust, and we expect that several extensions can be obtained with further calculation. One such model would include a more realistic model of inhibition through a Gaussian synaptic input whose mean increases with excitatory activity (and no fixed K). Another extension would be to show robustness of assembly formation to perturbation of the stimulus, initial random excitatory activity, and noise. Also, it would be interesting to compare our results with simulations of biologically more realistic models, and to test experimentally if indeed assemblies in brains are more densely connected than random.

It would be interesting to see if the kind of mechanism hypothesized in this paper is present in other cognitive functions besides long term memory. One such is that of *visual invariants*, the mysterious ability by humans to identify, e.g., various rotations, postures, zooms, and occlusions of a familiar face, or even identify those visual images with a voice waveform or a string of characters. Our experiments show that, if two stimuli are presented together repeatedly, then the corresponding assemblies keep coming closer and closer; eventually they may become indistinguishable, and one can wonder if this mechanism cannot be part of the neural basis of invariants.

More ambitiously, what if *two* stimuli — or existing assemblies, encoding let us say to two parts of a sentence — are projected to another brain area (the same way a single stimulus is projected in the basic mechanism of this paper)? The assembly thus formed can be thought of as encoding a *Merge* [1] of the other two, that is, the new root of a syntax tree. Also, a mechanism akin to our assembly creation called *assembly pointer* has been studied recently through computational experiments [10]. An assembly pointer creates a copy of an extant assembly in a different brain area — perhaps a copy of the assembly for the lexical element "give" in another brain area where verbs get ready for syntax (see [15] for recent experimental evidence suggesting such activities in various areas of the cortex). It would be exciting to explore whether variants of the proposed mechanism can be the basis of beginning to understand how language is implemented in the Brain.

Acknowledgment. An inspiring discussion with Richard Axel on assemblies in the mouse's piriform cortex is gratefully acknowledged. This work was partially supported by the Human Brain Project of the European Union #604102 and #720270, and NSF grants CCF-1408635, CCF-1563838 and CCF-1717349.

References

1. Robert C Berwick and Noam Chomsky. *Why only us: Language and evolution*. MIT Press, 2016.
2. Béla Bollobás. Random graphs. In *Modern Graph Theory*, pages 215–252. Springer, 1998.
3. G Buzsaki. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, 68(3), 2010.
4. Uriel Feige, David Peleg, and Guy Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.
5. Kevin M Franks, Marco J Russo, Dara L Sosulski, Abigail A Mulligan, Steven A Siegelbaum, and Richard Axel. Recurrent circuitry dynamically shapes the activation of piriform cortex. *Neuron*, 72(1):49–56, 2011.
6. Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2015.
7. S Guzman, A J Schlö gl, M Frotscher, and P Jonas. Synaptic mechanisms of pattern completion in the hippocampal ca3 network. *Science*, 353(6304):1117–1123, September 2016.
8. Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, 1949.
9. Matias J Ison, Rodrigo Quian Quiroga, and Itzhak Fried. Rapid encoding of new memories by individual neurons in the human brain. *Neuron*, 87(1):220–230, 2015.
10. Robert Legenstein, Christos H Papadimitriou, Santosh Vempala, and Wolfgang Maass. Assembly pointers for variable binding in networks of spiking neurons. *arXiv preprint arXiv:1611.03698*, 2016.
11. A. Litwin-Kumar and B. Doiron. Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature communications*, 5, 2014.
12. Rodrigo Quian Quiroga. Concept cells: the building blocks of declarative memory. *Nature Reviews Neurosci.*, 13(8):587–597, 2012.
13. Rodrigo Quian Quiroga. Neuronal codes for visual perception and memory. *Neuropsychologia*, 83:227–241, 2016.
14. S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biology*, 3(3):e68, 2005.
15. Emiliano Zaccarella, Lars Meyer, Michiru Makuuchi, and Angela D Friederici. Building by syntax: The neural basis of minimal linguistic structures. 27, 10 2015.
16. F. Zenke, E. J. Agnes, and W. Gerstner. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature communications*, 6, 2015.

10 Appendix

Proof of Lemma 2. Let $W = \sum_i \sigma_i u_i v_i^T$ be the SVD of W, and $v = \sum_i \alpha_i v_i$. Without loss of generality, assume $\sum_i \alpha_i^2 = 1$. Then for any integer k,

$$\|W^k v\|^2 = v^T (W^T)^k W^k v = \sum_i \alpha_i^2 \sigma_i^{2k} = \mathsf{E}(X^{2k})$$

where the random variable X is equal to σ_i with probability α_i^2. Then $x(t)$ is proportional to $(v + Wv + \dots W^t v)$ and the desired inequality can be stated as follows:

$$\frac{v^T(I + W + \dots + W^{t+1})^T W^T W (I + W + \dots + W^{t+1})v}{v^T(I + W + \dots + W^{t+1})^T (I + W + \dots + W^{t+1})v} \geq \frac{v^T(I + W + \dots + W^{t})^T W^T W (I + W + \dots + W^{t})v}{v^T(I + W + \dots + W^{t})^T (I + W + \dots + W^{t})v}$$

which is equivalent to:

$$\mathsf{E}(X^2(1+X+\dots+X^{t+1})^2)\mathsf{E}((1+X+\dots+X^t)^2) \geq \mathsf{E}(X^2(1+X+\dots+X^t)^2)\mathsf{E}((1+X+\dots+X^{t+1})^2)$$

or

$$\mathsf{E}(\frac{X^2(1-X^{t+2})^2}{(1-X)^2})\mathsf{E}(\frac{(1-X^{t+1})^2}{(1-X)^2}) \geq \mathsf{E}(\frac{X^2(1-X^{t+1})^2}{(1-X)^2})\mathsf{E}(\frac{(1-X^{t+2})^2}{(1-X)^2}).$$

