bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Effective synaptic interactions in subsampled nonlinear networks with strong coupling

Braden A. W. Brinkman,[1,2] Fred Rieke,[2,3,4] Eric Shea-Brown,[1,2,3,5] and Michael A. Buice[5,1]

[1] *Department of Applied Mathematics, University of Washington, Seattle, WA, 98195, USA*
[2] *Department of Physiology and Biophysics, University of Washington, Seattle, WA, 98195, USA*
[3] *Graduate Program in Neuroscience, University of Washington, Seattle, WA, 98195, USA*
[4] *Howard Hughes Medical Institute, University of Washington, Seattle, WA, 98195, USA*
[5] *Allen Institute for Brain Science, Seattle, WA, 98109, USA*

(Dated: February 18, 2017)

A major obstacle to understanding neural coding and computation is the fact that experimental recordings typically sample only a small fraction of the neurons in a circuit. Measured neural properties are skewed by interactions between recorded neurons and the "hidden" portion of the network. To properly interpret neural data, we thus need a better understanding of the relationships between measured effective neural properties and the true underlying physiological properties. Here, we focus on how the effective spatiotemporal dynamics of the synaptic interactions between neurons are reshaped by coupling to unobserved neurons. We find that the effective interactions from a presynaptic neuron r' to a post-synaptic neuron r can be decomposed into a sum of the true interaction from r' to r plus corrections from every directed path from r' to r through unobserved neurons. Importantly, the resulting formula reveals when the hidden units have—or do not have—major effects on reshaping the interactions among observed neurons. As a prominent example, we derive a formula for strong impact of hidden units in random networks with connection weights that scale with $1/\sqrt{N}$, where N is the network size—precisely the scaling observed in recent experiments.

Establishing relationships between a network's architecture and its function is a fundamental problem in neuroscience and network science in general. Not only is the architecture of a neural circuit intimately related to its function, but pathologies in wiring between neurons are believed to underlie diseases conditions [1–15].

A major obstacle to uncovering structure-function relationships is the fact most experiments can only directly observe small fractions of an active network. The state-of-the-art methods for determining connections between neurons in living networks is to infer them by fitting statistical models to neural spiking data [16–24]. However, the fact that we cannot observe all neurons in a network means that the statistically inferred connections are at best "effective" connections, representing some dynamical relationship between the activity of nodes but not necessarily a true physical connection [20, 25–31]. The exact relationship between the effective and true connections is in general unknown, making it difficult to extrapolate from the statistics of effective connections back to the statistics of the true connections. Establishing this relationship thus has immediate importance for interpreting experimental measurements of synaptic interactions. It also informs fundamental a question in network computation: how hidden units shape the dynamics of a subset of nodes.

To provide a foundation upon which we can begin to understand the relationship between true network connectivity and the effective connections between neurons, we analyze a probabilistic model of network activity in which all properties are known, and derive a novel approximation to the effective model for the network of subsampled observed neurons. This makes explicit how the synaptic interactions between neurons are modified by unobserved neurons in the network, and under what conditions these modifications are, and are not, significant. As an important example, we study a sparse, random Erdős-Réyni network with N cells and with synaptic weights that scale as $1/\sqrt{N}$ [32–35], as has been recently observed experimentally [36], and show how unobserved neurons significantly reshape the effective synaptic interactions away from the ground-truth interactions. This is not the case with more "classical" $1/N$ scaling. Hence, depending on the nature of the statistics of synaptic weights and how they scale with network size, hidden paths can have a major impact on the observed interactions between neurons, impacting both the local network computation and interpretation of experimental data.

