Mechanisms of distributed working memory in a large-scale model of the macaque neocortex

Jorge F. Mejias[1], Xiao-Jing Wang[2]

[1]Swammerdam Institute for Life Sciences, University of Amsterdam, 1098XH Amsterdam, Netherlands
[2]Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003, USA
Contact: j.f.mejias@uva.nl; xjwang@nyu.edu.

Working memory, the brain's ability to retain and manipulate information internally, has been traditionally associated with persistent neural firing in localized brain areas such as those in the frontal cortex (Fuster 1973, Funahashi et al., 1989; Goldman-Rakic 1995; Romo et al., 1999; Rigotti et al., 2013; Kopec et al., 2015; Inagaki et al., 2019). However, self-sustained neural persistent activity during working memory is present in multiple brain regions (Romo et al., 2004; Christophel et al., 2017; Leavitt et al., 2017; Sreenivasan and D'Esposito, 2019), the underlying mechanism is unknown. We developed an anatomically constrained large-scale computational model of the macaque cortex endowed with a macroscopic gradient of synaptic excitation, to investigate the origin of distributed working memory representation. We found that long-range inter-areal reverberation can support the emergence of persistent activity patterns across multiple cortical regions, by virtue of a robust bifurcation in space, even when none of isolated local areas is capable of generating persistent activity. The model uncovered a host of distinct persistent activity patterns (attractor states), and provides experimentally testable predictions that cannot be explained in local circuit models. Simulating cortical lesions reveals that distributed activity patters are resilient against simultaneous lesions of multiple cortical areas, but depend on areas that form the core of the entire cortex. This work provides a theoretical framework for identifying large-scale brain mechanisms and computational principles of distributed cognitive processes.

Our computational model includes 30 cortical areas distributed across all four neocortical lobes (Fig. 1a; see Supplementary Methods for further details). The inter-areal connectivity is based on quantitative connectomic data from tract-tracing studies of the macaque monkey (Markov et al., 2013; Extended Data Fig. 1). Each of the cortical areas is modeled as a neural circuit which contains two selective excitatory populations and one inhibitory population (Fig. 1b) (Wang, 2001; Wong and Wang, 2006). In addition, there is a macroscopic gradient of synaptic excitation (Chaudhuri et al., 2015; Joglekar et al., 2018; Extended Data Fig. 2), namely the number of spines, loci of excitatory synapses, per pyramidal cell (Elston 2007) was used as a proxy for the strength of recurrent and long-range excitation that increases along the cortical hierarchy (Felleman and van Essen, 1991; Markov et al., 2014; Fig. 1c). To allow for the propagation of activity from sensory to association areas, inter-areal long-distance connections target more strongly excitatory neurons than inhibitory neurons for more feedforward pathways, but biased in the opposite direction for more feedback pathways, in a graded fashion (Mejias et al., 2016; Fig. 1b).

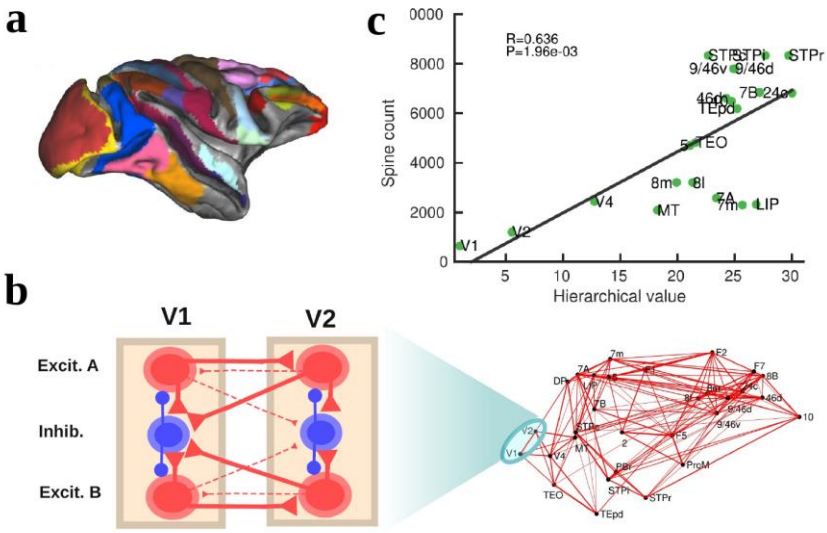


Figure 1: Scheme and anatomical basis of a multi-regional macaque neocortex model. *(a) Lateral view of the macaque cortical surface with areas in color. (b) In the model, inter-areal connections are calibrated by mesoscopic connectomic data (Markov et al., 2012), each parcellated area is modeled by a population firing rate description with two selective excitatory neural pools and an inhibitory neural pool (Wong and Wang, 2006). (c) Correlation between spine count data (Elston, 2007) and anatomical hierarchy as defined by layer-dependent connections in Markov et al., (2014).*

In local circuit models of working memory (WM) (Wang, 2001; Compte et al., 2006), areas high in the cortical hierarchy make use of sufficiently strong synaptic connections (notably involving NMDA receptors, see Wang et al., 2013) to generate self-sustained persistent activity. Specifically, the strength of local synaptic reverberation must exceed a threshold level (in our model, the local coupling parameter J_S must be larger than a critical value of 0.4655), for an isolated local area to produce stimulus-selective persistent activity states that coexist with a resting state of spontaneous activity (operating in a multistable regime rather than in a monostable regime (Fig. 2a)). However, there is presently no conclusive experimental demonstration that an isolated cortical area like dorsolateral prefrontal cortex (dlPFC) is indeed capable of generating mnemonic persistent activity. In this study, we first examined the scenario in which all areas, including dlPFC (9/46d) at the top of the hierarchy, have J_S values below the critical value for multistability (Fig. 2a). In this case, any observed persistent activity pattern must result from inter-areal connection loops. In a simulated delayed response task, a transient visual input excites a selective neural pool in the primary visual cortex (V1), which yielded activation of other visual areas such as MT during stimulus presentation (Fig. 2b, upper left). After stimulus withdrawal, neural activity persists in multiple areas across frontal, temporal and parietal lobes (Fig. 2b, lower right). This activation pattern was stimulus specific, so only the neural pool selective for the shown stimulus in each cortical area displayed elevated persistent activity (Fig. 2c; Extended Data Fig. 3). The same result is obtained when stimulating other sensory modalities (Extended Data Fig. 4) and, if simplified AMPA dynamics is considered, also for brief stimuli (Extended Data Fig. 5). We observed cross-area variations of neural dynamics: while areas like TEpd displayed a sharp binary jump of activity, areas like LIP exhibited a more gradual ramping activity, resembling temporal accumulation of information in decision-making (Shadlen and Newsome, 2001).

Consistent with experimental observations (Leavitt et al., 2017), early sensory areas such as V1 and MT did not display persistent activity. This is ensured in our model by a combination of two

factors. First, local synaptic coupling of early sensory areas at the bottom of the hierarchy are weak. Second, a certain level of preferential targeting inhibitory neurons by top-down projections prevents indiscriminate persistent activation across all cortical areas (Fig. 2d). Such bias towards inhibitory neurons of feedback projections, supported by experimental evidence (Tsushima et al., 2006), also allows the distributed WM patters to exist for a wide range of global scaling values for long-range synaptic strengths (Fig. 2e).

When we plotted the firing rate of stimulus-selective persistent activity across 30 areas along the hierarchy, our results revealed a gap that separated the areas displaying persistent activity and those that did not (Fig. 2f). This is a novel type of bifurcation or abrupt transition of behavior that takes place in space, rather than as a function of a network parameter like in Fig. 2a. As a matter of fact, the relevant parameter here is the strength of synaptic excitation that varies across cortical space (Extended Data Fig. 6), in the form of a macroscopic gradient. Therefore, bifurcation is robust and does not require precisely tuning a parameter to a threshold value.

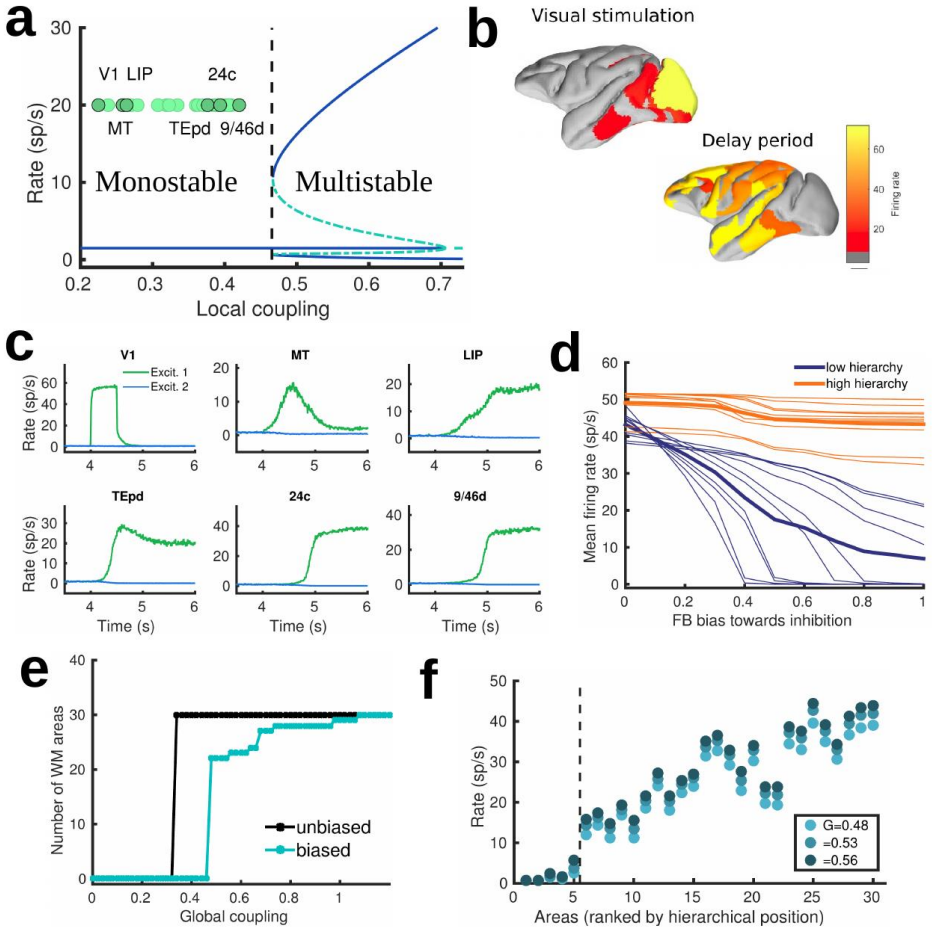


Figure 2: Distributed WM sustained via long-range loops in cortical networks*. (a) Bifurcation diagram for an isolated area. Here, all areas (green circles) are in the monostable regime when isolated. (b) Spatial activity map during visual stimulation (upper left) and delay period (lower right). (c) Activity of selected cortical areas during the WM task. (d) Firing rate of areas at the bottom and top of the hierarchy (10 areas each, thick lines denote averages) as a function of the feedback preference to inhibitory neurons. (e) Number of areas in the example distributed activity pattern vs global coupling strength for inhibitory-preferred FB vs neutral FB. (f) Firing rate for all areas ranked by hierarchical position, for several global coupling values.*

