Statistical mechanics of spike events underlying phase space partitioning and sequence codes in large-scale models of neural circuits

Maximilian Puelma Touzel*
Max Planck Institute for Dynamics and Self-Organization, Goettingen, Germany, and Mila, Université de Montréal, Montréal, Canada

Fred Wolf
Max Planck Institute for Dynamics and Self-Organization, Goettingen, Germany, Physics Department, Georg August University, Goettingen, Germany, Bernstein Center for Computational Neuroscience, Goettingen, Germany, and Kavli Institute for Theoretical Physics, University of California Santa Barbara, Santa Barbara, USA

Cortical circuits operate in an inhibition-dominated regime of spiking activity. Recently, it was found that spiking circuit models in this regime can—despite disordered connectivity and asynchronous, irregular activity—exhibit a locally stable dynamics that may be used for neural computation. The lack of existing mathematical tools has precluded analytical insight into this phase. Here we present analytical methods tailored to the granularity of spike-based interactions for analyzing attractor geometry in high-dimensional spiking dynamics. We apply them to reveal the properties of the complex geometry of trajectories of population spiking activity in a canonical model of locally stable spiking dynamics. We find that attractor basin boundaries are the pre-images of spike-time collision events involving connected neurons. These spike-based instabilities control the divergence rate of neighboring basins, and have no equivalent in rate-based models. They are located according to the disordered connectivity at a random subset of edges in a hypercube representation of the phase space. Backward-iterating these edges using the stable dynamics induces a partition refinement on this space that converges to the attractor basins. We formulate a statistical theory of the locations of such events relative to attracting trajectories via a tractable representation of local trajectory ensembles. Averaging over the disorder, we derive the basin diameter distribution, whose characteristic scale emerges from the relative strengths of the stabilizing inhibitory coupling and destabilizing spike interactions. Our study provides an approach to analytically dissect how connectivity, coupling strength, and single neuron dynamics shape the phase space geometry in the locally stable regime of spiking neural circuit dynamics.

Dynamics is determined by the geometry of trajectories in its associated phase space. In high-dimensional models of disordered neural circuits, for example, trajectories from phases of locally stable dynamics are structured around a large set of coexisting attractors dispersed throughout the phase space. Using a highly idealized model, Hopfield demonstrated how to construct a version of this neural dynamics that implements a high-capacity, error-correcting code [1], in which corrupted versions of the stable states are corrected by the locally stable dynamics, endowing it with robustness to noise. Could such a model be at work in the brain? Since many model details can be suppressed when describing collective behaviour, system idealization need not compromise the veracity of the description. However, models aiming to capture collective dynamics should qualitatively capture the activity statistics of the regime under study and, since collective states of many-body systems typically depend strongly on the form of the interactions between the bodies, do so using a faithful representation of the type of interaction. In this regard, Hopfield-like constructions relying on local stability in models more faithful to the underlying neurobiology have proved elusive.

While a continuously interacting *rate dynamics* admits powerful statistical methods through which results like Hopfield's have been well understood [2–4], neurons rather interact at a discrete set of spike times [5]. Despite proving computationally powerful [6], the granular character of spikes makes many of these methods inadmissible, complicating the analysis of *spiking dynamics*. Further complication arises since spiking dynamics deviates from the expectations of smooth dynamical systems theory. As a salient example, we consider asynchronous, irregular spiking activity, that is reminiscent of chaotic dynamics, but surprisingly does not preclude local stability. Indeed, a locally stable phase of spiking dynamics has been found in a variety of models [7–10] operating in the inhibition-dominated regime exhibited by cortical circuit activity [11, 12], achievable with $\mathcal{O}(1/\sqrt{K})$ interaction strength. In our current state of knowledge, weak ($\mathcal{O}(1/K)$) and strong ($\mathcal{O}(1)$) interaction strengths give only the more conventional pairing of stable dynamics with temporally regular and vanishing/exploding spiking activity, respectively, and are thus unsuitable for modeling cortex. Moreover and in contrast to rate-based encoding schemes employing the local stability of fixed points or limit cycles, cortical circuits rather encode inputs using time-varying, intrinsically-generated activity. While chaos produces such activity and emerges rather generically with sufficient sampling of a disordered connectiv-

ity in both rate and spiking models, its error-amplifying nature appears to negate the kind of error-correcting schemes required for robust encoding.

Spiking dynamics in the locally stable phase, by contrast, can partition the phase space into a large set of tube-shaped basins of attraction, termed *flux tubes*, each enclosing a single attracting trajectory [13, 14], and together in principle providing an error-correcting encoding useful for neural computation. Whether described by the state sequence at spike times or the index-sequence of spiking neurons, flux tube attractors thus reflect the activity statistics of cortical activity, and do so using a more faithful representation of the interaction.

Reference [14] significantly advanced our understanding of the geometry of this locally stable phase, with numerical scaling results for the divergence rate and average diameter of flux tube attractor basins. This advance stopped short, however, of providing a picture of the phase space geometry associated with these attractors, the topic of the present study, for three important reasons. First, the analysis of Ref.[14] only consider a single time slice of phase space around these attractors. The time-varying nature of the attractors strongly suggests a time-varying basin diameter. Second, the analysis of Ref. [14] never precisely locate the attractor boundaries, and so never access the discrete nature of the boundary, leaving the putative instability responsible for the boundary unexplained. These two missing pieces are not only fundamental to the phenomenology, but likely also provide important insights into the overall attractor geometry. Without them, Ref. [14] employed a numerical approach that, as a third gap, left the scaling dependencies unexplained (the experiments also omitted the dependency on the strength of neural interactions). Thus, Ref. [14] provided limited mechanistic insight into the phenomena, e.g. into how the model ingredients contribute and why. Since the constraints that dynamics places on computation and the capacity of any putative neural code are ultimately controlled by these dependencies, a theory is needed in which they are jointly derived and can be understood in terms more naturally related to the dynamics.

Here, we provide these missing pieces of the phenomenology and use them to build a theory with which we provide a more complete understanding of the phase space geometry of flux tubes in the networks considered by [14]. We first present a simulation study of flux tubes, uncovering the temporal variation of the attractor basins. Representing the activity using the spike interval sequence, we find that the attractor boundary is formed by pre-images of destabilizing events realized when an input and output spike collide. The properties of these collision events allow us to derive the rate of the mutual divergence of neighboring tubes. We develop a disorder-averaging scheme for trajectory ensembles and apply it to the boundary trajectories to obtain the distribution of flux tube diameters. Assembling these results, we provide a construction of the phase space organization based on a dynamics-induced partition refinement seeded by the disordered connectivity. Finally, we discuss the results, their generality, and applications of this ensemble averaging method. The proposed approach to revealing attractor structure from spiking activity informs how coupling strength, connectivity, single neuron dynamics and population activity control a circuit's sensitivity to perturbations. This is knowledge that can guide the burgeoning experimental approaches, such as bidirectional neural implants, that investigate neural computation by perturbing neural dynamics.

MODEL DEFINITION

N neurons are connected by an Erdős-Rényi graph with adjacency matrix $A = (A_{mn})$. $A_{mn} = 1$ denotes a connection from neuron n to m, realized with probability, p. The neurons' membrane potentials, $V_n \in (-\infty, V_{\rm thr}]$, are governed by LIF dynamics,

$$\tau \dot{V}_n(t) = -V_n(t) + RI_n\,(t)\ , \tag{1}$$

for $n \in \{1, \ldots, N\}$ ($\dot{x}$ is the time derivative of x). Here, $\tau = (RC)^{-1}$ is the membrane time constant for a membrane with capacitance, C, and resistance, R. $I_n\,(t)$ is the synaptic current received by neuron n; when V_n reaches a threshold, $V_{\rm thr}$, neuron n 'spikes', and V_n is reset to $V_{\rm res}$. Without loss of generality and for convenience, the voltage has been non-dimensionalized so that the reset is $V_{\rm res} = -1$ and $V_{\rm thr} = 0$, which zeros the offset of spike rate as a function of external current. At the spike time, t_s, the spiking neuron, n_s, delivers a current pulse of strength J to its $\mathcal{O}(K := pN)$ postsynaptic neurons, $\{m|A_{mn_s} = 1\}$, with spike index, s. The total synaptic current is

$$I_n\,(t) = I_{\rm ext} + \tau J \sum_s A_{nn_s}\delta(t - t_s)\ , \tag{2}$$

where $I_{\rm ext} > 0$ is a constant external current and $J < 0$ is the recurrent coupling strength. An $\mathcal{O}\left(1/\sqrt{K}\right)$-scaling of J maintains finite current fluctuations at large K and implies that the external drive is balanced by the recurrent input. As a consequence, firing in this network is robustly asynchronous and irregular [38–41]. Setting $I_{\rm ext} = \sqrt{K}I_0$, with $I_0 > 0$, and $J = -J_0/\sqrt{K}$ with $J_0 > 0$, the corresponding stationary mean-field equation for the population-averaged firing rate, $\bar{\nu}$, reflects a balance of the external drive and recurrent inhibition [14]

$$\bar{\nu}\tau = \frac{I_0}{J_0} + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)\ . \tag{3}$$

It is convenient to map the voltage dynamics to a

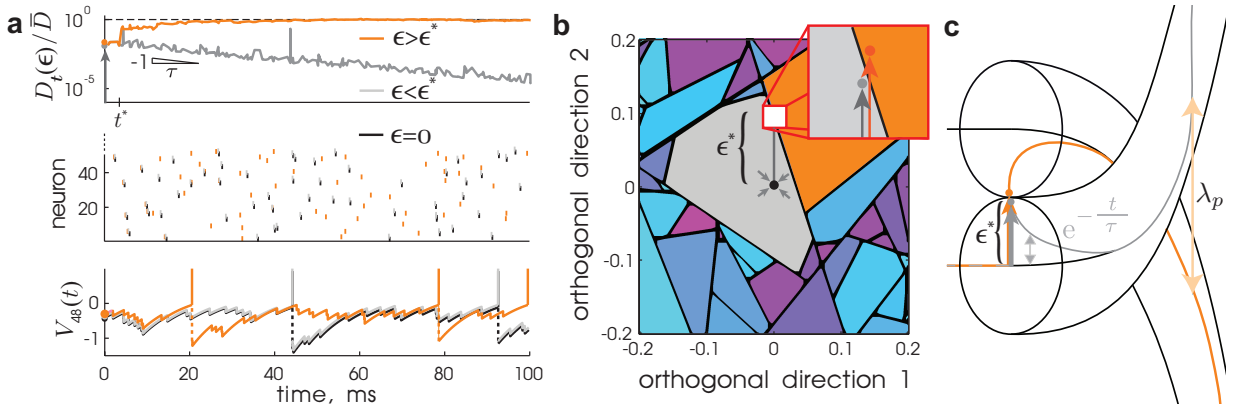

Figure 1. *Finite-size perturbation instability and phase space partitioning in spiking networks.* The three panels display the same two slightly subcritical and supercritical perturbations of strength $\epsilon^* \pm \delta$, $\delta \gtrsim 0$, respectively, applied once at $t = 0$ and in a random direction away from an attracting trajectory. **(a)** Temporal responses of the system. *Top*: The corresponding distance time series, $D_t(\epsilon)$, between the perturbed and unperturbed trajectories (gray: sub-critical, orange: super-critical). Arrows in all three panels indicate the respective perturbation. The divergence of $D_t(\epsilon^* + \delta)$ begins at $t^* \approx 3$ ms, and saturates at the average distance between randomly chosen trajectories, $\bar{D}$ (dashed line) [14], while $D_t(\epsilon^* - \delta)$ only decays exponentially. *Middle*: The spike times as vertical ticks of 50 randomly labeled neurons from the network. The unperturbed sequence ($\epsilon = 0$) is shown in black. *Bottom*: The subthreshold voltage time course of an example neuron. The spike sequence and membrane potentials of the sub and supercritical trajectories decorrelate after t^*. **(b)** A 2D cross-section $(\delta\phi_1, \delta\phi_2)$ of the pseudo phase representation of the phase space, orthogonal to and centered on the unperturbed trajectory from (a) at $t = 0$ (see also [14]). The black dot at the origin indicates the latter, whose attractor basin is colored gray. The other colors distinguish basins in the local neighborhood. The two perturbed trajectories from (a) were initiated from $(\delta\phi_1, \delta\phi_2) = (0, \epsilon^* \pm \delta)$, respectively (shown as gray and orange dots, respectively, in the inset, in (a,Top and Bottom), and in (c)). **(c)** Schematic provided by Ref. [14] of phase space caricature of two neighboring flux tubes with subcritical perturbations decaying on the order of the membrane time constant, τ, and typical basin diameter, ϵ^*. The pseudo Lyapunov exponent, λ_p, is the rate at which neighboring tubes separate from each other (parameters: $N = 200$, $K = 50$, $\bar{\nu} = 10$ Hz, $\tau = 10$ ms, $J_0 = 1$).

pseudo phase representation [14, 15] with

$$\phi_n(t) = \frac{\tau}{T_{\text{free}}} \ln\left[\frac{I_{\text{ext}} - V_{\text{res}}}{I_{\text{ext}} - V_n(t)}\right] , \tag{4}$$

where T_{free} is the oscillation period of a neuron driven only by I_{ext}. $\phi_n(t)$ evolves linearly in time,

$$\dot{\phi}_n(t) = T_{\text{free}}^{-1} , \tag{5}$$

between spike events, *i.e.* $t \notin \{t_s\}$, and undergoes shifts given by the phase response curve, $Z(\phi)$, across input spike times. Its argument ϕ is the state at spike reception. In the large-K limit, T_{free} and $Z(\phi)$ simplify to

$$T_{\text{free}} \approx \frac{\tau}{I_{\text{ext}}} = (\sqrt{K} J_0 \bar{\nu})^{-1} , \tag{6}$$

$$Z(\phi) \approx -d\phi + const.$$

$$\text{with } d := \frac{|J|}{I_{\text{ext}}} = (K\bar{\nu}\tau)^{-1} , \tag{7}$$

respectively. Event-based simulations of this model are described in Appendix A. See Ref. [14] for further details. They employ the phase representation for its computational efficiency and for viewing a cross section of the phase space. We will rely on its tractability to describe the phase space geometry.