Model—We model the full network of N neurons as a nonlinear Hawkes process [37]. This is commonly known as a "Generalized linear (point process) model" in neuroscience, and is broadly used to fit neural activity data [16–19, 21–24, 38]. Here we use it as a generative model for network activity, as it is a can be directly linked to common spiking models such as integrate and fire systems driven by noisy inputs [39, 40]. Similar studies of the impact of hidden neurons have been performed for the kinetic Ising model, another non-equilibrium model that has been applied to neural activity [41–43]. However, the versions of the kinetic Ising model studied thus far have a Markovian time-dependence, whereas the nonlinear Hawkes model we consider here may have arbitrary history dependence, important for correctly capturing the statistics of network dynamics.

To derive an approximate model for an observed subset of the network, we partition the network into two sets: recorded neurons (labeled by indices r) and hidden neurons (labeled by indices h). Each recorded neuron has an

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

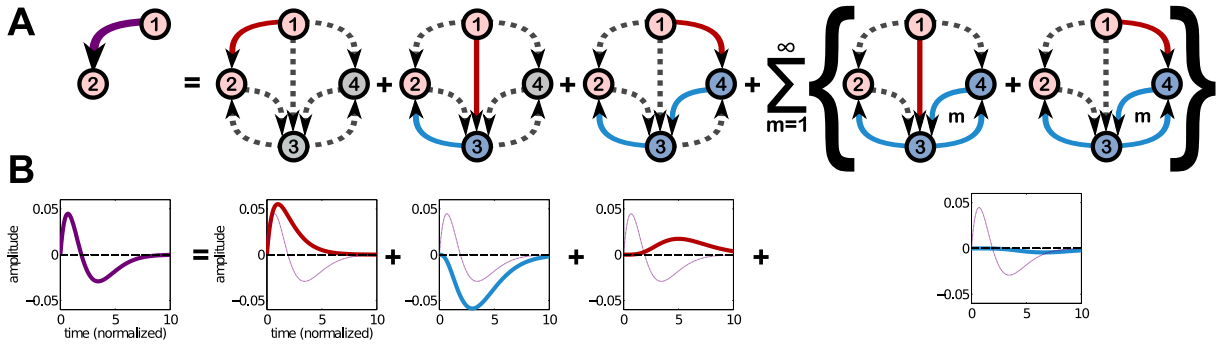

FIG. 1. The hidden unit problem: **A**: The effective interaction between two neurons can be decomposed into contributions from all paths from one neuron to the other through "hidden" intermediary neurons within the same circuit. In this schematic, neurons 1 and 2 are observed, while 3 and 4 are hidden. The effective interaction from 1 to 2 comprises a direct connection plus contributions from paths signals can travel through neurons 3 and 4 before arriving at neuron 2. **B**: Leftmost, the effective interaction from neuron 1 to 2. Subsequent plots decompose this interaction into contributions from each path in **A**. Only *directed* interactions from neuron 1 to 2 through hidden units contribute to the effective interaction. For instance, although neuron 2 makes a connection to neuron 3, this does not generate an effective interaction from 2 to 1 because neither neurons 3 nor 4 make a connection back to neuron 1, and hence there is no path that neuron 2 can send a signal to neuron 1.

instantaneous firing rate $\lambda_r(t)$ such that the probability that the neuron fires within a small time window $[t, t+dt]$ is $\lambda_r(t)dt$. The instantaneous firing rate in our model is

$$\lambda_r(t) = \lambda_0 \phi\left(\begin{array}{l} \mu_r + \sum_{r'} J_{r,r'} * \dot{n}_{r'}(t) \\ \qquad + \sum_h J_{r,h} * \dot{n}_h(t) \end{array} \right), \qquad (1)$$