We recognized that a large-scale circuit can potentially display a large number of distributed persistent activity patterns (attractors), some of them may not be accessible by stimulation of a primary sensory area. Note that distinct attractor states are defined here in terms of their spatial patterns, not stimulus tuning; for the sake of simplicity we assumed only two selective neural pools per local area. We developed a numerical approach to identify and count distinct attractors. Our aim is not to exhaustively identify all possible attractors (as the parameter space is too large for that), but to get insight on how our estimations depend on relevant parameters such as the global scaling factor G of all local and long-range excitatory connections, or and the maximum area-specific synaptic strength J_{max}. Five examples are shown in Fig. 3a. Our analysis included four cases; two of them with the strength of local and long-range connectivity above the bifurcation threshold for certain areas high in the hierarchy (Fig. 3b), so that some areas like dlPFC have strong enough local reverberation to sustain activity independently, while other areas like TEpd and LIP require long-range support to participate in WM. In all four cases, the number of attractors is a function of the scaling factor G, with an optimal G value maximizing the number of attractors (Fig. 3c). This optimal G value shifted towards lower values as the proportion of intrinsically multistable areas increased, with peak number of attractors simultaneously increasing (Fig. 3d). Across all four cases and G values considered, we found a significant positive correlation between the number of areas involved in a given attractor and the average activity level of these areas (Fig. 3e). With a high proportion of intrinsically multistable areas, attractors tend to be largely restricted to these multistable areas (located at the top of the hierarchy), while in cases with zero or low proportion of multistable areas attractors involve a larger number of areas and are more diverse in their area composition (Fig. 3f).

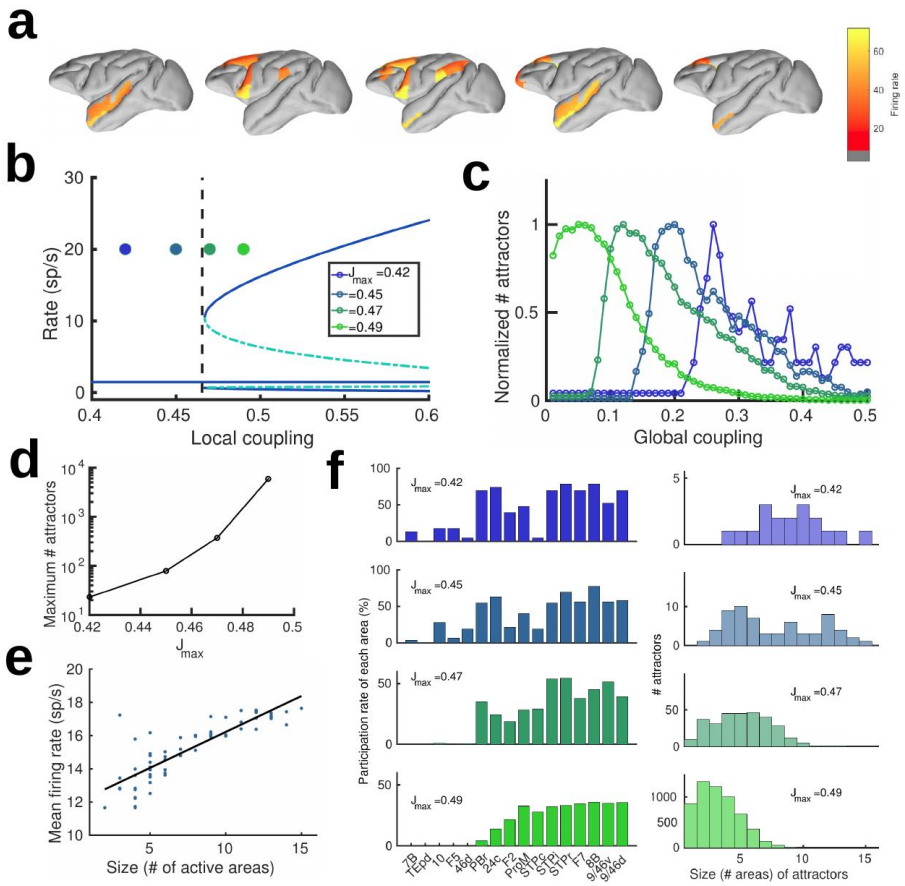


Figure 3: Distributed and local WM mechanisms can coexist in the model. *(a) Five example attractors of the network (Jmax=0.42). (b) Bifurcation diagram with the four cases considered. (c)*

Normalized number of attractors found via numerical exploration as a function of the global coupling for all four cases. (d) Maximum (peak) number of attractors for each one of the cases. (e) Correlation between size of attractors and mean firing rate of its constituting areas for Jmax=0.45 and G=0.2. (f) Participation index of each area (left) and distribution of attractors according to their size (right).

The distributed nature of WM has implications for the impact of perturbations on performance. We simulated a delayed response task with distractors (Fig. 4a), in which if the to-be-remembered cue is A, B is presented as distractor during the delay period (and vice versa), and found that only strong distractors (compared to the original visual cue) are able to destabilize the memory (Fig. 4b). This is due to the robustness of a distributed attractor compared to a local circuit mechanism, but also to the inhibitory effect of feedback projections which dampen the propagation of distractor signals (cf. responses in V1 and MT).

Finally, we tested the effect of lesioning cortical areas in WM tasks. Lesions of individual areas have only a local effect on the overall maintenance of the distributed attractor, even when lesioning prefrontal areas (Fig. 4c). The number of areas involved in the evoked attractor decreases only linearly with the number of simultaneously (randomized) lesioned areas (Fig. 4d) and decreases a bit more abruptly when lesioning areas in reverse hierarchical order (Extended Data Fig. 7). A more systematic evaluation revealed that lesioning most areas have limited consequences for the total number of available attractors, with the exception of several temporal and prefrontal areas for which the overall impact is large (Fig. 4e; Extended Data Fig. 7). Interestingly, the latter areas are part of the anatomical 'bowtie hub' of the macaque cortex (Fig. 4f, Markov et al., 2013).

In summary, we have presented a large-scale circuit mechanism of distributed working memory, realized by virtue of a new concept of robust bifurcation in space. The distributed WM scenario is compatible with recent observations of multiple cortical areas participating in WM tasks (Christophel et al., 2017; Leavitt et al., 2017; Sreenivasan and D'Esposito, 2019), even when such areas have not been traditionally associated with WM. Interestingly, the model uncovered a host of distinct persistent activity attractor states, defined by their spatial distributed patterns in the large-scale cortical system. Many of these persistent activity states are not produced by stimulation of primary sensory areas; they could represent various forms of internal computations and functions independent of direct sensory inputs. This model serves as a starting point for future research in several directions. First, the model presented here is limited to 30 areas, and can be expanded to include other cortical areas as their connectivity data become available. Second, the model can be improved by incorporating more biological detailed such as contributions of AMPA receptor and NMDA receptor mediated excitation, as well as various types of inhibitory neurons. Third, attractors do not have to be steady states, our model can be generalized to account for a rich repertoire of temporal dynamics during working memory (Mongillo et al., 2008; Lim and Goldman, 2013; Rigotti et al., 2013; Miller et al., 2018; Bouchacourt and Buschman, 2019). Our model yields several experimentally testable predictions in monkey and rodent experiments, including (i) the need of strong large-scale interactions to sustain distributed WM patterns, (ii) a positive correlation between the number of areas involved in a WM task and their average firing rate of persistent activity, and (iii) robustness of working memory encoding against distractors via inter-areal inhibitory feedback. This model and its extensions can serve as a computational platform to elucidate complex experimental observations, such as inactivation by optogenetic method of various cortical areas in behaving animals performing a working memory task, in future research. Conceptually, this work showed a novel form of bifurcation in space as a mechanism to generate differential functions across different cortical areas, and represents the first large-scale cortical model for distributed cognitive processes.