RESULTS

A. Phase-space contraction and partitioning

The circuit models exposed above have proved useful for understanding many aspects of the inhibition-dominated regime of cortical network activity. Despite having no recurrent excitatory connections, they serve as a limiting class of models for the fast action potential onset and pulse-coupling regime that also exhibits the mean activity statistics characteristic of the asynchronous, irregular activity of canonical excitatory-inhibitory circuits. It was in these models that a locally stable dynamics was first observed [15, 16].

The character of the resulting phase space partitioning is complicated here by the nonlinear time evolution of the network state voltages. In the phase representation $\vec{\phi}(t)$ (Eq. (4)) by contrast, the state evolves linearly in the unit-hypercube and parallel to its main diagonal. States evolved across a face of the cube are mapped to a different location in the opposite face. We exploit this representation to define measures of phase space contraction and partitioning, two important features of the dynamics that contribute to the evolution of nearby trajectories.

With vanishing coupling strength between neurons, $J = 0$, the dynamics reduces to $\dot{\vec{\phi}}(t) = const.$ and so

preserves phase space volume. For the recurrent dynamics emerging at finite J, however, the phase space volume is contracted in the $\mathcal{O}(K)$ dimensional subspace spanned by the post-synaptic neurons at each spike as a result of the derivative of the phase response curve being negative, $\frac{\mathrm{d}}{\mathrm{d}t}Z(\phi) \approx -d$ (Eq. (7)). Thus, trajectories from a small ball of initial conditions observed at the same future spike form a ball of states that contracts by a factor $1-d$ along each of these K dimensions. The volume thus contracts by $(1-d)^K \approx e^{\lambda_K}$ per spike, for $K \gg 1$, with exponential rate,

$$\lambda_K \approx -Kd < 0 \,. \tag{8}$$

λ_K captures how the model ingredients involved in the inhibitory interactions contribute to this dissipative dynamics (mean Lyapunov exponent, $\lambda_{\text{mean}} < 0$ [14]). The latter appears to be the dominant stabilizing contribution, and strong enough to stabilize the dynamics (maximum Lyapunov exponent,$\lambda_{\max} < 0$ [16]).

Larger phase space volumes, however, are not uniformly contracted but were previously found in simulations [13, 14] to be torn apart, with the pieces individually contracted but mutually dispersed across the entire traversed phase space volume. The elementary phenomenon in a single direction is illustrated in Fig. 1. To study the sharp onset of this tearing, we capture its discrete nature by introducing the *critical perturbation strength*, ϵ^*: the precise extent out from a given state $\vec{\phi}_0$ on the attracting trajectory, $\vec{\phi}_t$, and in a given orthogonal perturbation direction, $\vec{\xi}$, within which trajectories contract over time,

$$\epsilon^*(\vec{\phi}_0, \vec{\xi}) := \sup\left\{\epsilon \,\middle|\, \lim_{t\to\infty} D_t(\epsilon) = 0\right\} \,. \tag{9}$$

Here, $D_t(\epsilon)$ is the distance between perturbed and unperturbed trajectories using any conventional metric, since this definition only concerns the finiteness of the limiting behavior (we use the 1-norm to allow interpretation of the values relative to the distance between reset and threshold; see Appendix B for details). D_t initially decays exponentially. For $\epsilon = \epsilon^* - \delta$ ($\delta > 0$), this decay characterizes the long-time behavior. For $\epsilon = \epsilon^* + \delta$, in contrast, there exists a *divergence event time,* $t^* = t^*(\vec{\phi}_0, \vec{\xi}) > 0$, at which a sustained divergence in D_t begins (see Fig. 1a). In later sections we will show that this holds for $\delta \to 0$, and that this discreteness of ϵ^* arises from a discrete destabilizing location in the phase space traversed by the trajectory at t^*.

Since we will build on the picture established by [14] (Fig. 1c), we present and comment on it here. Since trajectories in the phase representation only change their relative positions at the boundary of the phase space, the geometry is more clearly reflected in the Poincare section obtained by projecting the hypercube phase space into the $N-1$-dimensional hyperplane orthogonal to its main diagonal (see Ref. [14]). The system's state between spikes becomes a point in the hyperplane, and the sequence of such points indexed by spikes corresponds to the trajectory. A small portion of a 2D projection of this hyperplane around $\vec{\phi}_0$ (Fig. 1b) reveals that the locations of these critical perturbations form lines that partition this plane into polygon-shaped basin boundaries formed by their intersections. The putative N-dimensional volumes serving as attractor basins were termed flux tubes (Fig. 1c) [14]. The smoothness of the caricature in Fig. 1c is misleading in two ways, however. First, the divergence of supercritically perturbed trajectories only begins at t^*. Initially, these attract alongside the subcritically perturbed trajectories. Second, as we will see, the flux tube boundaries are not uniformly smooth. We expect they are formed by sequences of random, but temporally correlated $(N-1)$-dimensional polytopes, each enclosing a state from the state sequence trajectory in the hyperplane. Before developing a theory for this phase space organization, we analyze two main features of its geometry: the punctuated exponential decay of a tube's cross-sectional volume and the exponential separation of neighboring tubes.

B. Punctuated tube geometry

By following a simulated trajectory, we find that the cross-sectional volume enclosed by the local flux tube exhibits exponential decay. This decay, expected from the typical phase space volume contraction (see Eq. (8)), must be opposed by some counteracting element to be consistent with the numerical observation in [14] of a finite average size. Indeed, our simulation shows that the volume decay is punctuated by events at which the volume blows up. Figure 2a displays the spiking activity produced by a typical trajectory, $\vec{\phi}_t$. The temporal evolution of the neighborhood around $\vec{\phi}_t$ in the hyperplane is more clearly represented in a *unfolded representation* in which copies of the space are aligned such that the trajectory passes through them continuously. For visualization, we show only a fixed, 2D projection of the hyperplane around $\vec{\phi}_t$ (Fig. 2b and *SI Movie*; see Appendix D for construction details). The boundary of the attractor basin surrounding $\vec{\phi}_t$ in this 2D projection consists of lines which remain fixed between spike times. Across spike times, new lines appear and existing lines disappear. At irregular intervals breaking up time windows of exponential contraction, large abrupt blowup events take the boundary away from the center trajectory. The area enclosed by the boundary increases sharply there as a result (Fig. 2c). It is important to note that these blowup events do not mean that the evolving phase space volume from an ensemble of nearby trajectories would expand. Such volumes only contract and converge to the same asymptotic trajectory. The basin of attraction it-

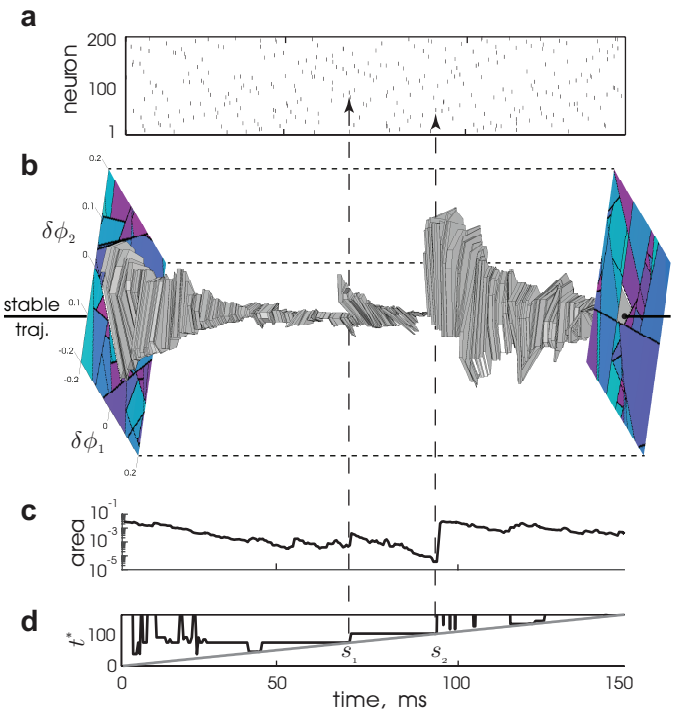

Figure 2. *The basin boundary contracts towards and can blowup away from the stable center trajectory.* **(a)** Spike times from all neurons of the simulated trajectory, $\vec{\phi}_t$, in a 150 ms window. **(b)** 2+1D unfolded phase space volume, $(\delta\phi_1, \delta\phi_2, t)$, centered around $\vec{\phi}_t$ located at $(0, 0, t)$ (black line) and extended in two fixed, random directions, $\delta\vec{\phi}_1$ and $\delta\vec{\phi}_2$. The center tube is filled gray in this volume, and the two cross-sections, $(\delta\phi_1, \delta\phi_2, 0)$ and $(\delta\phi_1, \delta\phi_2, 150)$, are shown. **(c)** Cross-sectional area of the center tube from (b) versus time. The area decays exponentially but can undergo abrupt expansions at blow-up times, e.g. at spikes s_1 and s_2 (note the logarithmic scale on the ordinate). **(d)** The absolute time of the next divergence event, t^* (see Fig. 1a, top), versus time, for perturbations along $\delta\vec{\phi}_1$. Note the step increase coincident with the blowup events seen in (b,c) (vertical, dashed lines). (Same parameters as Fig. 1.)

self, however, does not exclusively contract with time, but with these blowup events maintains a typical size on average.

The conspicuous blowup events typically coincide with a divergence event time, t^* (Fig. 1a). Two such coincidences are visible in Fig. 2c,d. This suggests the hypothesis that there exist destabilizing locations of the phase space that underlie both blowup and divergence events. How would such locations give rise to the observed time variation of the boundary? Due to the exponential expansion of the backward dynamics, the set of their pre-images naturally trace out the observed exponential shape. With sufficient backward iterations, however, another destabilizing event closer to the trajectory is passed and becomes the event determining the boundary. Thus, we conclude that a local basin's extent in a direction, at a given time, and out from the attracting trajectory is determined by a pre-image of a divergence event at a location in the phase space nearby the trajectory at a future time.

C. Decorrelating spike collision events

While our above analysis highlighting the existence of destabilizing events does not rely on what causes them, that knowledge is nevertheless important to further understand the origin and generality of flux tubes. We thus analyzed a set of divergence events from simulations to reveal that the collision of a pair of input and output spikes was responsible (see Appendix E). This occurs when synaptic input is received around when the voltage is near the threshold for spiking. Thus, the pair of spikes involved in a collision event are generated by connected pairs of neurons. As a phase space location, a collision event is then the $N-2$-dimensional 'hyperedge' subspace of the hypercube spanned by the remaining $N-2$ neurons and passing through $\vec{\phi} = \vec{1}$. Moreover, we found that a perturbation-induced collision of a input-output spike pair generated an abrupt spike time shift in one or both of these spike times depending on the motif by which the two neurons connect. The type of voltage dynamics and coupling interaction conspire to produce this shift, as described in Appendix F and shown in Fig. 3 for the *backward-connected* pair motif $n_{s^*} \leftarrow n_{s'}$, where s^*, the *divergence event index*, is the spike index of the earlier of the pair (note that $t^* \equiv t_{s^*}$), and $s' > s^*$ here labels the index of the later spike in the pair. The two other motifs (forward-connected and symmetrical) are discussed in Appendix F, where we also demonstrate this abrupt shift in two less idealized neuron models, each exhibiting a smoothness in one of the two limits of fast action potential onset and fast coupling, respectively, that characterize (non-smooth) pulse-coupled LIF networks.