where λ_0 is a characteristic firing rate, $\phi(x)$ is a non-negative, continuous function, μ_r is a tonic drive that sets the baseline firing rate of the neuron, and $J_{i,j} * \dot{n}_j(t) \equiv \int_{-\infty}^{\infty} dt' \; J_{i,j}(t-t')\dot{n}_j(t')$ is the convolution of the spike filter $J_{i,j}(t)$ with spikes $\dot{n}_j(t)$ *from* pre-synaptic neuron j *to* post-synaptic neuron i. In this work we will take the tonic drive to be constant in time, and focus on the steady-state network activity in response to this drive. We consider spike filters of the form $J_{i,j}(t) \equiv \mathcal{J}_{i,j} g_j(t)$, where the temporal waveforms $g_j(t)$ are normalized such that $\int_0^{\infty} dt \; g_j(t) = 1$ for all neurons j. Because of this normalization, the weight $\mathcal{J}_{i,j}$ carries units of time. Self-couplings $\mathcal{J}_{i,i}$ need not be interpreted as actual autapses, but rather account for the influence of a neuron's spiking history on its own firing rate (e.g., $\mathcal{J}_{i,i} < 0$ suppresses firing, like a refractory period, while $\mathcal{J}_{i,i} > 0$ promotes bursting). The firing rates for the hidden neurons follow the same expression with indices h and r interchanged.

We seek to describe the dynamics of the recorded neurons entirely in terms of their own set of spiking histories, eliminating the dependence on the activity of the hidden neurons. This demands averaging out the activity of the hidden neurons; in practice this is intractable to perform exactly [44–46]. Here, we use a mean field approximation and assume that the input from the hidden neurons can be approximated by its mean *conditioned on the activity of the recorded neurons*. This yields the following formula for the instantaneous firing rates of the recorded neurons:

$$\lambda_r(t) \approx \lambda_0 \phi\left(\mu_r^{\text{eff}} + \sum_{r'} J_{r,r'}^{\text{eff}} * \dot{n}_{r'}(t) \right).$$

Here, the effective baselines $\mu_r^{\text{eff}} = \mu_r + \sum_h \mathcal{J}_{r,h}\nu_h$ are simply modulated by the net input to the neuron, so we do not focus on them here. The effective coupling filters are given in the frequency domain by

$$\hat{J}_{r,r'}^{\text{eff}}(\omega) = \hat{J}_{r,r'}(\omega) + \sum_{h,h'} \hat{J}_{r,h}(\omega)\hat{\Gamma}_{h,h'}(\omega)\hat{J}_{h',r'}(\omega). \qquad (2)$$

Here, the ν_h are the steady-state mean firing rates of the hidden neurons and $\hat{\Gamma}_{h,h'}(\omega)$ is the linear response function of the hidden network to perturbations in the *input* [47]. Both ν_h and $\hat{\Gamma}_{h,h'}(\omega)$ are calculated *in the absence of the recorded neurons*, using a mean field theory approximation. The complete derivation is given in the Supplemental Material (SM).

In deriving these results, we have neglected both fluctuations around the mean input from the hidden neurons, as well as higher order filtering of the recorded neuron spikes. These results hold for any choice of network parameters for which the mean field steady state of the hidden network exists. For details on the justification of these approximations, see the SM.

The effective coupling filters are what we would—in principle—measure experimentally if we observe only a subset of a network. In practice, inferring these network properties from data is an extremely nontrivial task [16–24, 41, 42], and details of the fitting procedure could potentially further skew the inferred coupling filters. We will put aside these complications here, and assume we have access to an inference procedure that allows us to

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

measure $J^{\text{eff}}_{r,r'}(t)$ without error, so that we may focus on their properties and relationship to the ground-truth coupling filters.

Structure of effective coupling filters—The ground-truth coupling filters $\hat{J}_{r,r'}(\omega)$ are modified by a correction term $\sum_{h,h'} \hat{J}_{r,h}(\omega)\hat{\Gamma}_{h,h'}(\omega)\hat{J}_{h',r'}(\omega)$. The linear response function $\hat{\Gamma}_{h,h'}(\omega)$ admits a series representation in terms of paths through the network that neuron r' can send a signal to neuron r *through hidden neurons only.* This is shown schematically in Fig. 1. The effect of self-coupling from a neuron back to itself can be absorbed into the contribution of each hidden node in a path.