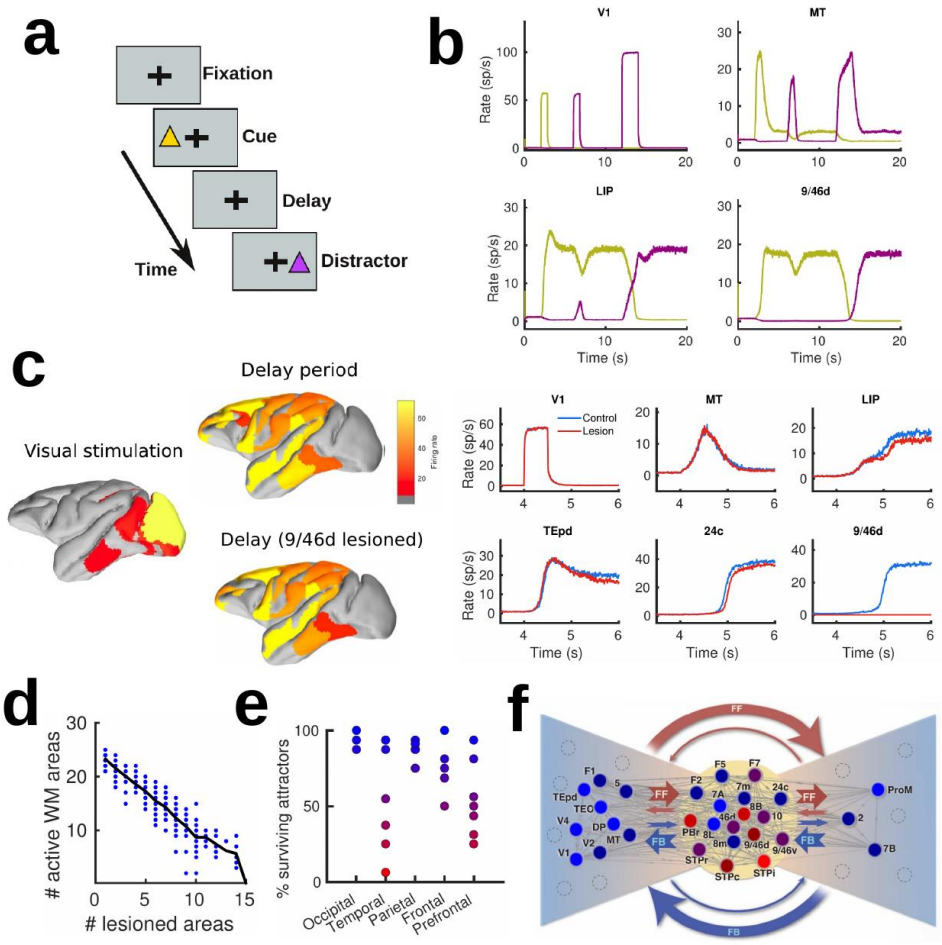


Figure 4: Effects of distractors and lesions on the network. *(a) Scheme of the WM task with distractors. (b) Activity traces of selected areas (color code denotes input selectivity). (c) Spatial activity maps during stimulation and delay periods for control (top) and 9/46d lesioned case (bottom). Traces on the right show the impact for selected areas. (d) Number of active areas in the example attractor as a function of the number of (randomly selected) lesioned areas. (e) Numerical exploration of the percentage of surviving attractors for area lesions in different lobes. (f) Lesions to areas at the center of the 'bowtie hub' have a stronger impact on WM (adapted from Markov et al., 2013).*

Acknowledgments: This work was supported by the NIH grant R01MH062349 and the Simons Foundation Collaboration on the Global Brain grant (to XJW).

References:

1. Bouchacourt, F., Buschman, T. J. A flexible model of working memory. Neuron 103, 147-160 (2019).
2. Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H., Wang, X.-J. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. Neuron 88, 419-431 (2015).
3. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., Haynes, J. D. The distributed nature of working memory. Trends Cog. Sci. 21 (2), 111-124 (2017).

4. Compte, A. Computational and in vitro studies of persistent activity: edging towards cellular and synaptic mechanisms of working memory. Neuroscience 139, 1:135-51 (2006).
5. Elston, G. N. Specialization of the neocortical pyramidal cell during primate evolution. In Evolution of Nervous Systems: a Comprehensive Reference, Volume 4. Kass JH and Preuss TM eds. (Elsevier), 191-242 (2007).
6. Felleman, D. J., Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1 (1), 1-47 (1991).
7. Funahashi, S. et al. Mnemonic coding of visual space in the monkey primate cerebral cortex. Neurophysiol. 61, 331-349 (1989).
8. Fuster, J. M. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. J. Neurophysiol., 36, 1:61-78 (1973).
9. Goldman-Rakic, P. S. Cellular basis of working memory. Neuron 14, 477-485 (1995).
10. Inagaki, H. K., Fontolan, L., Romani, S. and Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature, 566, 212-217 (2019).
11. Joglekar, M. R., Mejias, J. F., Yang, G. R. and Wang, X.-J. Inter-areal balanced amplification enhances signal propagation in a large-scale circuit model of the primate cortex. Neuron, 98, 222-234 (2018).
12. Kopec, C. D., Erlich, J. C., Brunton, B. W., Deisseroth, K. and Brody, C. D. Cortical and Sub-cortical Contributions to Short-Term Memory for Orienting Movements. Neuron, 88, 367-77 (2015).
13. Leavitt, M. L., Mendoza-Halliday, D., Martinez-Trujillo, J. C. Sustained Activity Encoding Working Memories: Not Fully Distributed. Trends Neurosci. 40, 328-346 (2017).
14. Lim, S. and Goldman, M. S. Balanced cortical microcircuitry for maintaining information in working memory. Nat. Neurosci. 16, 1306-1314 (2013).
15. Markov, N. T. et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cereb. Cortex 24, 17-36 (2012).
16. Markov, N. T. et al. Cortical high-density counterstream architectures. Science 342, 1238406, (2013).
17. Markov, N. T. et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J. Comp. Neurol. 522, 225-259 (2014).
18. Mejias, J. F., Murray, J. D., Kennedy, H. and Wang, X.-J. Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex, Sci. Adv. 2, e1601335 (2016).
19. Miller, E. K., Lundqvist, M. and Bastos, A. M. Working memory 2.0. Neuron 100, 463-475 (2018).
20. Mongillo, G., Barak, O. and Tsodyks, M. Synaptic theory of working memory. Science 319, 1543-1546 (2008).
21. Murray, J. D., Bernacchia, A., Roy, N. A. Constantinidis C, Romo R, Wang X-J, Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. USA 114, 394-399 (2017).
22. Rigotti, M. et al., The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585-590 (2013).
23. Romo, R., Brody, C. D., Hernández, A. and Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. Nature, 399, 470-473 (1999).
24. Romo, R., Hernández, A. and Zainos, A. Neuronal correlates of a perceptual decision in ventral premotor cortex. Neuron, 41, 165-73 (2004).
25. Shadlen, M. N. and Newsome, W. T. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. 86, 1916-1936 (2001).
26. Sreenivasan, K. K. and D'Esposito, M. The what, where and how of delay activity. Nat. Rev.

Neurosci. 20, 466-481 (2019).
27. Tsushima, Y., Sasaki, Y. and Watanabe, T. Greater disruption due to failure of inhibitory control on an ambiguous distractor. Science, 314, 1786-8 (2006).
28. Wang, X.-J. Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 24, 455-463 (2001).
29. Wong, K. F. and Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. 26, 1314-1328 (2006).
30. Wang, M. et al. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. Neuron 77, 736-749 (2013).

Methods

Anatomical data

The anatomical connectivity data used has been gathered in an ongoing track tracing study in macaque and has been described in detail elsewhere (Markov et al., 2012; 2013; 2014; Mejias et al., 2016). Briefly, retrograde tracer injected into a given target area labels neurons in a number of source areas projecting to the target area. By counting the number of labeled neurons on a given source area, Markov et al. defined the fraction of labeled neurons (FLN) from that source to the target area. FLN can serve as a proxy for the 'connection strength' between two cortical areas, which yields the connectivity pattern of the cortical network (Extended Fig. 1a, b). In addition, Markov et al. also measured the number of labeled neurons located on the supragranular layer of a given source area. Dividing this number over the total number of labeled neurons on that source area, we can define the supragranular layered neurons (SLN) from that source area to the target area (Extended Fig. 1c, d).

SLN values may be used to build a well-defined anatomical hierarchy (Felleman and Van Essen, 1991; Markov et al. 2014). Source areas located lower (higher) than the target area in the anatomical hierarchy, as defined by Felleman and Van Essen (1991), display a progressively higher (lower) proportion of labeled neurons in the supragranular layer. As a consequence, the lower (higher) the source area relative to the target area, the higher (lower) the SLN values of the source-to-target projection.

Iterating these measurements across other anatomical areas yields and anatomical connectivity matrix with weighted directed connections and an embedded structural hierarchy. The 30 cortical used to build our data-constrained large-scale brain network are, in hierarchical order: V1, V2, V4, DP, MT, 8m, 5, 8l, 2, TEO, F1, STPc, 7A, 46d, 10, 9/46v, 9/46d, F5, TEpd, PBr, 7m, LIP, F2, 7B, ProM, STPi, F7, 8B, STPr and 24c. Finally, data on wiring connectivity distances between cortical areas is available for this dataset as well, allowing to consider communication time lags when necessary (we found however that introducing time lags this way does not have a noticeable impact on the dynamics of our model). The connectivity data used here is available to other researchers from core-nets.org.

The corresponding 30x30 matrices of FLN and SLN are shown in Extended Fig. 1b, d. Areas in these matrices are arranged following the anatomical hierarchy, which is computed using the SLN values and a generalized linear model (Chaudhuri et al., 2015; Mejias et al., 2016). Surgical and histology procedures were in accordance with European requirements 86/609/EEC and approved by the ethics committee of the region Rhone-Alpes.