The spike time shift resulting from the collision is large enough that with saturating probability, the shifted spike collides with a spike from a neuron in its pre or post synaptic subpopulation, depending on the motif (see Appendix H for the analytical result). An approximately exponential cascade of collision events follows whose speed then depends on the average rate of spikes in these subpopulations of on average K neurons,

$$\omega_K = K\bar{\nu} \equiv p/\overline{\Delta t}\,, \tag{10}$$

where $\overline{\Delta t} = (N\bar{\nu})^{-1}$ is the average distance between successive spikes.

Thus, the total collision rate is ω_K multiplied by the number of source neurons. For most of the cascade, collisions involve a previously unaffected neuron, so the number of source neurons roughly increments with each collision. With collision event times, $\{t_m\}$ (reference time t^*),

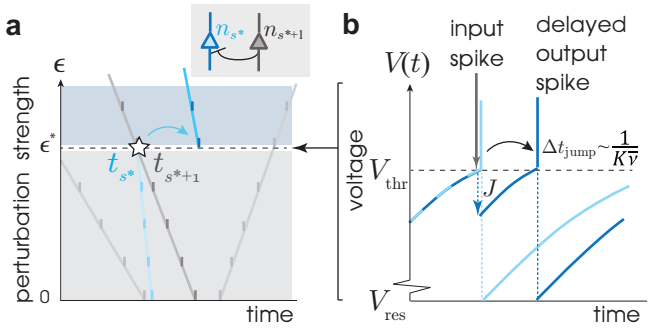

Figure 3. *The collision of a input-output spike pair causes an abrupt change in spike time.* **(a)** A schematic illustration of the collision event (☆) for the backward-connected pair motif (shown in inset). For this motif, the interval vanishes as $\epsilon \to \epsilon^*$ from below. The linearly varying locations of spike times as a function of perturbation strength, ϵ, are plotted in the ϵ-time plane. The spike times shift continuously for $\epsilon < \epsilon^*$. The next input spike time, $t_{s^*+1}(\epsilon^* - \delta)$, is advanced over the output spike, $t_{s^*}(\epsilon^* - \delta)$. A discontinuous jump of size $\Delta t_{\rm jump}$ occurs in the spike time of the post-synaptic neuron, n_{s^*} (light to dark blue) from $t_{s^*}(\epsilon^* - \delta)$ to $t_{s'>s^*}(\epsilon^* + \delta)$, $\delta \gtrsim 0$. **(b)** Schematic illustration of the voltage of the n_{s^*} neuron versus time for $\epsilon^* \pm \delta$. The inhibitory kick of size $J = -J_0/\sqrt{K}$ (not shown to scale) delays the spike time by an amount $\Delta t_{\rm jump} \sim (K\bar{\nu})^{-1}$.

the inverse total collision rate gives an estimate for the interval, $t_m - t_{m-1}$, between successive collision events. Using the approximation $\frac{1}{m} \sim \log(1-\frac{1}{m})^{-1}$ valid for $m \gg 1$, we can then write $t_m - t_{m-1} \sim (\omega_K)^{-1} \log(1-\frac{1}{m})^{-1}$, which can be rearranged as $m/(m-1) = e^{\omega_K(t_m - t_{m-1})}$. Over many realizations of the cascade, the average number of collisions, and thus the distance grows exponentially with rate, ω_K, providing the origin of the numerical scaling result for the pseudoLyapunov exponent, $\lambda_p = K\bar{\nu}$ [14], and the rate at which adjacent flux tubes diverge from one another.

Statistical theory of flux tube diameter

We capture the geometry of a flux tube by introducing the *flux tube indicator function*, $\mathbb{1}_{\rm FT}(\epsilon) = \Theta\left(\epsilon^*(\vec{\phi}_0, \vec{\xi}) - \epsilon\right)$, evaluated at a network state, $\vec{\phi}_0$, on the attracting trajectory inside a tube, and for a perturbation direction, $\vec{\xi}$, orthogonal to it. Using the Heaviside function, $\Theta(x)$, $\mathbb{1}_{\rm FT}(\epsilon) = 1$ for perturbation strengths remaining in the tube ($\epsilon < \epsilon^*$), and is 0 otherwise (see Fig.5a; Eq (9)). The average of $\mathbb{1}_{\rm FT}(\epsilon)$ over $\vec{\phi}_0$ and $\vec{\xi}$,

$$\hat{S}(\epsilon) = \langle \mathbb{1}_{\rm FT}(\epsilon) \rangle_{\rho(\vec{\phi}_0, \vec{\xi})} \ , \tag{11}$$

is the *survival function*: the probability that an ϵ-sized perturbation does not lead to a divergence event later in the perturbed trajectory. Formally, $\hat{S}(\epsilon) := 1 - \int_0^\epsilon \rho(\epsilon^*)\, d\epsilon^*$, with $\rho(\epsilon^*)$ the transformed density over ϵ^*. $\hat{S}(0) = 1$ and decays to 0 as $\epsilon \to \infty$. The scale of this decay defines the typical flux tube size. Calculating $\hat{S}(\epsilon)$ requires two steps: firstly, establishing a tractable representation of $\epsilon^*(\vec{\phi}_0, \vec{\xi})$ and secondly, performing the average in Eq. (11). Both of these in general pose intricate problems. However, as we will see next, they substantially simplify when generic properties of the asynchronous, irregular activity regime are taken into account.

Since the spike collision event underlying ϵ^* for each $(\vec{\phi}_0, \vec{\xi})$ pair can be identified through a vanishing spike interval, we represent trajectories using the perturbed spike interval sequence. The perturbation-induced spike time deviations, $\delta t_s(\epsilon) := t_s(\epsilon) - t_s(0)$, $s = 1, 2, \ldots$, provide this sequence,

$$\Delta t_s(\epsilon) = t_s(\epsilon) - t_{s-1}(\epsilon) = \Delta t_s(0) + \delta t_s(\epsilon) - \delta t_{s-1}(\epsilon) \, , \tag{12}$$

here with $s \geq 2$. In a linear approximation valid in our setting where $\bar{\epsilon}^* \ll 1$,

$$\delta t_s(\epsilon) \approx -\frac{T_{\rm free}}{\sqrt{N}} a_s \epsilon \, , \tag{13}$$

where a_s is a recursively defined, dimensionless susceptibility (see Appendix I).

$$\begin{aligned} a_s :=& \xi_{n_s} \prod_{j=1}^{s-1} \left(1 + d_{\phi_s^j}\right)^{A_{n_s n_j}} \\ &+ \sum_{j=1}^{s-1} A_{n_s n_j} d_{\phi_s^j} a_j \left(\prod_{k=j+1}^{s-1} \left(1 + d_{\phi_s^k}\right)^{A_{n_s n_k}} \right) \, , \end{aligned} \tag{14}$$

depending on the adjacency matrix, $A = (A_{mn})$, the perturbation direction $\vec{\xi}$, and derivatives of the phase response curve evaluated at the previous states when input spikes were received, $d_{\phi_s^j} := Z'(\phi_{n_s}(t_j))$. $T_{\rm free}$ converts phase to time (cf. Eq. (5)). With $\mathcal{O}(1)$-mean, random elements, $|\vec{\xi}| \propto \mathcal{O}(\sqrt{N})$, so $-\frac{T_{\rm free}}{\sqrt{N}}$ simply converts the units into an $\mathcal{O}(1)$ spike time deviation. Note from Eqs. 12 and 13 that $\Delta t_s(\epsilon)$ can have a zero, *i.e.* a spike time collision only when $\Delta a_s = a_s - a_{s-1} > 0$.

We now focus on collisions of backward-connected input-output spike pairs. They obey a simple, implicit definition of $\epsilon^*(\vec{\phi}_0, \vec{\xi})$, expressed using the perturbed spike intervals and connectivity alone: the smallest ϵ for which $\Delta t_s(\epsilon - \delta) \to 0$ as $\delta \to 0$ for any s satisfying $A_{n_{s-1} n_s} = 1$. Using this definition we can write,

$$\mathbb{1}_{\rm FT}(\epsilon) = \prod_{s=2}^{\infty} \Theta\left(\Delta t_s(\epsilon)\right)^{A_{n_{s-1} n_s}} \ , \tag{15}$$

with $\Delta t_s(\epsilon) = \Delta t_s - \frac{T_{\rm free}}{\sqrt{N}} \Delta a_s \epsilon$. $\mathbb{1}_{\rm FT}$ and its average, $\hat{S}(\epsilon)$ (Eq. (11)) will depend on the adjacency matrix,

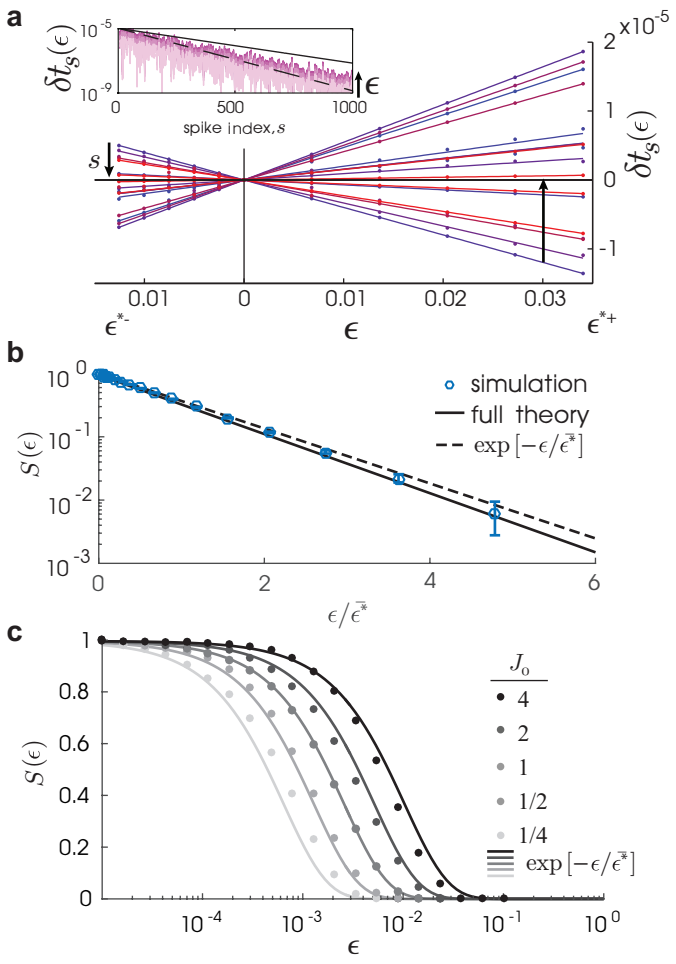

Figure 4. *The survival probability to remain in a flux tube.* **(a)** Spike-time deviations, $\delta t_s(\epsilon)$ (dots), as a function of perturbation strength up to the positive and negative critical strength, ϵ^{*-} and ϵ^{*+}, respectively, for $s = 1, \ldots, 15$ (colors) with their linear approximation (lines) given by Eq. (13). Inset: $\delta t_s(\epsilon)$ as a function of s (shown for $\epsilon = 0.2\epsilon^{*\pm}, 0.4\epsilon^{*\pm}, 0.6\epsilon^{*\pm}, 0.8\epsilon^{*\pm}$) decays exponentially at a rate near the maximum and mean Lyapunov exponent, $\lambda_{\max}$ (black line) and λ_{mean} (black-dashed line) respectively [14]. **(b)** The survival probability function $S(\epsilon)$ from simulations (dots, Eq. (11); bars are standard error), theory (line, equations (19),(20), and the simplified theory at large K, $\exp\left[-\epsilon/\overline{\epsilon^*}\right]$ (dotted line, Eq. (22)), where $\overline{\epsilon^*} = \left(\sqrt{KN}\bar{\nu}\tau/J_0\right)^{-1}$. **(c)** $S(\epsilon)$ from simulations (dots) and $\exp\left[-\epsilon/\overline{\epsilon^*}\right]$ (lines) for $J_0 = 2^n, n = -2, -1, 0, 1, 2$. (Same parameters as Fig. 1 except $N = 10^4$, $K = 10^3$.)