We may write down a set of Feynmanesque rules for explicitly calculating terms in this series. First, we define the gain, $\gamma_h \equiv \lambda_0 \phi' \left(\mu_h + \sum_{h'} \mathcal{J}_{h,h'}\nu_{h'}\right)$. The contribution of each term can then be written down using the following rules: *i*) for the edge connecting recorded neuron r' to a hidden neuron h_i, assign a factor $\hat{J}_{h_i,r'}(\omega)$; *ii*) for each node corresponding to a hidden neuron h_i, assign a factor $\gamma_{h_i}/(1-\gamma_{h_i}\hat{J}_{h_i,h_i}(\omega))$; *iii*) for each edge connecting hidden neurons $h_i \neq h_j$, assign a factor $\hat{J}_{h_j,h_i}(\omega)$; and *iv*) for the edge connecting hidden neuron h_j to recorded neuron r, assign a factor $\hat{J}_{r,h_j}(\omega)$. All factors for each path are multiplied together, and all paths are then summed over.

In practice, the linear response matrix $\hat{\Gamma}_{h,h'}(\omega)$ can be calculated directly by numerical matrix inversion and multiplied with $\hat{J}_{r,h}(\omega)$ and $\hat{J}_{h',r'}(\omega)$ to form the modification to $\hat{J}^{\text{eff}}_{r,r'}(\omega)$, then inverse Fourier transformed to return to the time domain. However, the utility of the path-length series is the intuitive understanding of the origin of contributions to the effective coupling filters and our ability to analytically analyze the strength of contributions from each path [48]. This is reminiscent of recent works expanding correlations functions of linear models of network spiking in terms of network "motifs" [49–51]. One immediate insight the path decomposition offers is that unconnected neurons only develop effective interaction between one other if there is a path that one neuron can send a signal to the other.

Strongly coupled networks—We can now investigate under what conditions hidden paths significantly skew measured neural interactions. The synaptic weights $\mathcal{J}_{i,j}$ (for $i \neq j$) generally must scale with the size of the network in order for network activity to be stable. The more inputs a neuron receives, the weaker the incoming synaptic connection weights must be so as to not drive the neuron to constantly fire. While a glance at Eq. (1) suggests we might expect that $\mathcal{O}(N)$ synaptic inputs to a neuron demands $\mathcal{J}_{i,j} \sim 1/N$ for the inputs to a neuron to be $\mathcal{O}(1)$, this will in fact be too weak. A balance of positive and negative independent couplings reduces the typical magnitude, requiring only $\mathcal{J}_{i,j} \sim \mathcal{O}(1/\sqrt{N})$ for network stability. Hence, we term these two cases "weak" coupling ($\mathcal{J}_{i,j} \sim \mathcal{O}(1/N)$) and "strong" coupling ($\mathcal{J}_{i,j} \sim \mathcal{O}(1/\sqrt{N})$). Previous work has studied the hidden-neuron problem in the weak coupling limit [52–55]; here we study the strong coupling limit. The $1/\sqrt{N}$ scaling of the strong coupling limit has been theoretically predicted to be important for network computations in balanced networks [32–35]. Moreover, this scaling has recently been observed experimentally in cultured neural tissue [36], indicating that it may have intrinsic physiological importance.

We consider the case, ubiquitous in neural modeling, of an Erdős-Réyni (ER) network of N neurons, with connection sparsity p (only $100p\%$ of connections are non-zero). The baselines of the neurons are taken to be equal for all neurons, $\mu_i = \mu_0$, for which there will exist a time-independent steady state. We choose an exponential nonlinearity, $\phi(x) = e^x$. This is the "canonical" choice of nonlinearity used in applications of this model [16–18, 21, 56]. We will further assume $\exp(\mu_0) \ll 1$, so that we may use this as a small control parameter. For $i \neq j$, the non-zero synaptic weights between neurons $\mathcal{J}_{i,j}$ independently drawn from a normal distribution with zero mean and standard deviation $J_0/(pN)^{2a}$, where J_0 controls the overall strength of the weights and $a = 1$ or $1/2$, corresponding to "weak" and "strong" coupling, respectively, for this choice of synaptic weight statistics. For simplicity we take $\mathcal{J}_{i,i} = 0$, for all neurons i. To simplify the analytic analysis, we do not impose Dale's Principle, which requires a neuron's outgoing connections to all be the same sign. We have checked the case of Dale's Principle numerically, and find similar results to the ER network [56]. Numerical values of all parameters are given in Table SI in the SM.