In addition to the data on FLN and SLN across 30 cortical areas, we used additional data to constrain the area-to-area differences in the large-scale brain network. In particular, we have collected data on the total spine count of layer 2/3 pyramidal neuron basal dendrites across different cortical areas, as the spine count constitutes a proxy for the density of synaptic connections within a given cortical area (see Elston, 2007 for a review). A full list of all area-specific values of spine densities considered and their sources is given below:

Rank in SLN hierarchy	Area name	Measured spine count	Age correction factor*	Source
1	V1	643	1	Elston and Rosa, 1997; Elston et al., 1999, p1369
2	V2	1201	1	Elston and Rosa 1997, p444
3	V4	2429	1	Elston and Rosa, 1998b, p287
4	DP	-	-	
5	MT	2077	1	Elston et al., 1999, p1369
6	8m	3200	1.30	Elston and Rosa, 1998a, p128
7	5	4689	1	Elston and Rockland, 2002, p1073
8	8l	3200	1.30	Elston and Rosa, 1998a, p128
9	2	-	-	
10	TEO	4812	1	Elston and Rosa, 1998b, p287
11	F1	-	-	
12	STPc	8337	1	Elston et al., 1999, p1369
13	7a	2572	1	Elston and Rosa, 1997, p444
14	46d	6600	1.15	estim. Elston 2007 (fig. 17); Elston et al., Frontiers 2011
15	10	6488	1.15	Elston et al., Frontiers 2011
16	9/46v	7800	1.15	estim. Elston, 2007 (fig. 17)
17	9/46d	7800	1.15	estim. Elston, 2007 (fig. 17)
18	F5	-	-	
19	TEpd	7260	1	Elston et al., 1999, p1369
20	PBr	-	-	
21	7m	2294	1.30	Elston, 2001, p146
22	LIP	2316	1	Elston and Rosa, 1997, p444
23	F2	-	-	
24	7B	6841	1	Elston and Rockland, 2002, p1073
25	ProM	-	-	
26	STPi	8337	1	Elston et al., 1999, p1369
27	F7	-	-	
28	8B	-	-	
29	STPr	8337	1	Elston et al., 1999, p1369
30	24c	6825	1.15	Elston et al., 2005, p67

The age correction factor is meant to correct for the decrease of spine counts with age for data obtained from old monkeys. A plausible estimate would be a ~30% decrease for a 10y difference (Duan et al., 2003, p955; Young et al., 2014, p36). See Extended Fig. 2 for the effect of this correction on the overall gradient established by the spine count data, and the correlation of such gradient with the SLN hierarchy.

Computational model

Local neural circuit

We describe the neural dynamics of the local microcircuit representing a cortical area with the Wong-Wang model (Wong and Wang, 2006). In its three-variable version, this model describes the temporal evolution of the firing rate of two input-selective excitatory populations as well as the evolution of the firing rate of an inhibitory population. All populations are connected to each other (see Fig. 1a). The model is described by the following equations:

$$\frac{dS_A}{dt} = -\frac{S_A}{\tau_N} + \gamma\,(1 - S_A)\,r_A \qquad \text{(Eq. 1)}$$

$$\frac{dS_B}{dt} = -\frac{S_B}{\tau_N} + \gamma\,(1 - S_B)\,r_B \qquad \text{(Eq. 2)}$$

$$\frac{dS_C}{dt} = -\frac{S_C}{\tau_G} + \gamma_I\,r_C \qquad \text{(Eq. 3)}$$

Here, S_A and S_B are the NMDA conductances of selective excitatory populations A and B respectively, and S_C is the GABAergic conductance of the inhibitory population. Values for the constants are τ_N=60 ms, τ_G=5 ms, γ=1.282 and γ_I=2. The variables r_A, r_B and r_C are the mean firing rates of the two excitatory and one inhibitory populations, respectively. They are obtained by solving the transcendental equation $r_i = \phi_i(I_i)$ at each time step, with Ii being the input to population I, given by

$$I_A = J_s S_A + J_c S_B + J_{EI} S_C + I_{0A} + I^A_{net} + x_A(t) \qquad \text{(Eq. 4)}$$

$$I_B = J_c S_A + J_s S_B + J_{EI} S_C + I_{0B} + I^B_{net} + x_B(t) \qquad \text{(Eq. 5)}$$

$$I_C = J_{IE} S_A + J_{IE} S_B + J_{II} S_C + I_{0C} + I^C_{net} + x_C(t) \qquad \text{(Eq. 6)}$$

In these expressions, J_s, J_c are the self- and cross-coupling between excitatory populations, respectively, J_{EI} is the coupling from the inhibitory populations to any of the excitatory ones, J_{IE} is the coupling from any of the excitatory populations to the inhibitory one, and J_{II} is the self-coupling strength of the inhibitory population. The parameters I_{0i} with i=A, B, C are background inputs to each population. Parameters are J_s=0.3213 nA, J_c=0.0107 nA, J_{IE}=0.15 nA, J_{EI}=-0.31 nA, J_{II}=-0.12 nA, I_{0A}=I_{0B}=0.3294 nA and I_{0C}=0.26 nA. Later we will modify some of these parameters in an area-specific manner (in particular J_s and J_{IE}) to introduce a gradient of properties across the cortical hierarchy. The term I^i_{net} denotes the long-range input coming from other areas in the network, which we will keep as zero for now but will be detailed later. Sensory stimulation can be introduced here as extra pulse currents of strength I_{pulse}=0.2 and duration T_{pulse}=0.5 sec (unless specified otherwise).

The last term $x_i(t)$ with i=A, B, C is an Ornstein-Uhlenbeck process, which introduces some level of stochasticity in the system. It is given by

$$\tau_{noise}\,\frac{dx_i}{dt} = -x_i + \sqrt{\tau_{noise}}\,\sigma_i\,\xi_i(t) \qquad \text{(Eq. 7)}$$

Here, $\xi_i(t)$ is a Gaussian white noise, the time constant is τ_{noise}=2 ms and the noise strength is $\sigma_{A,B}$=0.005 nA for excitatory populations and σ_C=0 for the inhibitory one.

The transfer function $\phi_i(t)$ which transform the input into firing rates takes the following form for the excitatory populations (Abbott and Chance, 2005):

$$\phi_{A,B}(I) = \frac{1}{2}\,\frac{aI-b}{1-\exp[-d\,(aI-b)]} \qquad \text{(Eq. 8)}$$

The values for the parameters are a=135 Hz/nA, b=54 Hz and d=0.308 s. For the inhibitory population a similar function can be used, but for convenience we choose a threshold-linear function:

$$\phi_C(I) = \frac{1}{g_I}\,(c_1 I - c_0) + r_0 \qquad \text{(Eq. 9)}$$

The values for the parameters are g_I=4, c_1=615 Hz/nA, c_0=177 Hz and r_0=5.5 Hz. Finally, it is sometimes useful for simulations (although not a requirement) to replace the transcendental equation $r_i = \phi_i(I_i)$ by its analogous differential equation, of the form

$$\tau_r \frac{dr_i}{dt} = -r_i + \phi_i(I_i) \qquad \text{(Eq. 10)}$$

The time constant can take a typical value of τ_r=2 ms.

Gradient of synaptic strengths

Before considering the large-scale network and the inter-areal connections, we look into the area-to-area heterogeneity to be included in the model.

Our large-scale cortical system consists of N=30 local cortical areas, for which inter-areal connectivity data is available. Each cortical area is described as a Wong-Wang model of three populations like the ones described in the previous section. Instead of assuming areas to be identical to each other, here we will consider some of the natural area-to-area heterogeneity that has been found in anatomical studies. For example, work from Elston (2007) has identified a gradient of dendritic spine density, from low spine numbers found in early sensory areas to large spine counts found in higher cognitive areas. This may reflect an increase of local recurrent strength as we move from sensory to association areas. In addition, cortical areas are distributed along an anatomical hierarchy (Felleman and Van Essen, 1991; Markov et al. 2012). The position of a given area within this hierarchy can be computed via a logistic regression of the SLN (fraction of supragranular layer neurons) projecting to and from that area (as in Chaudhuri et al., 2015).

In the following, we will assign the incoming synaptic strength (both local and long-range) of a given area as a linear function of the dendritic spine count values observed in anatomical studies, with age-related corrections when necessary. Alternatively, when spine count data is not available for a given area, we will use its position in the anatomical hierarchy, which displays a high correlation with the spine count data, as a proxy for the latter. After this process, the large-scale network will display a gradient of local and long-range recurrent strength, with sensory/association areas showing weak/strong local connectivity, respectively. We denote the local strength value of a given area i in this gradient as h_i, and this value normalized between zero (bottom of the gradient, area V1) and one.

We assume therefore a gradient of values of J_s, with its value going from J_{min} to J_{max}. Having large values of J_s for association areas strongly affects the spontaneous activity of these areas, even without

considering inter-areal coupling. A good way to keep the spontaneous firing rate of these areas within physiologically realistic limits is to impose that the spontaneous activity fixed point is the same for all areas (Murray et al., 2017b). To introduce this into the model, we take into account that the solutions in the spontaneous state are symmetrical: S_A=S_B=S (we assume zero noise for simplicity). The current entering any of the excitatory populations is then (assuming I_{0A}=I_{0B}=I_0):

$$I = (J_s + J_c)S + J_{EI}S_C + I_0 \qquad \text{(Eq. 11)}$$

Assuming a fast dynamics for r_C and S_C (mediated by GABA) as compared to S_A and S_B (mediated by NMDA) we can obtain the approximate expression for S_C:

$$S_C \simeq \tau_G \gamma_I r_C = 2SJ_{IE}\zeta + \beta \qquad \text{(Eq. 12)}$$

with

$$\zeta = \frac{\tau_G \gamma_I c_1}{g_I - J_{II}\tau_G \gamma_I c_1} \qquad \text{(Eq. 13)}$$

$$\beta = \tau_G \gamma_I \frac{c_1 I_{0C} + g_I r_0 - c_0}{g_I - J_{II}\tau_G \gamma_I c_1} \qquad \text{(Eq. 14)}$$

The equation for the excitatory current has then the form

$$I = (J_s + J_c)\, S + 2\, J_{EI}\, J_{IE}\zeta\, S + J_{EI}\, \beta + I_0 \qquad \text{(Eq. 15)}$$

To maintain the excitatory input (and therefore the spontaneous activity level S) constant while varying J_s across areas, we just have to keep the quantity $J_s + J_c + 2\, J_{EI}\, J_{IE}\, \zeta \equiv J_0$ constant (for the original parameters of the isolated area described above, we obtain J_0=0.2112 nA). A good choice, but not the only one, is to assume that the excitatory synapses to inhibitory neurons, J_{IE}, also scales with the ranks and with J_s accordingly:

$$J_{IE} = \frac{1}{2\, J_{EI}\, \zeta}\, (J_0 - J_s - J_c) \qquad \text{(Eq. 16)}$$

This linear relationship ensures that the spontaneous solution is the same for all areas in the network. Since J_{IE} needs to be non-negative, this imposes a minimum value of J_{min}=0.205 nA for J_s. The particular maximum value of J_s, namely J_{max}, will determine the type of WM model we assume. Since the bifurcation point of an isolated area is at J_s=0.4655 nA for this set of parameter values, setting J_{max} below that value implies that all areas in the network are monostable in isolation. In this situation, any persistent activity displayed by the model will be a consequence of a global, cooperative effect due to inter-areal interactions. On the other hand, having J_{max} above the bifurcation point means that some areas will be multistable when isolated, e.g. they will be intrinsically multistable and compatible with classical WM theories.