$A = (A_{mn})$, of the network realization. Conveniently, removing this dependence by averaging over the ensemble of graphs, $P(A)$, simplifies the calculation of the survival function,

$$S(\epsilon) = \left\langle \hat{S}(\epsilon) \right\rangle_{P(A)} . \tag{16}$$

Evaluating the right-hand side of Eq. (16) using the linearized perturbed spike intervals requires knowledge of the joint probability density of the variables on which these intervals depend,

$$\rho\left(\{\Delta a_s\}, \{\Delta t_s\}, \left\{A_{n_{s+1}n_s}\right\}, M, \vec{\phi}_0 \middle| \vec{\xi}, A\right) \rho(\vec{\xi})\, P(A) , \tag{17}$$

where we have chosen the perturbation direction, $\vec{\xi}$, to be statistically independent of the initial perturbed state, $\vec{\phi}_0$. Taken over a time window of size, T, we hereafter refer to this density as ρ_T. Here, the unperturbed spike pattern is represented by two random variables: M, the number of spikes in the time interval $[0, T]$ after the perturbation, and $\{\Delta t_s\}$, the set of all $M-1$ inter-spike intervals in this window.

We now exploit the properties of the asynchronous, irregular phase. It is well understood that in the large-system limit for a sparse graph, $1 \ll K \ll N$, the currents driving individual neurons in the network converge to independent, stationary Gaussian random functions [17]. For low average firing rates, this implies that the pattern of network spikes $(M, \{\Delta t_s\})$ resembles a Poisson process with weak serial correlations and exponential spike interval distribution [18]. These weak serial correlations are absent in $\mathbb{1}_{\text{FT}}$ at short range by the sparsity $(p \ll 1)$ of the surviving $(A_{n_{s+1}n_s} = 1)$ factors and are further suppressed at longer range by the irregular activity and the fact that $s \neq s^*$ indexed variables contribute to $\mathbb{1}_{\text{FT}}$ only insofar as they determine s^* via an extremum condition, not via their actual values. Thus, we neglect serial index correlations. Moreover, the linearization of the phase response curve for the weak coupling in this limit implies that its derivative, and thus the susceptibilities a_s, are state-independent (see Eq. (7)). Finally, we neglect the weak dependence between the distribution of network spike patterns and $A = (A_{mn})$.

Using the above assumptions (see Appendix J for details), we have the factorized density

$$\rho_T \approx P(A_{mn})\, P_T(M) \prod_{s=2}^{M} \rho(\Delta t)\, 2\Theta(\Delta a_s) \rho(\Delta a_s) , \tag{18}$$

with distribution of an adjacency matrix element, $P(A_{mn} = 1) = p$, $P(A_{mn} = 0) = 1 - p$, count distribution of spikes in the observation window, $P_T(M)$, and exponential distribution of single inter-spike intervals, $\rho(\Delta t)$ with scale parameter $\overline{\Delta t}$ (see Fig.J.1). All dependencies on the distribution of perturbation direction are now mediated by the susceptibilities, $\{\Delta a_s\}$. For any isotropic $\rho(\vec{\xi})$ having finite-variance, we find $\rho(\Delta a_s)$ has zero mean and standard deviation proportional to $\exp\left[\frac{\lambda_K}{N} s\right]$ with the average contraction rate per neuron, $\frac{\lambda_K}{N} = -\frac{Kd}{N} = -pd$, due to the inhibition (see Eq. (8); Appendix J). The factor $2\Theta(\Delta a_s)$ places support only at positive values of Δa_s as required.

As ρ_T factorizes, so does $S(\epsilon)$,

$$S(\epsilon) = \lim_{T\to\infty} \left\langle \prod_{s=1}^{M} S_s(\epsilon) \right\rangle_{P_T(M)} = \prod_{s=1}^{\infty} S_s(\epsilon) \ , \qquad (19)$$

where $S_s(\epsilon)$ is the probability that a perturbation of strength ϵ does not lead to a collision event involving the s^{th} spike. With the above simplifications,

$$S_s(\epsilon) = \left\langle \Theta\left(\Delta t - \frac{T_{\text{free}}}{\sqrt{N}} \Delta a_s \epsilon\right)^{A_{mn}} \right\rangle_{\rho(\Delta t)\rho(\Delta a_s)P(A_{mn})} . \qquad (20)$$

Evaluating Eq. (20) (see Appendix J for details), we find

$$S(\epsilon) \approx \prod_{s=1}^{\infty} \left(1 - \frac{T_{\text{free}}}{\sqrt{N}} \omega_K e^{\frac{\lambda_K}{N} s} \epsilon\right) , \qquad (21)$$

where we have identified the rate of spikes from connected sub-populations, ω_K (Eq. (10)). Employing the logarithm and $\frac{T_{\text{free}}}{\sqrt{N}}\omega_K \epsilon \propto \sqrt{p}\epsilon \ll 1$,

$$S(\epsilon) \approx \exp\left[-\frac{\epsilon}{\overline{\epsilon^*}}\right] \qquad (22)$$

with

$$\overline{\epsilon^*} = \sqrt{N} \frac{\left|\frac{\lambda_K}{N}\right|}{\omega_K T_{\text{free}}} = \sqrt{N} \frac{Kd/N}{K\bar{\nu} T_{\text{free}}} = \sqrt{N} \frac{|J|/N}{\bar{\nu}\tau}$$
$$\overline{\epsilon^*} = \frac{J_0}{\sqrt{KN}\bar{\nu}\tau} , \qquad (23)$$

where K and I_{ext} (Eq. (2)) cancel, and we have used Eqs. (6) and (7). With the survival probability in hand, the density, $\rho(\epsilon^*)$, is simply obtained as the negative of its derivative. Equation (22) shows for $1 \ll K \ll N$ that the basin diameter, ϵ^*, is exponentially distributed and so completely determined by its characteristic scale, $\overline{\epsilon^*}$ (Eq. (23)). $\overline{\epsilon^*}$ is smaller for larger network size, higher average in-degree, higher population activity, and larger membrane time constant, τ. Diameters tend to be larger, however, for stronger synaptic coupling strength, J_0. This previously unknown dependence of $\overline{\epsilon^*}$ is crucial to its scaling with the stabilizing rate, λ_K and its interpretation as the ratio of the stabilizing and destabilizing rates. In Fig. 4b, we show quantitative agreement in simulations between the definition of $\hat{S}(\epsilon)$ (Eq. (11) using the definition of ϵ^*, Eq. (9)) and its approximate microstate parametrization (Eqs. (19), (20)). These results also confirm the exponential form of our reduced expression (Eqs. (22), (23)) and a scaling dependence on J_0 (Fig. 4c). The latter holds while the system is in the asynchronous and irregular activity regime of $J \propto \mathcal{O}\left(1/\sqrt{K}\right)$. The other scalings agree with previous numerical simulations [14].

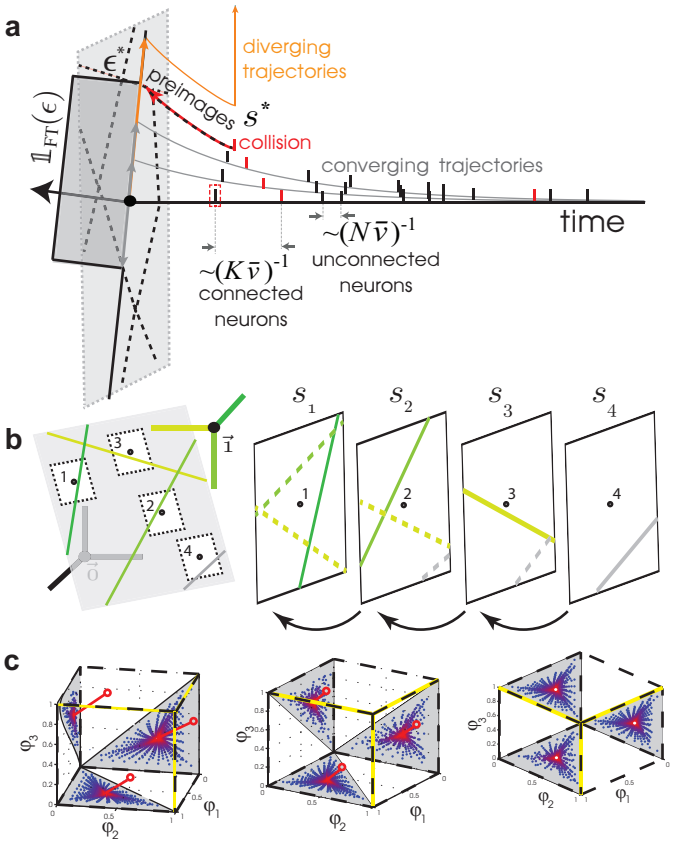

Figure 5. *Flux tube boundaries are the pre-images of future input-output spike collisions.* **(a)** A unfolded phase space representation of a input-output spike collision. Spikes (ticks) occur at a rate $N\bar{\nu}$ in the unperturbed trajectory (black line). For an example spike (red-bordered) from some neuron, the spikes from connected neurons (red) occur at rate $K\bar{\nu}$. Small perturbations (gray arrows) lead to trajectories (gray) with spike time deviations that decay (tick alignment). A larger perturbation (orange arrow) generates an input-output spike collision (at s^*) in the subsequent trajectory. The indicator function, $\mathbb{1}_{\text{FT}}(\epsilon)$, has support (dark gray) only over the local tube. **(b)** Constructing the local flux-tube partition in the non-unfolded phase space. *Left*: Input-output spikes are represented by unstable edges (thick green lines) of the unit hypercube having $\vec{1} = (1, \ldots, 1)$ (black dot) as an endpoint. An intrinsic partition (thin green lines) is generated by projecting these edges onto the hyper-plane orthogonal to $\vec{1}$ (light gray). A given trajectory (numbered sequence of small dots) and its local neighborhood (within black dashed lines) is shown. *Right*: The flux tube partition for this trajectory at a given spike (here s_1) is obtained from back-iterating the intrinsic partition from all future spikes (dashed lines). **(c)** A fully-connected 3-neuron network phase space viewed from rotated perspectives (from left to right) so that the main diagonal aligns perpendicular to the page, showing that a 2D projection captures the dynamics. All states on the reset manifold are attracted in time (blue to red) to a unique trajectory (red line), emitting spikes on the threshold manifold (the red outlined dots). The unstable edges (yellow) and their preimages (black-dashed) form the basin boundaries.

Global geometry of phase space partitioning

Using our results, we build a summary the phase space organization of these spiking circuits as we have revealed it (Fig. 5). Figure 5a schematizes the phase space geometry local to a basin in the unfolded Poincare sections where an attracting trajectory is located at the center, as in Fig. 2a. Sub-critical perturbations push the perturbed state within the tube and vanish over time, while supercritical perturbations lead to a divergence event at some future spike time. The preimages of the divergence event in previous sections determine the flux tube boundary back to the perturbation time. The flux tube indicator function, extending out in one perturbation direction, has support only on the local tube. Averaging this function over perturbation directions and states gave the survival probability of remaining in a flux tube after a perturbation.

This trajectory-centered view is useful for quantifying local deviations away from stable trajectories. It is nevertheless artificial because the location of the stable trajectory is not established *a priori*, but is determined by the combination of the spatial layout of spike collisions (and thier pre-images) in the phase space and the projected effects of the inhibition-induced volume contraction of the postsynaptic subspace. Our results suggest a picture of the geometry based on the global partitioning of the phase space by these events (Fig. 5b). Here, input-output spike collisions are represented by the subset of hyperedges of the N-dimensional unit hypercube of phases where the corresponding voltages of two connected neurons both approach threshold. A perturbed trajectory will diverge once one of these unstable edges is crossed. The projections of these unstable edges into the Poincare section of the dynamics generates a partition (Fig. 5b(left)). The flux tube partition emerges as the refinement of this partition obtained by iterating it backwards in time using the inverse of the Poincare map. Different parts of this refined partition are sampled by the trajectory's neighborhood as it evolves. The polytope basin boundaries thus arise as backward iterates of the unstable edges lying nearby the trajectory (Fig. 5b(right)). Unstable edges at sufficiently distant future spikes (the gray edge at s_4) will no longer refine the partition in the local neighborhood (at s_1), since the expansive backwards dynamics maps the projected edges outside the neighborhood. The refinement thus converges to a unique global partition of the phase space.

For concreteness, in Fig. 5c we use rotated perspectives to show how the projection captures the full phase space dynamics in the flux tube partition of a fully-connected, 3-neuron network. There are two stable spike index sequences for this network, $\ldots 123 \ldots$ and $\ldots 213 \ldots$. Permutation of any adjacent spikes thus changes the encoding symbol, i.e. tube, in which the trajectory resides.