Define $f_1, f_2, g_1, g_2 : \mathbf{R}_+ \to \mathbf{R}_+$ to be each of the functions inside the expectations in the order above, so that the inequality is

$$\mathsf{E}(f_1(X))\mathsf{E}(f_2(X)) \geq \mathsf{E}(g_1(X))\mathsf{E}(g_2(X)).$$

Observe that for any X, we have

$$f_1(X)f_2(X) = g_1(X)g_2(X).$$

Moreover, for any X, Y, we claim that

$$f_1(X)f_2(Y) + f_1(Y)f_2(X) \geq g_1(X)g_2(Y) + g_1(Y)g_2(X).$$

For our choice of functions, this is

$$\frac{X^2(1-X^{t+2})^2}{(1-X)^2}\frac{(1-Y^{t+1})^2}{(1-Y)^2} + \frac{Y^2(1-Y^{t+2})^2}{(1-Y)^2}\frac{(1-X^{t+1})^2}{(1-X)^2} \geq \frac{X^2(1-X^{t+1})^2}{(1-X)^2}\frac{(1-Y^{t+2})}{(1-Y)^2} - \frac{Y^2(1-Y^{t+1})^2}{(1-Y)^2}\frac{(1-X^{t+2})^2}{(1-X)^2}$$

which is equivalent to

$$X^2(1-X^{t+2})^2(1-Y^{t+1})^2+Y^2(1-Y^{t+2})^2(1-X^{t+1})^2 \geq X^2(1-X^{t+1})^2(1-Y^{t+2})^2+Y^2(1-Y^{t+1})^2(1-X^{t+2})^2$$

or

$$(X^2 - Y^2)\frac{(1-X^{t+2})^2}{(1-Y^{t+2})^2} \geq (X^2 - Y^2)\frac{(1-X^{t+1})^2}{(1-Y^{t+1})^2}$$

which is always true. Therefore, we have

$$\begin{aligned}
&\mathsf{E}(f_1(X))\mathsf{E}(f_2(X)) - \mathsf{E}(g_1(X))\mathsf{E}(g_2(X)) \\
&= \sum_i \alpha_i f_1(X_i) \sum_i \alpha_i f_2(X_i) - \sum_i \alpha_i g_1(X_i) \sum_i \alpha_i g_2(X_i) \\
&= \sum_{i<j} \alpha_i \alpha_j (f_1(X_i) f_2(X_j) + f_1(X_j) f_2(X_i) - g_1(X_i) g_2(X_j) - g_1(X_j) g_2(X_i)) \\
&\quad + \sum_{i=j} \alpha_i^2 (f_1(X_i) f_2(X_i) - g_1(X_i) g_2(X_i)) \\
&\geq 0.
\end{aligned}$$

Moreover this holds with strict inequality unless $X = Y$, i.e., two of the singular values of W are equal. Thus the rate of convergence is at least the minimum singular value gap of W.

For the second part, note that the vector w consists of only the nonzero synapses into some cell j, and so the update rule on each synapse can be written as $w_{ij} = w_{ij} + \beta_j x_i$ where $\beta_j = \beta x_j$. Treating w and x as indexed only by i, the cells with synapses to a fixed j, we write

$$\begin{aligned}
\left(\frac{\tilde{w} \cdot x}{\|\tilde{w}\|_2}\right)^2 &= \frac{((w + \beta_j x) \cdot x)^2}{\|w + \beta_j x\|_2^2} \\
&= \frac{(w \cdot x)^2 + \beta_j^2 \|x\|_2^4 + 2\beta_j (w \cdot x)\|x\|_2^2}{\|w\|_2^2 + \beta_j^2 \|x\|_2^2 + 2\beta_j (w \cdot x)}
\end{aligned}$$

and need to show that this is greater than

$$\frac{(w \cdot x)^2}{\|w\|_2^2}.$$

Comparing, the inequality becomes

$$((w \cdot x)^2 + \beta_j^2 \|x\|_2^4 + 2\beta_j (w \cdot x)\|x\|_2^2)\|w\|_2^2 > (w \cdot x)^2 (\|w\|_2^2 + \beta_j^2 \|x\|_2^2 + 2\beta_j (w \cdot x))$$

which is implied by the Cauchy-Schwartz inequality:

$$\|w\|_2^2 \|x\|_2^2 \geq (w \cdot x)^2$$

applied twice. ◀