Unless specified otherwise, we assume a range of J_{min}=0.21 nA and J_{max}=0.44 nA (i.e. below the critical value), so that the model displays distributed WM.

Inter-areal projections

We now consider the inter-areal projections connecting isolated areas to form the large-scale cortical

network. Assuming that inter-areal projections stem only from excitatory neurons (as inhibitory projections tend to be local in real circuits) and that such projections are selective for excitatory neurons, the network or long-range input term arriving at each of the populations of a given area *y* from all other cortical areas is given by

$$I^y_{A,net} = G \sum_x W^{xy} SLN^{xy} S^y_A \quad \text{(Eq. 17)}$$

$$I^y_{B,net} = G \sum_x W^{xy} SLN^{xy} S^y_B \quad \text{(Eq. 18)}$$

$$I^y_{C,net} = \frac{G}{Z} \sum_x W^{xy} (1 - SLN^{xy})(S^y_A + S^y_B) \quad \text{(Eq. 19)}$$

Here, a superindex denotes the cortical area and a subindex the particular population within each area. The sum in all equations runs over all cortical areas of the network (N=30). Excitatory populations A and B receive long-range inputs from equally selective units from other areas, while inhibitory populations receive inputs from both excitatory populations. Therefore, neurons in population A of a given area may be influenced by A-selective neurons of other areas directly, and by B-selective neurons of other areas indirectly, via local interneurons.

G is the global coupling strength, which controls the overall long-range projection strength in the network (G=0.48 unless specified otherwise). Z is a factor that takes into account the relative balance between long-range excitatory and inhibitory projections. Setting Z=1 means that both excitatory and inhibitory long-range projections are equally strong, but this does not guarantee that their effect is balanced in the target area, due to the effect of local connections. Following Murray et al., (2017b), we choose to impose a balance condition that guarantees that, if populations A and B have the same activity level, their net effect on other areas will be zero –therefore highlighting the selectivity aspect of the circuits. Considering that the transfer function of inhibitory populations is linear and their approximately linear rate-conductance relationship, it can be shown that

$$Z = \frac{2 c_1 \tau_G \gamma_I J_{EI}}{c_1 \tau_G \gamma_I J_{II} - g_I} \quad \text{(Eq. 20)}$$

Aside from global scaling factors, the effect of long-range projections from population *x* to population *y* is influenced by two factors. The first one, W^{xy}, is the anatomical projection strength as revealed by tract-tracing data from Markov et al. (2013). We use the fraction of labelled neurons (FLN) from population *x* to *y* to constrain our projections values to anatomical data. We rescale these strengths to translate the broad range of FLN values (over five orders of magnitude) to a range more suitable for our firing rate models. We use a rescaling that maintains the proportions between projection strengths, and therefore the anatomical information, that reads

$$W^{xy} = k_1 \, (FLN^{xy})^{k_2} \quad \text{(Eq. 21)}$$

Here, the values of the rescaling are k_1=1.2 and k_2=0.3. The same qualitative behavior can be obtained from the model if other parameter values, or other rescaling functions, are used as long as the network is set into a standard working regime (i.e. signals propagate across areas, global synchronization is avoided, etc). FLN values are also normalized so that $\sum_y FLN^{xy} = 1$. In addition, and as done for the local connections, we introduce a gradient of long-range projection strengths using the spine count data: $W^{xy} \rightarrow \lambda^x W^{xy}$, where λ_x is one for the target area *x* with the maximal spine count, and decreases for other areas with the same slope as the gradient of the local connectivity presented above.

The second factor that needs to be taken into account is the directionality of signal propagation across the hierarchy. Feedforward (FF) projections that are preferentially excitatory constitute a reasonable assumption which facilitate signal transmission from sensory to higher areas. On the other hand, having feedback (FB) projections with a preferential inhibitory nature contributes to the emergence of realistic distributed WM patterns (Fig. 2d,e) (see also Markov et al., 2014; Tsushima et al., 2006). This feature can be introduced, in a gradual manner, by linking the different inter-areal projections with the SLN data, which provides a proxy for the FF/FB nature of a projection (SLN=1 means purely FF, and SLN=0 means purely FB). In the model, we assume a linear dependence with SNL for projections to excitatory populations and with (1-SLN) for projections to inhibitory populations, as shown above.

We limit the modulation of the FB projections between frontal areas to a maximum of 0.25 so that interactions between frontal areas are never strongly inhibitory, in agreement with evidence of frontal networks having strong excitatory loops (Markowitz et al., 2015). This consideration is important to allow FEF (areas 8l and 8m) to exhibit some level of persistent activity during distributed WM –as their hierarchical position and recurrent strength are not strong enough to sustain activity otherwise – but it does not affect the behavior of our model otherwise.

Data analysis

We developed a numerical method to estimate the number of stable distributed WM attractors for a particular set of parameters values of our large-scale model. This method is used to obtain the results shown in Figs. 3 and 4. To allow for a cleaner estimation, we do not consider noise in the neural dynamics during these simulations.

Our large-scale cortical model has 30 areas, with each of them having two selective excitatory populations A and B. Simply assuming that each of the areas can reach one of three possible states (persistent activity in A, persistent activity in B, or spontaneous activity) means that our model can potentially display up to 3^{30} attractor combinations. This number can be even larger if we refine the firing rate reached by each area rather than simply its persistent/non-persistent activity status. Since it is not possible to fully explore this extremely large number of possible attractors, we devised a strategy based on the exploration of a sample of the input space of the model. The core idea is to stimulate the model with a certain input pattern (targeting randomized areas) and registering the fixed point that the dynamics of the model converges to. By repeating this process with a large number of input combinations and later counting the number of different attractors from the obtained pool of fixed points, we can obtain an estimate of the number of attractors for a particular set of parameter values.

Stimulation protocol

A given input pattern is defined as a current pulse of fixed strength (I_{pulse} =0.2) and duration (T_{pulse} =1 sec) which reaches a certain number P of cortical areas. Only one population (A or B, randomized) in each area receives the input, and the P cortical areas receiving the input are randomly selected across the top 16 areas of the spine count gradient. This decreases the amount of potential input combinations we have to deal with by acknowledging that areas with stronger recurrent connections (such as 9/46d) are more likely to be involved in distributed WM patterns than those with weaker connections (such as MT). P can take any value between one and P_{max} =16, and we run a certain number of trials (see below) for each of them. Different values of I_{pulse} and T_{pulse}, as well as setting the randomly selected areas at a high rate initial condition instead of providing an external input, have been also explored and lead to

qualitatively similar results.

It is also important to consider that not all values of P have the same number of input combinations. For example, P=1 allows for 16*2=32 different input combinations (if we discriminate between populations A and B), while P=2 allows for 16*(16-1)*2=480 input combinations, and so on. For a given value of P, the number of possible input combinations N_c is given by

$$N_c = 2^P \binom{P_{max}}{P} = 2^P \frac{P_{max}!}{(P_{max}-P)!\,P!} \qquad \text{(Eq. 22)}$$

By summing all values of N_c for P=1, …P_{max}, we obtain around 43 million input combinations, which are still too many trials to simulate for a single model configuration. To simplify this further, we consider a scaling factor F_c on top of N_c to bring down these numbers to reasonable levels for simulations. We use F_c =0.0002 (or 0.02% of all possible combinations) for our calculations, which brings down the total number of simulated input combinations to around 9000. Other options, such as decreasing P_{max} and using a larger scaling factor (P_{max} =12, F_c =0.01 or 1% or all possible combinations) give also good results. Since the rescaling can have a strong impact for small P (yielding a number of trials smaller than one), we ensure at least one trial for these cases.

To guarantee the stability of the fixed points obtained during these simulations, we simulate the system during a time windows of 30 seconds (which is much larger than any other time scale in the system), and check that the firing rates have not fluctuated during the last 10 seconds before we register the final state of the system as a fixed point.

Estimating the number of attractors

The final step is to count how many different attractors have been reached by the system, by analyzing the pool of fixed points obtained from simulations. A simple way to do this is to consider that, for any fixed point, the state of each area can be classified as persistent activity in population A (i.e. mean firing rate above a certain threshold of 10 spikes/s), persistent activity in population B, or spontaneous activity (both A and B are below 10 spikes/s). This turns each fixed point into a vector of 30 discrete states, and the number of unique vectors among the pool of fixed points can be quickly obtained using standard numerical routines in Matlab (such as the 'unique' function).

A more refined way to count the number of attractors, which we use in this work, is to define an Euclidean distance to discriminate between an attractor candidate and any previously identified attractors. Once the first attractor (i.e. the first fixed point analyzed) is identified, we test whether the next fixed point is the same than the first one by computing the Euclidean distance E_d between them:

$$E_d = \frac{1}{n} \sum_{i=1}^{n} (r_i^{new} - r_i^{old})^2 \qquad \text{(Eq. 23)}$$

where n=30 is the total number of areas in the network (only one of the populations, A or B, needs to be considered here). If E_d is larger than a certain threshold distance ε, we consider it a new attractor. We choose ε=0.01, which grossly means that two fixed points are considered as different attractors if, for example, the activity of one of their cortical areas differs by 0.5 spikes/s and the activity on all other areas is the same for both. The particular value of ε does not have a strong impact on the results (aside from the fact that smaller values of ε gives us more resolution to find attractors). When several attractors are identified, each new candidate is compared to all of them using the same method.