More generally, the spike-based code emerging from this partitioning is insensitive to permutations of adjacent spikes from unconnected neurons, and thus exhibits error-corrective properties. This insensitivity necessarily lowers the spike sequence entropy. In a large window T, there are $N\bar{\nu}T$ number of spikes, $K\bar{\nu}T$ of which arrive at a given neuron, which itself emits $\bar{\nu}T$ spikes (ratio: N to K to 1, respectively). The number of all possible distinct spike-index sequences in this window scales roughly as N^N. An upper bound for the entropy then scales superextensively as $N \log N$, similar to the scaling found in an estimate of the entropy of the sequence of network states under this partition [14]. However, the entropy of the spike index sequence is constrained by the dynamics. An index sub-sequence of the network activity can be associated with each neuron by combining its input and output spikes. Enumerating the possibilities for this index subsequence that are considered distinct under the flux tube partition, only the positions of its output spikes relative to its input spikes need to be considered. The number of such positions scales with K and there is such a subsequence for each of the N neurons so that an upper bound on the entropy is $N \log K = N \log pN \ll N \log N$ for $K \ll N$. We conclude that the partitioning of the phase space by the dynamics, and the insensitivity of this partition to permutations in the spike sequence involving unconnected neurons, constrains the capacity of the associated spike-sequence neural code, while providing a robustness to sub-critical perturbations in the encoding.

DISCUSSION

Attractor states and their basins of attraction play a fundamental role in theories of neural computation. Methods from the physics of disordered systems have served these theories by mathematically characterizing the statistics of dynamics and phase space organization of rate networks (*e.g.* Ref. [2] calculates their typical basin diameter in the limit of high gain). Insight is gained by tracking the parameter dependencies in the resulting expressions back to the ingredients used to specify the model. In this contribution, we have used this approach for the treatment of the attractor geometry of flux tubes in the inhibitory LIF networks recently considered by [14]. This dynamics serves as the limiting example of inhibition-dominated spiking circuit models with fast action potential onset and fast synapse kinetics that is relevant to cortical dynamics.

Through massive activity simulations we present the phenomenology of the time-variation of flux tube attractors. The flux tube diameter enclosing an attracting trajectory contracts with the rate of volume contraction per neuron that we derive as, $\lambda_K/N = pd = (N\bar{\nu}\tau)^{-1}$, and is due to the inhibition received across the subspace of

post-synaptic neurons. This contraction is punctuated, however, by blowup events occurring when an initially adjacent flux tube diverges. We provide a formal definition of the attractor boundary with which we identify the boundary trajectories, and track the source of the blowup and divergence instability to a collision of an input and output spike. The exponential shape of the boundary is thus formed by the pre-images of these collisions and the blowup events occur as the network state passes a nearby collision event, and the basin boundary expands out to a pre-image of the next nearest collision event. The rate of spikes in the sub-populations connected to a given neuron, $\omega_K = K\bar{\nu}$, controls the probability of that neuron being involved in a collision event. Once a collision event occurs, it sets off an exponential (with rate, ω_K) cascade of such events afterward that is responsible for the tearing away of some adjacent tube.

Using the nature of these collision events to mathematically identify the spiking trajectories lying on flux tube boundaries, we were able to calculate the size distribution of these basins as a survival probability of a perturbation of a given size and direction remaining with in the local basin. The average basin diameter is controlled by the ratio of these per neuron rates, $(\lambda_K/N)/\omega_K$, as the two dominant opposing contributions to the stability. Both rates depend in the same way on the relative number of interactions, i.e. the dimension of the pre and post synaptic subspace to any neuron, K. It then cancels in the resulting expression, appearing in $\overline{\epsilon^*}$ only implicitly in the scaling of the synaptic coupling, J, and so does not directly control the attractor geometry. The remaining ratio reveals the parameters controlling these two rates,

$$\overline{\epsilon^*} \propto \frac{|J|/N}{\bar{\nu}\tau} \equiv \frac{\text{inhibitory coupling strength: stabilizing}}{\text{rate of spikes: destabilizing}} .$$

Namely, the synaptic coupling strength controls the stabilizing contraction per neuron, while the non-dimensionalized, single neuron spike rate controls the number of candidates for a collision event and thereby the destabilizing contribution to the dynamics. In the final expression, this is multiplied by $\sqrt{N}$ to accounts for the projection of the perturbation onto a single neuron.

From these results, we formed a geometric picture of the structure of the high dimensional phase space using a Poincare map obtained from the phase representation. Trajectories evolve parallel to the main diagonal and hit the sides of the unit hypercube at spikes times. Edges of the cube with $(1, \ldots, 1)$ as a vertex are where the voltages of two neurons neurons reach threshold. Collision events are localized in the space to the subset of these edges associated with connected neurons. Successive preimages of these edges generate successively refined partitions of the phase space. On account of the expansive reverse-time dynamics, this refinement converges to the unique flux tube partition on the space. Together, the elements of this dynamics-generated partition form a finite resolution code of the input signal (for this autonomous dynamical system the input is simply the initial condition). For spike sequence-based codes, sequences are distinct under this code only if they differ in the ordering of spikes in any of the $\mathcal{O}(KN)$ sub-sequences of spiking activity from pairs of connected neurons. Thus, we find that the non-commutability of the spike sequence to adjacent transpositions, previously proposed as applying generically[14], applies only to these subsequences. The code is insensitive to permutations of adjacent spikes from unconnected pairs of neurons and a reduced entropy of the code results.

We emphasize that collision events structure the phase space in this way only when the dynamics is linearly stable. For inhibition-dominated circuits, this phase has been found in the biologically relevant regime of fast but finite action potential onset and fast but finite synapse kinetics [7, 9, 19, 20]. We have applied the theory to explain phase space partitioning in the case of inhibitory, pulse-coupled LIF networks, a limiting model for this regime. Using natural extensions of the LIF to fast but finite action potential onset and fast but finite synapse kinetics, respectively, we find that the abrupt spike time jumps underlying the instability structuring the attractor basins in LIF networks nevertheless persist when using these non-limiting neuron models in this regime. As expected from our theory, divergence events have been observed in both these models [7, 19]. We leave understanding how collision events are involved in the transition out of this regime to future work, but we do show that they persist in two canonical relaxations of the LIF neuron limit.

Applying our approach in a relatively idealized context allowed for a tractable assessment of phase space organization. We have nevertheless neglected additional heterogeneity in many properties. For instance, in contrast to the locally stable regime studied here, mixed networks of excitatory and inhibitory neurons can instead be conventionally chaotic when the excitation is strong enough [21]. It appears this chaos can nevertheless be suppressed in the ubiquitous presence of fluctuating external drive [26–29] and with spatially-structured connectivity [10]. These observations, as well as the stable embedding of spiking patterns into recurrent circuits [30], suggest locally stable dynamics and phase space partitioning are more general features of spiking circuit dynamics than the specific setting studied here.

Our theory of destabilizing collision events treats instabilities by locating in the phase space abrupt changes in subsequent spike times produced as the network state is perturbed off an attracting trajectory. The existence of such instabilities is thus intimately tied to the granularity of spikes, and has no equivalent in rate networks. Despite the instability, chaos is kept at bay in this regime by the dominating effect of the contraction at spike times arising from the inhibitory and pulse-like form of the coupling. With increasingly smooth versions of the coupling or the

hard threshold, this stabilizing contraction is smeared in time, and presumably eventually succumbs to other destabilizing effects, yet to be well-characterized, and the dynamics turns chaotic. Our theory also applies to other, as yet unknown, and even coexisting instabilities involving spike collision events, the existence of which requires further investigation. Our work demonstrates that both unit dynamics and the type of interaction coupling will play a role.

Our approach, in particular the way we have quantified the ensemble of perturbed spiking trajectories, can inform formulations of local stability in less idealized contexts. Of particular interest are extensions where a macroscopic fraction of tubes remain large enough to realize encoding schemes tolerant of intrinsic and stimulus noise. For example, using random dynamical systems theory [29, 31] to incorporate stochastic external drive could provide theoretical control over spiking dynamic variants of rate network-based learning schemes to generate stable, input-specific trajectories [32]. We note that our expression for the survival probability (Eq. 21) takes the form of a q-Pochhammer symbol enumerating all partitions of a set. How exactly this relates to our enumeration of paths through the network, which we needed to compute the spike time deviations due to a previous perturbation to the network state, is left to future work.

Our calculations can be performed for different disordered connectivity ensembles (*e.g.* correlated entries from annealed dilution processes [33] and structured second-order statistics [34]). That the spatial structure of cortical circuits [8, 10] is stabilizing suggests that destabilizing collision events will be relevant for extensions to more realistic connectivities. Different activity regimes (*e.g.* non-Markovian spike interval processes [35]) as well as any hard threshold neuron model with known phase response curve are also amenable to our approach, so long as the averages remain mathematically tractable.

We note that flux tubes are not in a formal sense basins of attraction. The locally stable trajectories they enclose are in fact transients. They are nevertheless made quasistationary by a transient time growing exponentially with network size [9]. Formally, the linear stability of the dynamics precludes a finite-value for the Kolmogorov-Sinai entropy rate. Nevertheless, the partition refinement picture we provide in Fig. 5b is very much analogous to the formal partition refinement used in symbolic dynamics to define trajectories in chaotic systems. The difference is that in our setting the the refinement process converges in a finite number of steps and to partition elements having finite measure (i.e. not points), suggesting that there is transient production of information about a perturbation that persists up to timescales of the order of the divergence event time, t^*. Formally establishing this connection to ergodic theory is an interesting direction for future research.

Recent advances in experimental neuroscience have allowed for probes of the finite-size stability properties of cortical circuit dynamics *in vivo*. For example, simultaneous intra- and extra-cellular recordings in the whisker motion-sensing system of the rat reveal that the addition of a single spike makes a measurable impact on the underlying spiking dynamics of the local cortical area [36]. Indeed, rats can be trained to detect perturbations to single spikes emitted in this area [37]. Toy theories explicitly representing spiking interactions, such as the one presented here, can inform future experimental studies by highlighting the features of spiking neural circuits that contribute to these response properties. This combined theory-experiment approach promises to elucidate a rich substrate for collective computation in terms faithful to the way neurons actually interact.

ACKNOWLEDGEMENTS

MPT and FW would like to acknowledge discussions with Michael Monteforte, Sven Jahnke, Rainer Engelken, and Guillaume Lajoie. This work was supported by BMBF (01GQ07113, 01GQ0811, 01GQ0922, 01GQ1005B), GIF (906-17.1/2006), DFG (SFB 889), VW-Stiftung (ZN2632), and the Max Planck Society.

AUTHOR CONTRIBUTIONS

MPT conceived the project, developed the concepts, and performed analytical and numerical calculations. FW supervised the project and contributed calculations. MPT and FW discussed the results and wrote the manuscript.

ADDITIONAL INFORMATION

The authors declare no competing financial interests.

Appendix A: Event-based simulations

A true phase representation is defined on a circular domain, *e.g.* $[0,1]^N$ with 0 and 1 identified. The phase dynamics we analyze is termed a *pseudo*phase representation, $\vec{\phi}(t) \in (-\infty, 1]^N$ since the phase can be kicked to a negative value by an inhibitory input arriving when the voltage is near its reset value, $V \approx V_{\text{res}}$. We hereon drop *pseudo* from the term.

The complete phase representation dynamics is given by:

$$\dot{\phi}_n(t) = T_{\text{free}}^{-1} + \sum_s A_{nn_s}\delta(t - t_s)Z(\phi_n(t_s)) \qquad \text{(A1)}$$

with constant phase velocity, T_{free}^{-1}, the phase response curve, $Z(\phi)$, and a spike-reset rule: when $\phi_n = 1$, ϕ_n is reset to 0. Note that in the large-K limit the phase and the voltage representation converge onto one another (see [14]).

Event-based simulations of Eq. (A1) were implemented by iterations of a spike-time map that takes the network state from just after one spike, t_s^+, to just after the next, t_{s+1}^+, where s is the index of the network spike sequence. The next spike time, t_{s+1}, and next spiking neuron in the sequence are obtained simply in the phase representation via

$$t_{s+1} = t_s + \min_{n\in\{1,\ldots,N\}} \left(1-\phi_n(t_s)\right) T_{\text{free}} \,,$$
$$n_{s+1} = \text{argmin}_{n\in\{1,\ldots,N\}} \left(1-\phi_n(t_s)\right) T_{\text{free}} \,,$$

respectively. An iteration consists of evolving the network phases to this next spike time, t_{s+1}, applying the pulse of size $Z(\phi_m(t_{s+1}))$ to the postsynaptic neurons, $\{m|A_{mn_{s+1}} = 1\}$, and then resetting the phase of the spiking neuron, n_{s+1}. For further details, as well as methods for computing the Lyapunov spectrum for this network, see [14].

This implementation was used to apply perturbations to the system, and measure the subsequent activity. The model was simulated in isolation for a time before the application of the perturbation to allow it time to relax onto the stationary measure.

Appendix B: Determining the critical perturbation strength

The separation of trajectories is quantified using the 1-norm distance,

$$D_t(\epsilon) := \frac{1}{N}\sum_{n=1}^{N} |\phi_{n,t} - \phi_{n,t}(\epsilon)| \,, \tag{B1}$$

between $\vec{\phi}_t$ and the perturbed trajectory, $\vec{\phi}_t(\epsilon)$, evolving freely from the perturbed state, $\vec{\phi}_0(\epsilon) := \vec{\phi}_0 + \vec{\epsilon}$ with the perturbation time set as $t = 0$ and the perturbation vector, $\vec{\epsilon} := \frac{\epsilon}{\sqrt{N}}\vec{\xi}$. The norm of $\vec{\epsilon}$ is order 1 in N for $\sigma_{\xi_n}^2$ order 1. ϵ^* is the largest value below which $D_t(\epsilon)$ vanishes in time. We use a bisection method on ϵ (described in Appendix C) to obtain ϵ^*.

Appendix C: Estimation of the critical perturbation, ϵ^*

A random perturbation direction, $\vec{\xi}$, was obtained by sampling $N-1$ times from a standard normal distribution, normalizing this vector, and projecting it into the N-dimensional phase space such that it was orthogonal to the phase velocity vector $\vec{\omega} = \left(T_{\text{free}}^{-1}, \ldots, T_{\text{free}}^{-1}\right)$. Constrained to this hyper-plane, the perturbation alters only relative spike time differences, i.e. there is no global shift in spike times. For $\epsilon > 0$, the critical perturbation size, ϵ^*, in that direction was obtained using a bisection method. The initial estimate of ϵ^*, $\epsilon_0^* = J_0/\left(\sqrt{KN}\bar{\nu}\tau\right)$ was lower-bounded by $\epsilon_{low}^* = 10^{-4}\cdot\epsilon_0^*$, and upper-bounded by $\epsilon_{up}^* = 1$. The estimate ϵ_0 was iteratively refined based on a divergence flag on the distance between the perturbed and unperturbed trajectories at time T after the perturbation:

$$\begin{aligned} &\text{If } D_T(\epsilon_i) > D_{thresh}, \\ &\quad \text{then } \epsilon^*_{\text{upp}} \leftarrow \epsilon_i^* \;; \\ &\text{else} \\ &\quad \epsilon^*_{\text{low}} \leftarrow \epsilon_i^* \quad , \end{aligned}$$

for iteration index i, where $D_{thresh} = 0.01$ denotes the threshold chosen to lie between the two well-separated modes of the end-distance distribution. (D_t eventually saturates due to the bounded phase-space at the average distance, $\bar{D}$, between a pair of random trajectories, and computed in Ref. [14].) A bisection step was then made,

$$\epsilon_{i+1}^* = \frac{\epsilon^*_{\text{upp}} + \epsilon^*_{\text{low}}}{2} \,,$$

to obtain the estimate of the next iteration. The procedure was repeated until the differences in successive values of ϵ_i^* fell below a tolerance threshold of 10^{-8}, and the final estimate taken as ϵ^*.

Appendix D: Constructing the folded phase space representation

Here, we describe the procedure used to construct the folded representation of the phase space around the attracting trajectory shown in Fig.2b and the *SI Movie*. Similar to Fig. 7 in Ref. [14], the same, random 2D projection of the $(N-1)$-dimensional subspace orthogonal the trajectory was applied at each iteration of the event map. This subspace remains unchanged by the evolution since in the phase representation the trajectory is always parallel to the main diagonal of the unit hypercube. Then, a rectilinear grid of initial conditions were generated in these planes. The network was simulated from each initial condition and the corresponding grid of end-states stored. A corresponding grid of the pairwise distances between end-states of all adjacent initial conditions was computed. Distances falling in the finite-distance mode of the resulting bi-modal end-state distance distribution centered around the average distance, $\bar{D}$, were used to identify adjacent initial conditions spanning a putative flux tube boundary. A putative tube identity label was assigned to each continuous region of corresponding initial conditions enclosed by these putative boundaries in the grid. We occasionally observed

single tubes segregated into disjoint pieces in our 2D representation by the occlusion of another tube, consistent with the layering of projections as proposed in Fig. 5b. For robustness then, a round of amalgamation of tube identities was performed by identifying as the same any two tubes whose centers of mass gave a end-state distance which fell below a threshold of 0.01. Again, that the modes were well separated made for unambiguous flagging.

This algorithm to compute a single cross-section was then repeated at each spike of the network activity in a simulated time window to obtain a set of successive cross sections orthogonal to and centered on the stable trajectory. To present this data, a folded representation is used in which these cross sections are placed contiguously so that the center trajectory passes through them continuously. This gives a 2+1D representation of the tube and its neighborhood along the stable trajectory, oriented such that the line $(0, 0, t)$ is horizontal with time increasing to the right. The identity of the center tube is trivially maintained across sections since the $(0, 0)$-perturbation leaves the stable trajectory unchanged. To keep track of the identities of the surrounding tubes represented in the successive sections requires an identity list passed forward and updated from section to section. We constructed such a list by again comparing all pairwise end-state distances of the center of masses of all cells of the previous and current cross sections. We identified successive cells as coming from the same tube if this distance fell below a threshold. Identities were added when a current cell had no match in the previous section corresponding to the event of a new tube entering the section. Identities were removed when a cell in the previous section had no match in the current section corresponding to the event of an existing tube leaving the section. We then used this identity list to color the cells, using an adaptive color assignment scheme in order to keep the range of colors reasonably bounded. This scheme randomly assigned unused colors, orphaned from tubes that had exited the section, to the cells of new tubes that had entered the section.

Appendix E: Instability caused by spike collision events

In this section, we determine from simulations of the dynamics that (1) the perturbed trajectories begin to diverge where a difference in the spike sequence appears; (2) this change is associated with a vanishing interval; and (3) this interval is between susceptible spikes, i.e spikes from a pair of neurons that exhibit one of the three connected-pair motifs.

Over perturbation directions, an ensemble of pairs of perturbed trajectories were simulated using a perturbation strength just above, ϵ^{*+}, and just below, ϵ^{*-}, the estimate obtained according to the procedure described in Appendix C (notation: $x^{\pm} = \lim_{\delta\to 0} x \pm \delta$). From the simulation started at ϵ^{*^+}, the decorrelation index, s^*, was extracted as the index in the spike sequence at which a sustained difference between the pair of sequences began. We denote elements of the perturbed spiking neuron sequence and spike times as $n_s(\epsilon)$ and $t_s(\epsilon)$, respectively.

We first show that the sustained jump in distance begins at the decorrelation index, s^*. We aligned by s^* across trials the distances, $D_s(\epsilon^{*+})$, to the unperturbed trajectory from the perturbed trajectory started from ϵ^*. The result, in Fig. E.1a shows the high correlation across trials.

Next, in Fig. E.1b,c we see that the spike time interval, $t_{s^*+1}(\epsilon) - t_{s^*}(\epsilon)$ corresponding to s^* before ($\epsilon = \epsilon^{*-}$) and after ($\epsilon = \epsilon^{*+}$) the collision event, respectively, vanishes only when $A_{n_{s^*} n_{s^*+1}} = 1$. In addition, $t_{s^*+1}(\epsilon) - t_{s^*}(\epsilon)$ scales inversely with the precision of the bisection algorithm used to obtain ϵ^*, demonstrating that the event is indeed generated as two spikes become coincident, $t_{s^*+1}(\epsilon) \to t_{s^*}(\epsilon)$ as $\epsilon \to \epsilon^*$ (see Fig. E.1d).

Appendix F: Spike collision motifs

In the main text we focused on the backward-connected motif. In this section, we discuss the forward-connected and symmetrically-connected motif. Across these motifs, under consideration is a situation where an output spike time of a given neuron, t_{out}, is near in time to an input spike time, t_{in}, that this neuron receives. When the output spike is generated before the input spike, $t_{out} < t_{in}$ (the backward-connected motif), a collision can occur when a perturbation leads to the vanishing of the interval between them, an example of which is shown in Fig. 3 in main text. If $t_{in} < t_{out}$ (the forward-connected motif), however, the inhibition means that t_{in} already delays t_{out} for $\epsilon < \epsilon^*$ so that t_{out} can occur no closer to t_{in} than Δt_{jump}, for the same reason that t_{out} undergoes a jump forward in the backward-connected motif. Thus, a collision event occurs in this motif when the perturbation brings t_{in} and t_{out} to within Δt_{jump} of each other.

The two asymmetric motifs give collision scenarios that are identified under a reversal of the direction of change in perturbation strength. The forward and backward connected motif can be distinguished by whether the collision event is approached by an input spike moving forward, $\mathrm{d}t_{in}/\mathrm{d}\epsilon > 0$, or backward, $\mathrm{d}t_{in}/\mathrm{d}\epsilon < 0$, over t_{out} with t_{out} as the reference time. In the forward-connected motif, the interval vanishes, $t_{in} \to 0^+$, for $\epsilon \to \epsilon^{*+}$, *i.e.* just after the collision. In the backward-connected motif, the vanishing interval, $t_{in} \to 0^+$, occurs as $\epsilon \to \epsilon^{*-}$, *i.e.* just before the collision. For either case, when on the side of ϵ^* where the interval is vanishing, the input spike comes after the output spike, $t_{in} > 0$, in this reference

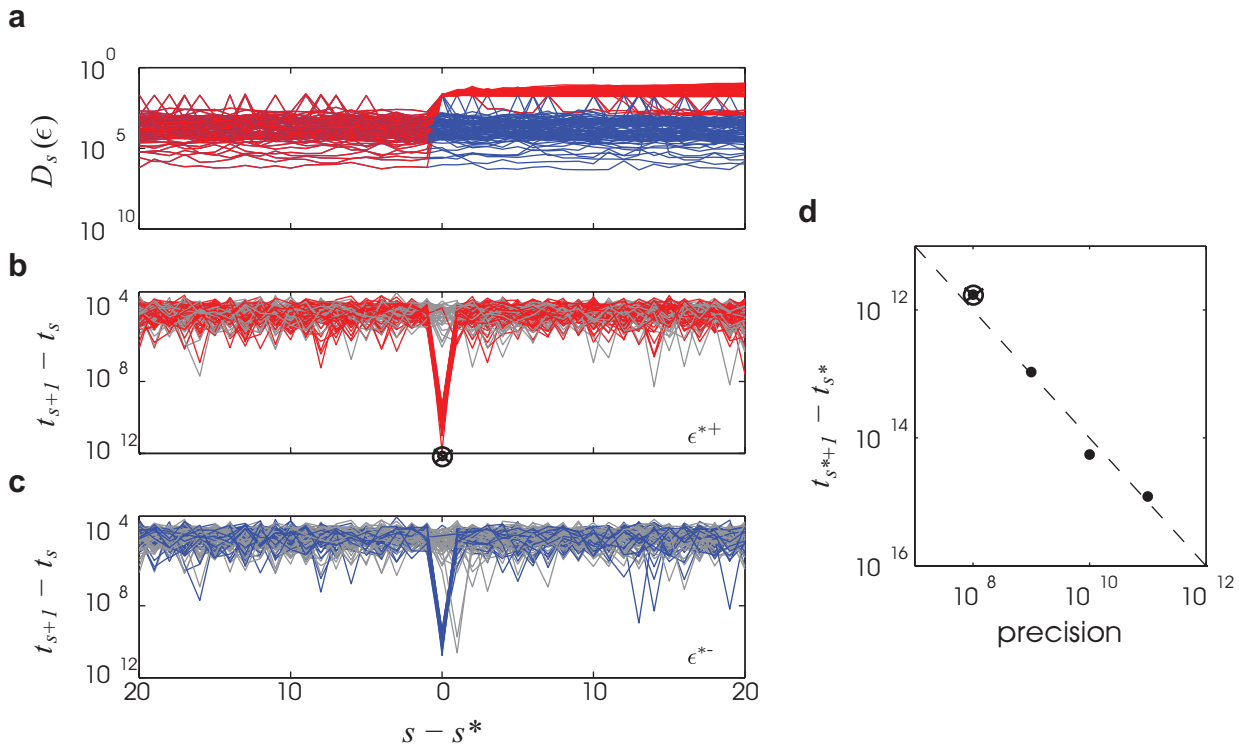

Figure E.1. Characteristics of divergence events. **(a)** A small window of the distance time series aligned to index, s^*, at which the decorrelation of the spike sequence begins. The supercritical perturbed (red) trajectory started from ϵ^{*+} jumps up away from the subcritical perturbed (blue) trajectory started at ϵ^{*-} at s^* (note the logarithmic scale on the ordinate). **(b)** and **(c)** show the spike intervals, $t_{s+1} - t_s$, for the $\epsilon = \epsilon^{*+}$ and $\epsilon = \epsilon^{*-}$ trajectories, respectively (colors as in (a)). Realizations not exhibiting the $n_{s^*-} \to n_{s^*-+1}$ and $n_{s^*-} \leftarrow n_{s^*-+1}$, respectively, have been grayed out. Note that those left colored have a significantly smaller interval at index s^*. **(d)** Coincidence of successive spikes with increasing precision (decreasing tolerance) of the bisection algorithm used to find ϵ^*. Here, a shrinking interval taken from a $\epsilon = \epsilon^{*+}$ realization has been used (see the identified minimum in panel b).

frame.