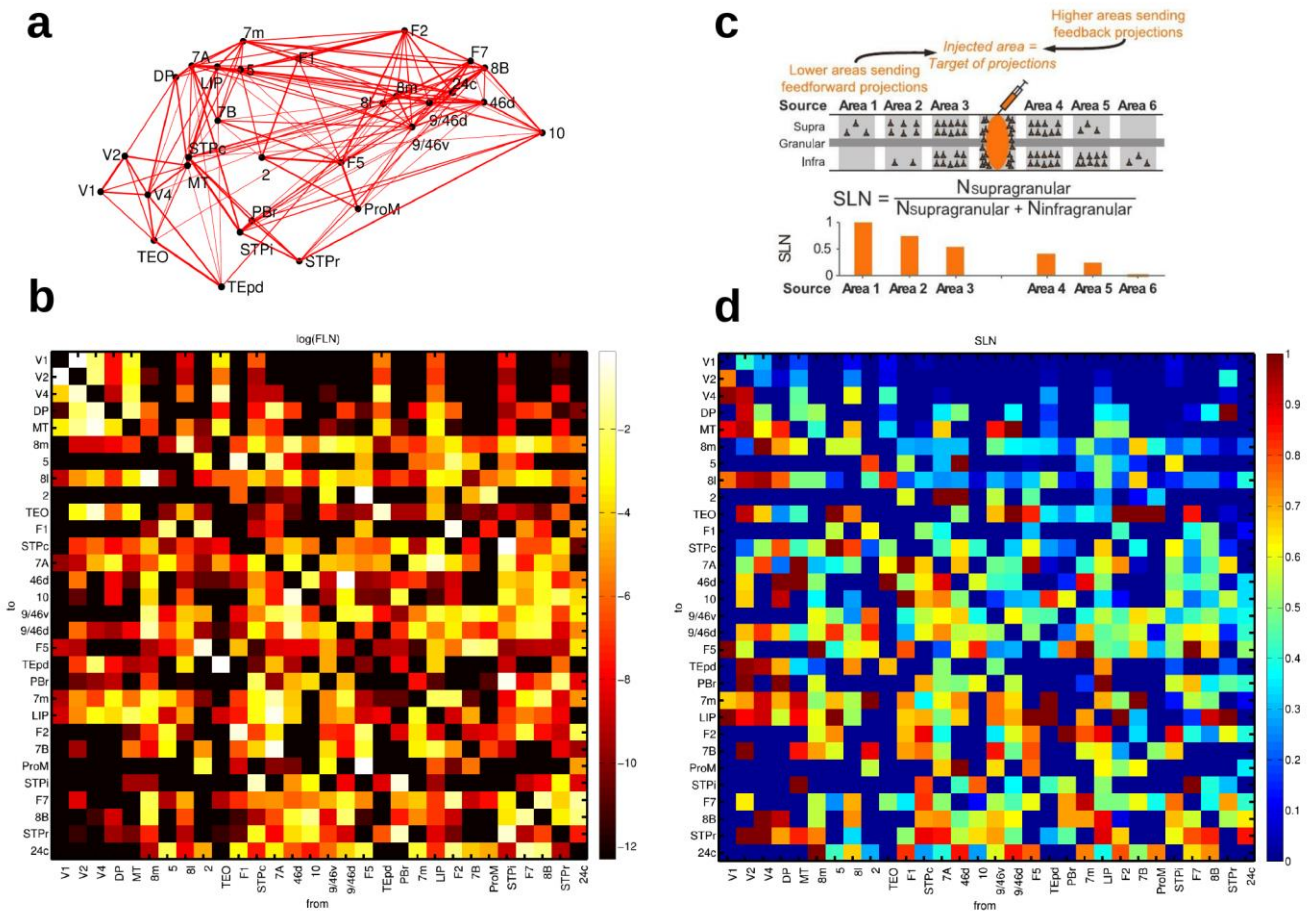
Both the first and the second method to count attractors deliver qualitatively similar results (in terms of the dependence of the number of attractors with model parameters), although as expected the second method yields larger numbers due to its higher discriminability.

Code availability
The code of the present model will be released upon publication of this manuscript.

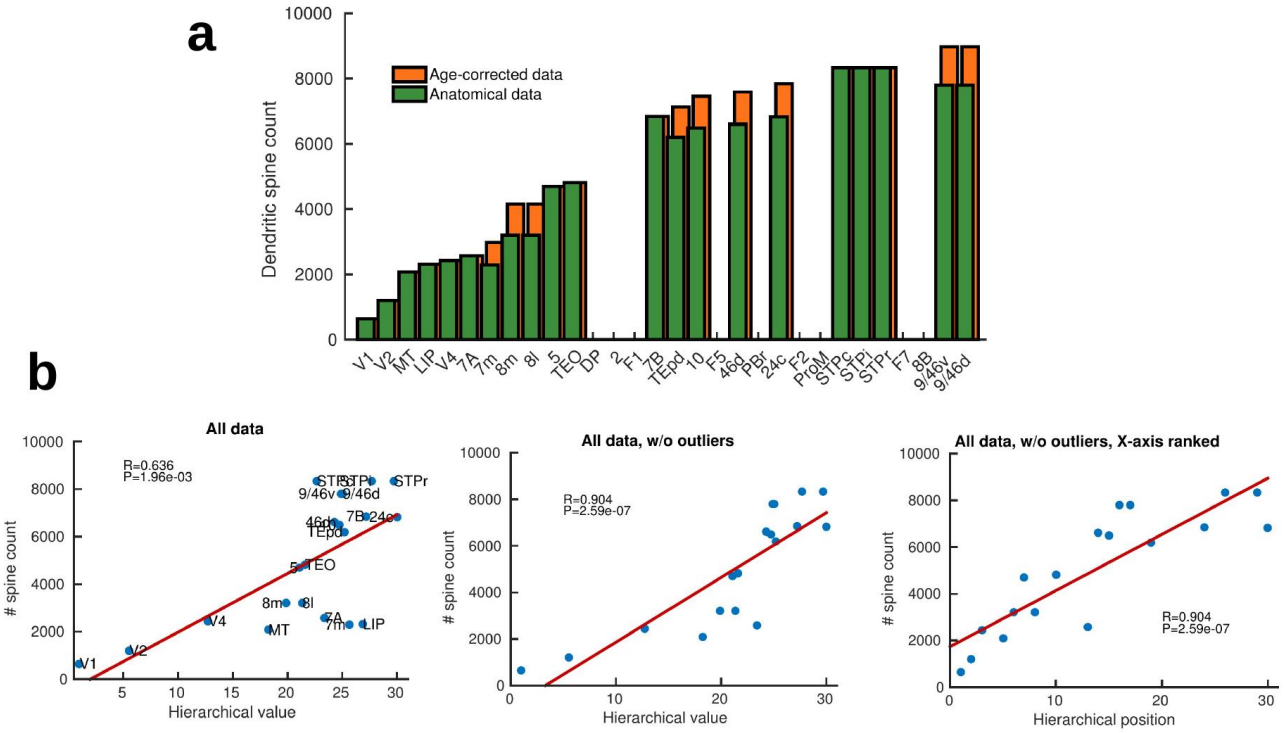
Data availability
All data used in this manuscript is already available at core-nets.org or from the cited literature.

Extended references

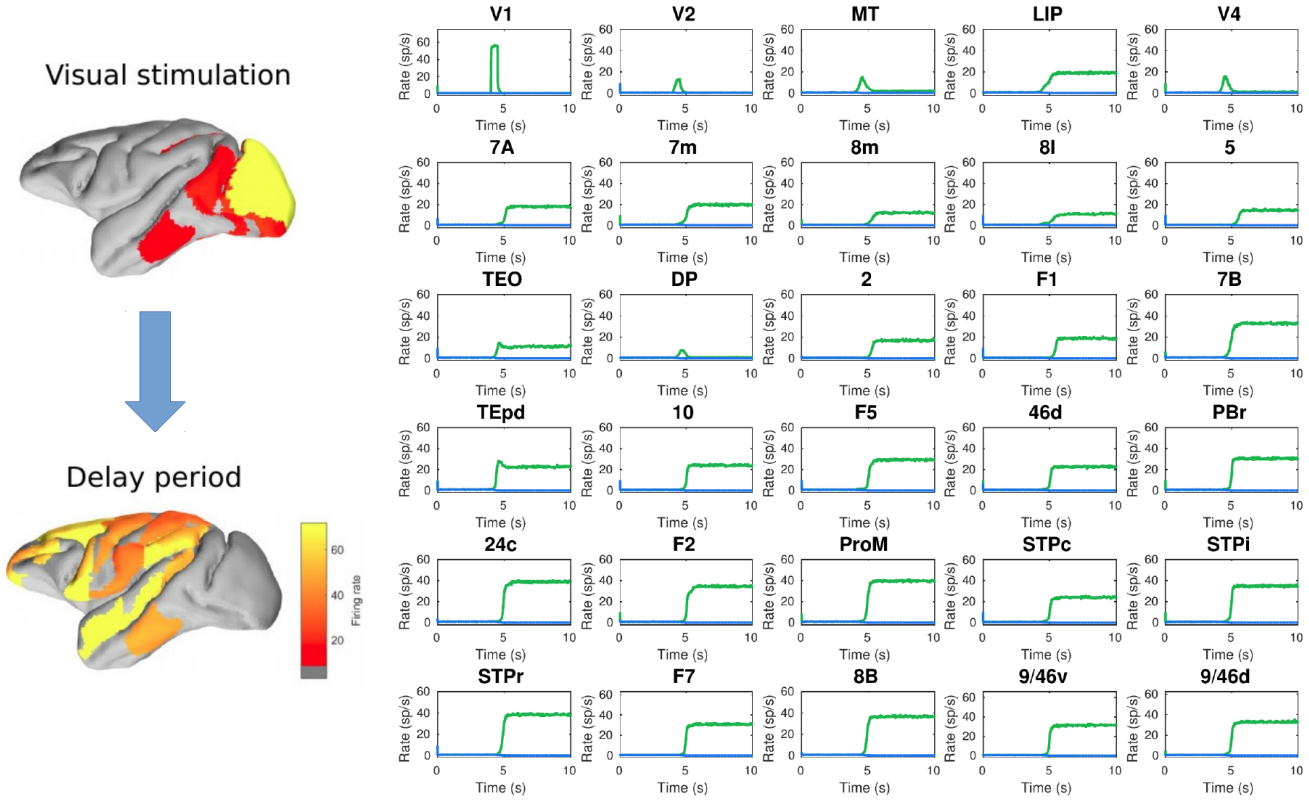
31. Abbott, L. F. and Chance, F. S. Drivers and modulators from push-pull and balanced synaptic input. Prog. Brain Res. 149, 147-155 (2005).
32. Duan, H. et al. Age-related dendritic and spine changes in corticocortically projecting neurons in macaque monkeys. Cereb. Cortex 13, 950-971 (2003).
33. Elston, G. N. and Rosa, M. G. The occipitoparietal pathway of the macaque monkey: comparison of pyramidal cell morphology in layer III of functionally related cortical visual areas. Cereb. Cortex 7, 432-452 (1997).
34. Elston, G. N. and Rosa, M. G. Complex dendritic fields of pyramidal cells in the frontal eye field of the macaque monkey: comparison with parietal areas 7a and LIP. Neuroreport 9, 127-131 (1998a).
35. Elston, G. N. and Rosa, M. G. Morphological variation of layer III pyramidal neurones in the occipitotemporal pathway of the macaque monkey visual cortex. Cereb. Cortex 8, 278-294 (1998b).
36. Elston, G. N., Tweedale, R. and Rosa, M. G. Cortical integration in the visual system of the macaque monkey: large-scale morphological differences in the pyramidal neurons in the occipital, parietal and temporal lobes. Proc. Biol. Sci. 266, 1367-1374 (1999).
37. Elston, G. N. Interlaminar differences in the pyramidal cell phenotype in cortical areas 7 m and STP (the superior temporal polysensory area) of the macaque monkey. Exp. Brain Res. 138, 141-152 (2001).
38. Elston, G. N. and Rockland, K. S. The pyramidal cell of the sensorimotor cortex of the macaque monkey: phenotypic variation. Cereb. Cortex 12, 1071-1078 (2002).
39. Elston, G. N., Benavides-Piccione, R. and De Felipe, J. A study of pyramidal cell structure in the cingulate cortex of the macaque monkey with comparative notes on inferotemporal and primary visual cortex. Cereb. Cortex 15, 64-73 (2005).
40. Elston, G. N., Benavides-Piccione, R., Elston, A., Manger, P. R. and De Felipe, J. Pyramidal cells in prefrontal cortex of primates: marked differences in neuronal structure among species. Front. Neuroanat. 5:2 (2011).
41. Markowitz, D. A., Curtis, C. E. and Pesaran, B. Multiple component networks support working memory in prefrontal cortex. Proc. Natl. Acad. Sci. USA 112, 11084-11089 (2015).
42. Murray, J. D., Jaramillo, J. and Wang, X.-J. Working Memory and Decision-Making in a Frontoparietal Circuit Model. J. Neurosci. 37, 12167-12186 (2017b).
43. Young, M. E., Ohm, D. T., Dumitriu, D., Rapp, P. R. and Morrison, J. H. Differential effects of aging on dendritic spines in visual cortex and prefrontal cortex of the rhesus monkey. Neuroscience 274, 33-43 (2014).