In each of these two asymmetric motifs, only one of the pair of spikes undergoes a jump of size Δt_{jump}. For the bidirectionally connected motif, however, both spikes undergo a jump of size Δt_{jump} simultaneously, by which they exchange spike times, and so no vanishing interval exists on either side of the flux tube boundary. A collision event occurs in this motif with reduced relative frequency, p, compared with the two asymmetric cases and so is negligible for sparse networks, $p \ll 1$.

The characteristics of an inhibitory event at threshold is a single neuron property, dependent on the neuron model, and so can be investigated for many neuron models. Since the LIF solution, is invertible, one can explicitly solve for the time, Δt_{jump}, that the inhibitory event has delayed the spike. With initial condition, $V(0) = V_T^- + J \approx J$,

$$V_T = \sqrt{K} I_0 - \left(\sqrt{K} I_0 + J\right) e^{-\frac{\Delta t_{\text{jump}}}{\tau}}$$
$$\Delta t_{\text{jump}} = \tau \ln\left(1 + \frac{J}{\sqrt{K} I_0}\right) \tag{F1}$$

Using the balance equation, Eq. 3, we obtain $\Delta t_{\text{jump}} \sim \tau \ln\left(1 + (K\bar{\nu}\tau)^{-1}\right) \sim (K\bar{\nu})^{-1}$, for $K \gg 1$, as stated in the main text.

The inhibition prohibits susceptible spike pairs in the forward-connected (and bidirectional) motif that occur closer than $1/(K\bar{\nu})$. Thus, these pairs are separated in time by on average $2/(K\bar{\nu})$ in the unperturbed trajectory. However, since they collide when they come within $1/(K\bar{\nu})$ from one another, the susceptible pairs in a collision event for the forward-connected and symmetric motifs are effectively separated by the same perturbation distance as those pairs satisfying the backward-connected motif.

Appendix G: Collision events for neuron models with smooth interaction and threshold dynamics

Here, we demonstrate that for two natural extensions of the LIF neuron model into the non-limiting regime of finite derivatives in the dynamics, the abrupt spike time shifts perists.

We simulated the inhibitory input spike at threshold event for a neuron model with an active spike-generating mechanism, the rapid-theta neuron (18). The theta neuron model on which it is based is the phase representation of the normal form of a saddle node bifurcation to periodic firing and thus its features are universal to all neuron models operating near this transition. It has an additional parameter relative to that model, the rapidness r, that controls the speed at which the voltage diverges. With the addition of finite speed of action potential onset, *i.e.* with a smooth threshold, the spike jump has a similar magnitude as in the LIF case, but now grows smoothly with perturbation strength after the collision (Fig.3d, bottom). As the action potential onset rapidness increases, however, this growth becomes sharper, approaching the discontinuous jump for the LIF neuron (Fig. H.1c). Divergence events thus result from spike collisions in this regime and the divergence rates have been quantified (18) similarly to the case of the LIF (14).

We also simulated the inhibitory event at threshold also for the LIF neuron with the addition of an integrating synaptic current compartment so that $I_n(t) = I_{\text{ext}} + I_{\text{syn}}(t)$ where $I_{\text{syn}}(t)$ is governed by $\tau_{\text{syn}} \frac{\mathrm{d}}{\mathrm{d}t} I_{\text{syn}} = -I_{\text{syn}} + \tau J \sum_s A_{nn_s} \delta(t - t_s)$. With the addition of finite synaptic current kinetics to the model that low-pass filter the input and smooth the interaction dynamics, the spike time jump is still instantaneous with magnitude approaching that of the LIF neuron for vanishing synaptic time constant (Fig.H.1d, top). Thus, collision events for finite-speed kinetics can induce divergence events. Since the corresponding jump size decreases as the kinetics are slowed, however, this collision-based instability is less likely to induce a cascade for slower kinetics and so its destabilizing effect on the dynamics is weakened the further away the system is poised from the LIF regime.

Appendix H: Cascade probability

For ϵ approaching ϵ^*, the presynaptic spike time, $t_{s'=s^*+1}$, is advanced relative to the postsynaptic spike time, t_{s^*}, until the two spikes collide (see Fig. 3a). At collision ($\epsilon = \epsilon^*$), the pulsed inhibition and the voltage's rate of approach to threshold cause an abrupt delay of t_{s^*} by Δt_{jump} (Fig. 3b). Using Eq. (3) and the single neuron dynamics we obtain

$$\Delta t_{\text{jump}} = \tau \ln [1 + d] \approx \tau d = (K\bar{\nu})^{-1} \tag{H1}$$

for $d \ll 1$.

Since $\Delta t_{\text{jump}} \approx \omega_K^{-1}$, the spike time of neuron n_{s^*} is typically shifted far enough forward to cross a spike emitted by a neuron in its postsynaptic population. Formally, the probability, p_{post}, of a spike emitted by any of the post-synaptic neurons during the window of size $(K\bar{\nu})^{-1}$ over which the output spike has jumped is $K\left(1 - e^{-1/K}\right) = 1 - \frac{1}{2K} + \mathcal{O}\left(K^{-2}\right)$. Since the subsequent activity of a neuron involved in a collision event is irreversibly altered, there are on average $\log N / \log K < N/K = 1/p$ number of these events until the activities of all neurons have been altered. A lower bound on the probability of a cascade is then $(p_{\text{post}})^{1/p} \approx \left(1 - \frac{1}{2pN}\right)^{1/p} \to 1$, in both the sparse ($N \to \infty$ and pN fixed) and dense ($N \to \infty$ and p fixed) thermodynamic limit.

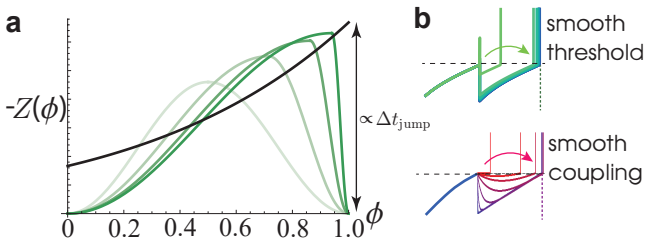

Figure H.1. *Spike crossing instability in other models.* **(a)** Phase response curves for a smooth threshold neuron model (green) qualitatively approximate that of the LIF neuron (black) as action potential onset rapidness is increased (light to dark). **(b)** Parametrized family of voltage traces of an inhibitory event near threshold for other neuron models. *Top*: A smooth threshold model parametrized by increasing action potential onset rapidness (green to blue). *Bottom*: A smooth synaptic coupling model parametrized by decreasing synaptic time constant (red to blue).

Appendix I: Susceptibilities, a_s

The spike time deviation, $\delta t_s(\epsilon) := t_s(\epsilon) - t_s(0)$, is composed of a contribution by the direct perturbation to n_s, and a contribution from the indirect effects of the perturbation via deviations of the input spike times to n_s. The deviations from both of these contributions will be contracted across subsequent input spikes to that neuron. The derivative with respect to perturbation strength thus consists of a differential due to changing initial state with fixed input spike times and due to changing input spike times with the initial state fixed, respectively:

$$\frac{\mathrm{d}t_s}{\mathrm{d}\epsilon} \approx \frac{\partial t_s}{\partial \epsilon} + \sum_{j=1}^{s-1} \frac{\mathrm{d}t_j}{\mathrm{d}\epsilon} \frac{\partial t_s}{\partial t_j} \tag{I1}$$

The chain-rule calculation is

$$
\begin{aligned}
\frac{\mathrm{d}t_s}{\mathrm{d}\epsilon} &\approx \frac{\partial t_s}{\partial \epsilon} + \sum_{j=1}^{s-1} \frac{\mathrm{d}t_j}{\mathrm{d}\epsilon}\frac{\partial t_s}{\partial t_j} \\
&= \frac{\mathrm{d}\phi_{n_s}\left(t_1^-\right)}{\mathrm{d}\epsilon}\left(\prod_{j=1}^{s-1}\frac{\partial\phi_{n_s}\left(t_j^+\right)}{\partial\phi_{n_s}\left(t_j^-\right)}\right)\frac{\partial t_s}{\partial\phi_{n_s}\left(t_{s-1}^+\right)} + \sum_{j=1}^{s-1}\frac{\mathrm{d}t_j}{\mathrm{d}\epsilon}\frac{\partial\phi_{n_s}\left(t_j^-\right)}{\partial t_j}\left(\prod_{k=j+1}^{s-1}\frac{\partial\phi_{n_s}\left(t_k^+\right)}{\partial\phi_{n_s}\left(t_k^-\right)}\right)\frac{\partial t_s}{\partial\phi_{n_s}\left(t_{s-1}^+\right)} \\
&= \left(\frac{\xi_{n_1}}{\sqrt{N}}\right)\left(\prod_{j=1}^{s-1}\left(1+d_{\phi_s^j}\right)^{A_{n_s n_j}}\right)(-T_{free}) + \sum_{j=1}^{s-1}\left(-\frac{1}{T_{free}}A_{n_s n_j}d_{\phi_s^j}\right)\left(\prod_{k=j+1}^{s-1}\left(1+d_{\phi_s^k}\right)^{A_{n_s n_k}}\right)(-T_{free})\frac{\mathrm{d}t_j}{\mathrm{d}\epsilon} \\
\frac{\mathrm{d}t_s}{\mathrm{d}\epsilon} &= \frac{-T_{free}}{\sqrt{N}}\xi_{n_1}\left(\prod_{j=1}^{s-1}\left(1+d_{\phi_s^j}\right)^{A_{n_s n_j}}\right) + \sum_{j=1}^{s-1}A_{n_s n_j}d_{\phi_s^j}\left(\prod_{k=j+1}^{s-1}\left(1+d_{\phi_s^k}\right)^{A_{n_s n_k}}\right)\frac{\mathrm{d}t_j}{\mathrm{d}\epsilon}
\end{aligned}
$$

where $d_{\phi_s^j}$ is shorthand for the derivative of the PRC,

$$d_{\phi_s^j} := Z'(\phi_{n_s}(t_j)) ,$$

evaluated at the phase of the n_s neuron at the time of the j^{th} spike in the network spike sequence, and where the perturbation direction vector, $\vec{\xi}$, is not normalized but explicitly divided by $\sqrt{N}$, preserving the $\mathcal{O}\left(1/\sqrt{N}\right)$-scaling of a unit vector.

The result contains three contributions: perturbations to n_s, perturbations to neurons connected to n_s, and the contraction events from input spikes to n_s. Dividing the result by $\frac{-T_{free}}{\sqrt{N}}$, and rescaling the perturbation to $\tilde{\epsilon} = \frac{-T_{free}}{\sqrt{N}}\epsilon$, we obtain

$$\delta t_s(\epsilon) = \frac{\mathrm{d}t_s}{\mathrm{d}\tilde{\epsilon}}\frac{\mathrm{d}\tilde{\epsilon}}{\mathrm{d}\epsilon}\epsilon = -\frac{T_{free}}{\sqrt{N}}a_s\epsilon , \tag{I2}$$

as stated in the main text (Eq. (14)).

Appendix J: Details of Derivation of flux tube diameter distribution

The state being perturbed at $t = 0$, $\vec{\phi}_0$, is an equilibriated state whose probability density function depends in general on the realization of the connectivity, $A = (A_{mn})$. For large, sparse connectivities, however, the self-averaging properties of A leave the invariant density $\rho\left(\vec{\phi}_0\right)$ dependent only on the parameters of the connectivity ensemble and not the particular realization. A closed form for this density has been previously derived (see [14]), though we will not need it here since the dependence of $d_{\phi_s^j}$ on ϕ_s^j becomes negligible at large K (Eq. (7)).