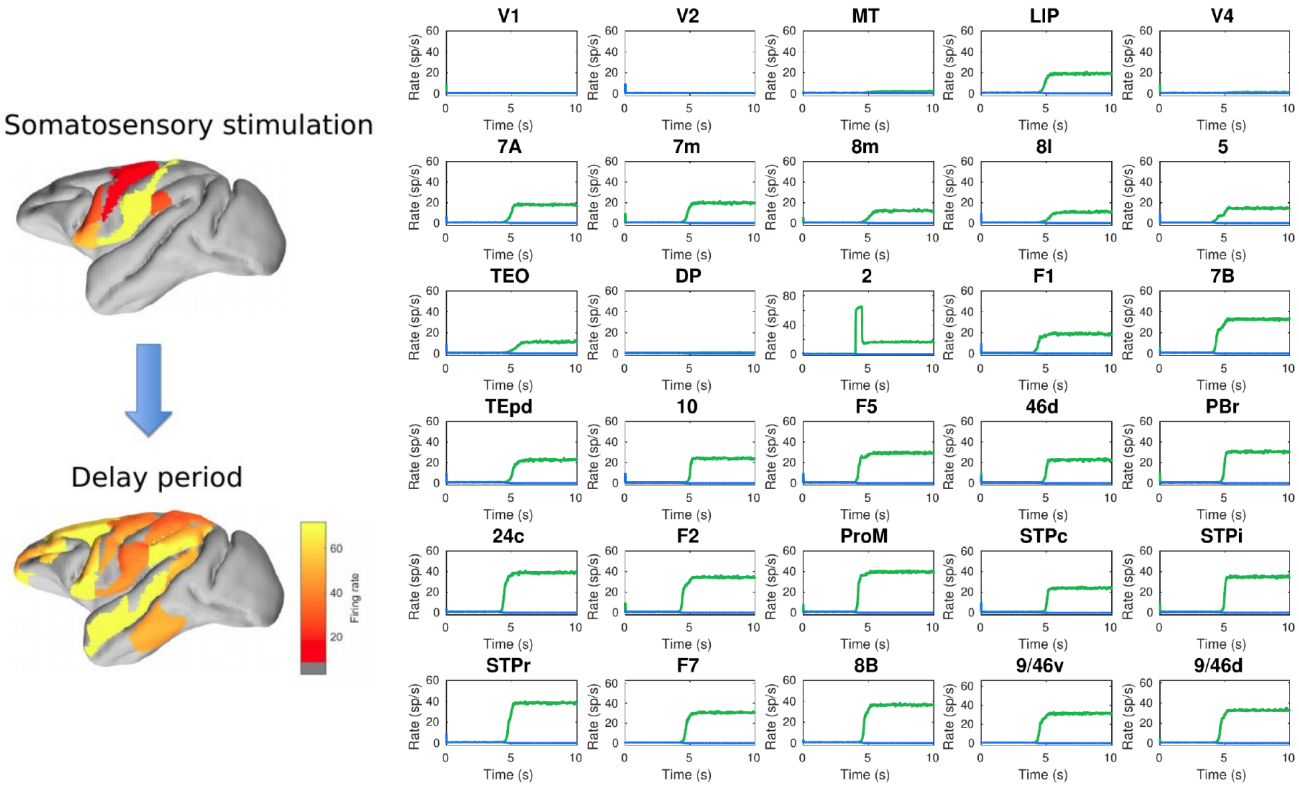
Extended Data Fig. 1: Anatomical connectivity data of the macaque cortex*. Data from Markov et al., 2013; 2014. (a) Connectivity of the 30 areas (positioned in 3D space following injection coordinates of experiments). Width of the lines denote two-way averaged FLN values (i.e. average strength of the projection). (b) Map of FLN values for all connections considered. (c) The proportion of supragranular vs infragranular neurons projecting to the injection site allowed to define an anatomical hierarchy. (d) Map of SLN values for all connections considered.*



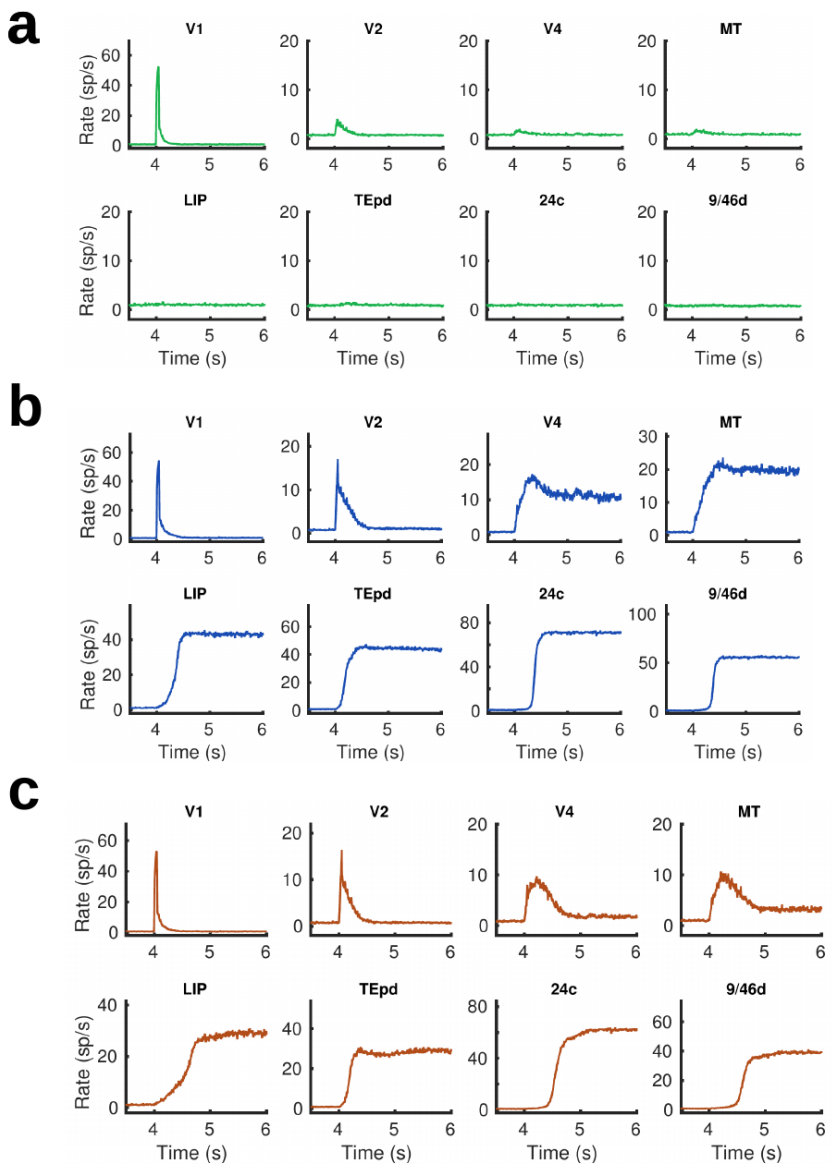
Extended Data Figure 2: Spine count data used to constrain connectivity strength. *Data from Elston, 2007 and others. (a) Spine count of the basal dendrites of layer 2/3 neurons across cortical areas of young (2~3 years old) macaques. When data from older macaques had to be considered, an age correction was introduced (orange bars). (b) Correlation between the spine count data and the hierarchical value for all areas (left), the hierarchical value for all areas except outliers (7m, LIP, and STPc), and the hierarchical position (rank) for all areas except outliers. Pearson correlation and corresponding P-value are shown in each panel. In the model, the connectivity strength of areas for which spine count data was not available was estimated using their hierarchical value.*



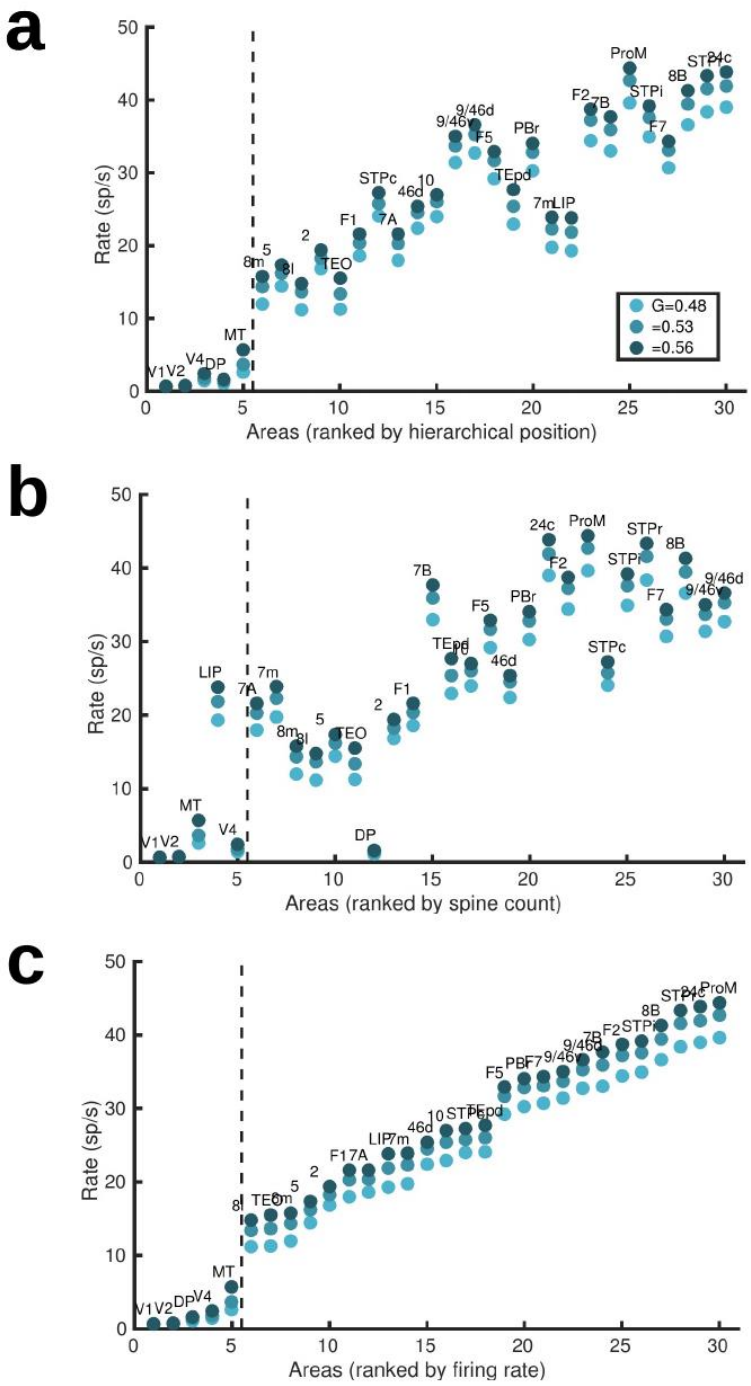
Extended Data Figure 3: Behavior of all areas in the network during the visual WM task. *Left: spatial maps of the simulated macaque brain during stimulation and delay period, with activity color coded. Right: evolution of the firing rate of all areas in the network. Stimulation occurs in V1 at t=4 seconds and has a duration of 500ms.*



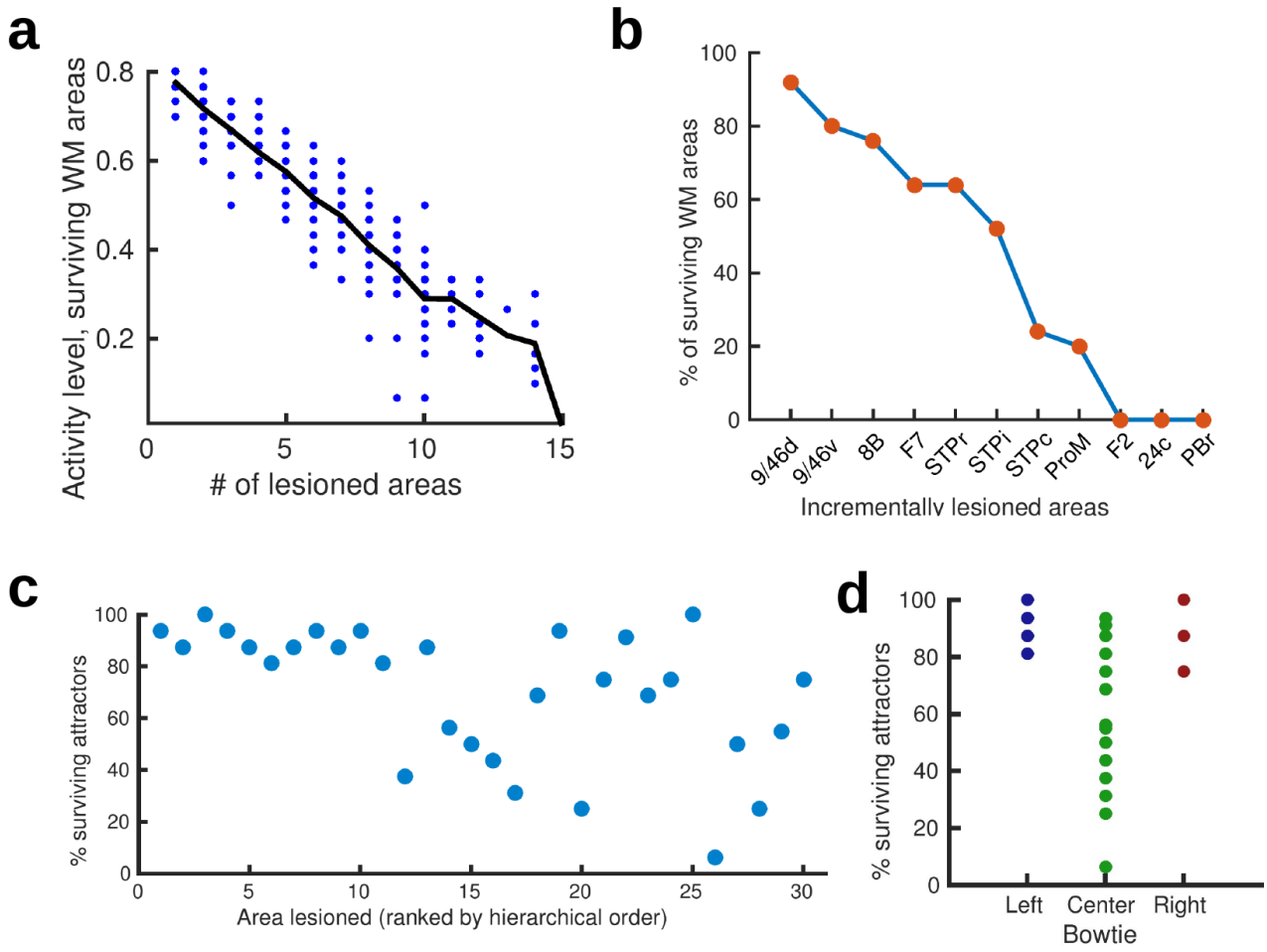
Extended Data Figure 4: Behavior of all areas in the network during the somatosensory WM task. *Left: spatial maps of the simulated macaque brain during stimulation and delay period, with activity color coded. Right: evolution of the firing rate of all areas in the network. Stimulation occurs in area 2 (primary somatosensory area) at t=4 seconds, and has a duration of 500ms.*



Extended Data Figure 5: Firing rates for selected areas during a visual WM task with a short (50ms) stimulus duration. *(a) Only NMDA and GABA synapses are considered in the network, the stimulus does not reach frontal areas. (b) By introducing simplified AMPA-like synapses (proportional to the firing rates) on all excitatory projections, we obtain distributed WM patterns with some early sensory areas (such as V4 and MT) now displaying persistent activity. (c) By limiting the AMPA-like synapses to feedforward projections, the original distributed WM patterns (with early sensory areas remaining within spontaneous levels) are recovered. This suggests that AMPA/NMDA asymmetry along the FF/FB projections in the hierarchy has to be considered for brief stimulation patterns.*



Extended Data Figure 6: Firing rate of ranked cortical areas reveals a robust bifurcation in space with different ranking systems. *(a) Areas ranked following their hierarchical (as obtained from SLN data) position. (b) Areas ranked by spine count. (c) Areas ranked by displayed firing rate. All panels show data for three different values of the global coupling strength.*



Extended Data Figure 7: Effect of lesioning areas on the stability of attractors. *(a) The activity level, or firing rate, or areas involved in a visually evoked distributed WM attractor decreases linearly with the number of lesioned areas. (b) The percentage of areas involved in the distributed WM attractor decreases more sharply when areas are consecutively lesioned in reverse hierarchical order. (c) The percentage of attractors surviving lesions is specific of the area lesioned, with lesions in areas higher in the hierarchy having a stronger impact. (d) Lesioning areas at the center of the anatomical bowtie hub has a stronger impact than lesioning the 'side areas'. Both (c) and (d) were obtained during the numerical exploration of the stable distributed attractors in the network.*