We rather require the distribution of unperturbed intervals. In a diffusion approximation, applicable to large, sparse graphs, the inputs to different neurons are negligibly correlated. Each of the set of unperturbed inter-spike intervals, $\{\Delta t_s\}$, of the compound spike sequence obeys a distribution that with increasing N rapidly approaches the same exponential form with rate $N\bar{\nu}$, $\rho(\Delta t_s) = N\bar{\nu}e^{-N\bar{\nu}\Delta t_s}$ for all s (see Fig. J.1).

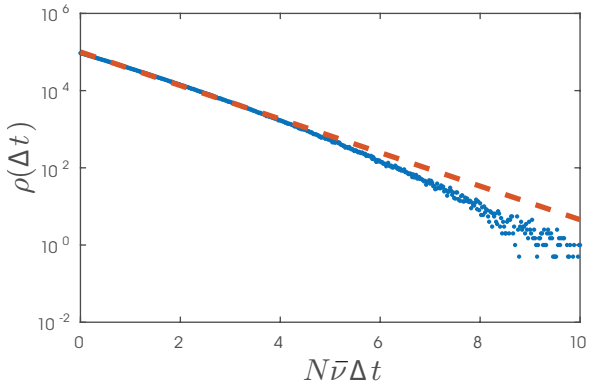

Figure J.1. *Network spike time interval probability density,* $\rho(\Delta t)$. $\rho(\Delta t)$ is distributed exponentially ($N = 10^4$, $\bar{\nu} =$ 10 Hz, 10^7 network intervals). Dashed line is the prediction, $\rho(\Delta t) = N\bar{\nu}e^{N\bar{\nu}\Delta t}$. Note that the abscissa is scaled by $N\bar{\nu}$.

The distribution of $\{\Delta t_s\}$ is then

$$\rho(\{\Delta t_s\}) = \prod_{s=2}^{M}\rho(\Delta t_s) . \tag{J1}$$

The susceptibilities, Δa_s, simplify in three ways. The size of indirect effects (the second term in Eq. (14)) are suppressed for large K, since they additionally contain $d_{\phi_s^j} \propto K^{-1}$ as a factor. Thus,

$$a_s \approx \xi_{n_s}\prod_{j=1}^{s-1}\left(1+d_{\phi_s^j}\right)^{A_{n_s n_j}} . \tag{J2}$$

The small synaptic strength linearizes $Z(\phi)$ for $K \gg 1$ so that $d_{\phi_s^j} \approx -d$ with $d := (K\bar{\nu}\tau)^{-1} > 0$ (Eq. (7)) so a_s no longer depends on the distribution of states. Third, for non-small s a fraction p of the earlier spikes $\{1, \ldots, s-1\}$ are from neurons presynaptic to

n_s so that $a_s \approx \xi_{n_s}(1-d)^{\sum_{j=1}^{s-1} A_{n_s n_j}} \approx \xi_{n_s}(1-d)^{ps} = \xi_{n_s}\left(\left(1-(\bar{\nu}\tau)^{-1}/K\right)^K\right)^{\frac{s}{N}} \approx \xi_{n_s} e^{-pds}$ for $K \gg 1$. Thus, $\Delta a_s \approx e^{-pds}\xi_{n_s} - e^{-pd(s-1)}\xi_{n_{s-1}} \approx e^{-pds}\left(\xi_{n_s} - \xi_{n_{s-1}}\right)$, since $e^{pd} \approx 1$ for $N \gg 1$. We note that $-pds \approx \lambda t$ where $\lambda = -\tau^{-1}$ serves here as an estimate for mean Lyapunov exponent, λ_{mean}, at large K, calculated in [14].

σ_ξ then determines the numeric prefactor in the standard deviation of Δa_s, $\sigma_{\Delta a_s}$, and so can be set to make this prefactor unity. $\rho(\xi)$ was chosen as a centered normal distribution in order to generate isotropic perturbation directions. The difference of two independent centered normal random variables has 0 mean and twice the variance. Thus, $\sigma_{\Delta a_s} = \sqrt{2}\sigma_\xi e^{-pds}$. We also note that $[a_s]_{\rho(\vec{\xi})} = 0$ when serial correlations are negligible, $\left[\xi_{n_s}\xi_{n_{s-1}}\right] = \delta_{n_s n_{s-1}}$, as assumed in the main text, and $[\xi_{n_s}]_{\rho(\vec{\xi})} = 0$ for the isotropic perturbation direction distributions used here.

The expectation of $S_s(\epsilon)$ (Eq. (20)) is then evaluated as

$$\begin{aligned} S_s(\epsilon) &= \left[\Theta\left(\Delta t - \frac{T_{free}}{\sqrt{N}}\Delta a_s \epsilon\right)^{A_{mn}}\right]_{\rho(\Delta t)\rho(\Delta a_s)P_{A_{mn}}(A_{mn})} \\ &= (1-p) + p\frac{2}{\sqrt{\pi}}e^{c_s^2}\int_{c_s}^{\infty} e^{-y_s^2}\mathrm{d}y_s \\ &= (1-p) + pe^{c_s^2}\left(1 - \int_0^{c_s} e^{-y_s^2}\mathrm{d}y_s\right) \\ &= 1 + p\left(\mathrm{Erfcx}\left[c_s\right] - 1\right) \end{aligned}$$

with $y_s = x_s + c_s$, $x_s = \Delta a_s/\left(\sqrt{2}\sigma_{\Delta a_s}\right)$, and $c_s = \frac{T_{free}}{\Delta t}\frac{\epsilon}{\sqrt{N}}\sigma_{\Delta a_s}$ and where $\mathrm{Erfcx}[x] = e^{x^2}\left(1 - \frac{2}{\sqrt{\pi}}\int_0^x e^{-y^2}\mathrm{d}y\right)$ is the scaled complementary error function. Using the approximation $\mathrm{Erfcx}[c_s] - 1 \approx -c_s$ for $c_s \ll 1$ (true if $\epsilon/\sqrt{p} \ll 1$), we obtain Eq. (21).

* puelma@lpt.ens.fr

[1] Hopfield JJ Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79(8)**, 2554–2558 (1982).

[2] Gardner E Optimal basins of attraction in randomly sparse neural network models. *J. Phys. A* **22(12)** 1969-1987(1989).

[3] Sompolinsky H, Crisanti A, Sommers HJ Chaos in random neural networks. *Phys. Rev. Lett.* **61(3)** 259–262 (1988).

[4] Kadmon J, Sompolinsky H (2015) Transition to chaos in random neuronal networks. *Phys. Rev. X* 5(4):1–28.

[5] Abbott LF, DePasquale B, Memmesheimer RM Building Functional Networks of Spiking Model Neurons. *Nat. Neurosci.* **19**, 1–16 (2016).

[6] Gutig R Spiking neurons can discover predictive features by aggregate-label learning. *Science* **351(6277)** 1–13 (2016).

[7] Jahnke S, Memmesheimer RM, Timme M How Chaotic is the Balanced State? *Frontiers in computational neuroscience* **3** 13 (2009).

[8] Engelken R, Wolf F Dynamical entropy production in cortical circuits with different network topologies in *COSYNE Abstracts* (2013).

[9] Zillmer R, Brunel N, Hansel D (2009) Very long transients, irregular firing, and chaotic dynamics in networks of randomly connected inhibitory integrate-and-fire neurons. *Phys. Rev. E* 79(3):1–13.

[10] Rosenbaum R, Doiron B Balanced networks of spiking neurons with spatially dependent recurrent connections. *Phys. Rev. X* **4(2)** 1–9 (2014).

[11] Ozeki H, Finn, IM, Schaffer ES, Miller KD, Ferster D Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* **62(4)** 578 (2009).

[12] Wolf F, Engelken R, Puelma Touzel M, Weidinger JDF, Neef A Dynamical models of cortical circuits. *Curr. Opin. Neurobiol.* **25** 228–236 (2014).

[13] Jahnke S, Memmesheimer RM, Timme M Stable irregular dynamics in complex neural networks. *Phys. Rev. Lett.* **100(4)**:2–5 (2008).

[14] Monteforte M, Wolf F Dynamic flux tubes form reservoirs of stability in neuronal circuits. *Phys. Rev. X* **2(4)** 041007 (2012) .

[15] Jin DZ Fast convergence of spike sequences to periodic patterns in recurrent networks. *Phys. Rev. Lett.* **89(20)** 208102 (2002).

[16] Zillmer R, Livi R, Politi A, Torcini A Desynchronization in diluted neural networks. *Phys. Rev. E* **74(3)** 1–10 (2006).

[17] Tuckwell HC *Introduction to Theoretical Neurobiology: Volume 2, Nonlinear and Stochastic Theories*, Cambridge Studies in Mathematics. (Cambridge University Press, 2005).

[18] Lindner B Superposition of many independent spike trains is generally not a Poisson process. *Phys. Rev. E* **73(2)** 1–4 (2006).

[19] Monteforte M Ph.D. thesis (Georg-August University, 2011).

[20] Puelma Touzel M Ph.D. thesis (Georg-August University, 2015).

[21] Monteforte M, Wolf F Dynamical entropy production in spiking neuron networks in the balanced state. *Phys. Rev. Lett.* **105(26)** 1–4 (2010).

[22] Harish O, Hansel D Asynchronous Rate Chaos in Spiking Neuronal Circuits. *PLOS Comput. Biol.* **11(7)**, e1004266 (2015).

[23] Mastrogiuseppe F, Ostojic S Intrinsically-generated fluctuating activity in excitatory-inhibitory networks. *PLOS Comput. Biol.* **13(4)** 1–33 (2017).

[24] Engelken R, Farkhooi F, Hansel D, van Vreeswijk C, Wolf F A reanalysis of "Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons". *F1000Res* **5(0)** 2043 (2016).

[25] Engelken R, Wolf F (2015) Input spike trains reduce dynamical entropy production in balanced networks in *Bernstein Conference Abstracts*.

[26] Molgedey L, Schuchhardt J, Schuster HG Suppressing chaos in neural networks by noise. *Phys. Rev. Lett.* **69(26)** 3717–3719 (1992).

[27] Toyoizumi T, Abbott LF Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Phys. Rev. E* **84(5)** 1–8

(2011).
[28] Lajoie G, Lin KK, Shea-Brown E Chaos and reliability in balanced spiking networks with temporal drive. *Phys. Rev. E* **87(5)** 1–5 (2013).
[29] Lajoie G, Thivierge JP, Shea-Brown E Structured chaos shapes spike-response noise entropy in balanced neural networks. *Front. Comput. Neurosci.* **12(12)** e1005258 (2014).
[30] Memmesheimer RM, Rubin R, Ölveczky BP, Sompolinsky H Learning Precisely Timed Spikes. *Neuron* **82(4)** 925–938 (2014).
[31] Arnold L *Random Dynamical Systems.* (Springer, 1991).
[32] Laje R, Buonomano DV Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat. Neurosci.* **16(7)** 925–33 (2013).
[33] Boutent M, Engels A, Komodat A, Serneelst R Quenched versus annealed dilution in neural networks. *J.Phys. A* **23(20)** 4643-4657 (1990).
[34] Zhao L, Beverlin B, Netoff T, Nykamp DQ Synchronization from second order network connectivity statistics. *Front. Comput. Neurosci.* **5** 28 (2011).
[35] Schwalger T, Droste F, Lindner B Statistical structure of neural spiking under non-poissonian or other non-white stimulation. *J. Comput. Neurosci.* **39(1)**:29–51 (2015).
[36] London M, Roth A, Beeren L, Häusser M, Latham PE Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* **466(7302)** 123–7 (2010).
[37] Houweling AR, Brecht M Behavioural report of single neuron stimulation in somatosensory cortex. *Nature* **451** 65–68 (2008).
[38] van Vreeswijk C, Sompolinsky H Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274(5293)** 1724–6 (1996).
[39] Brunel N, Hakim V Fast Global Oscillations in Networks of Integrate-and-Fire Neurons with Low Firing Rates. *Neural Comput.* **11(7)** 1621–1671 (1999).
[40] Renart A, et al. The asynchronous state in cortical circuits. *Science* **327(5965)**:587–90 (2010).
[41] Barral J, D Reyes A Synaptic scaling rule preserves excitatory–inhibitory balance and salient neuronal network dynamics. *Nat. Neurosci.* **19(12)** 1690–1696 (2016).