Consider an example case of a network of $N = 1000$ neurons and $N_{\text{rec}} = 3$ recorded neurons. The resulting effective coupling filters (solid blue curves) for these three neurons are plotted against their true ground-truth coupling filters in. Although the true network architecture is sparse, the resulting effective couplings are not sparse—a major qualitative difference in the apparent network architectures. While for many neuron pairs corrections to true non-zero filters were slight, for others, the effective coupling filters deviate significantly away from the true ones, indicating that contributions from paths through the hidden network are comparable to the direct connections.

This result demonstrates that the effective coupling filters can deviate significantly from the direct coupling, but what is the typical deviation? To assess this, we study the statistics of the integrated coupling strengths $\mathcal{J}^{\text{eff}}_{r,r'} \equiv \int_0^\infty dt\, J^{\text{eff}}_{r,r'}(t)$. The expected value of $\mathcal{J}^{\text{eff}}_{r,r'}$ is zero, so we compute the standard deviation of $\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}$, labeled $\sigma[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}]$. Because the self-coupling weights are $\mathcal{J}_{r,r} = 0$, we consider only the standard deviations for $r \neq r'$ pairs. When the hidden network is large, $\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}$ becomes increasingly Gaussian, and

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

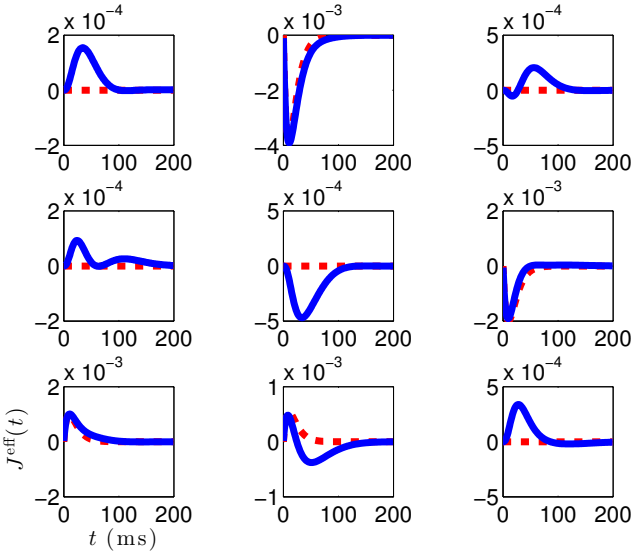

FIG. 2. Effective coupling filters $J^{\rm eff}_{r,r'}(t)$ (solid blue) versus true coupling filters (dashed red) for $N_{\rm rec} = 3$ recorded neurons in a network of $N = 1000$ total neurons. The simulated network has an Erdős-Réyni connectivity with sparsity $p = 0.2$ and normally distributed non-zero weights with zero mean and standard deviation $J_0/\sqrt{pN}$ (strong coupling). All self-couplings are set to zero (indicated by all true coupling filters being zero in plots along grid diagonal).

$\sigma[\mathcal{J}^{\rm eff}_{r,r'} - \mathcal{J}_{r,r'}]$ represents the width of this distribution. We first estimate these standard deviations by numerically computing $\mathcal{J}^{\rm eff}_{r,r'} - \mathcal{J}_{r,r'} = \sum_{h,h'} \mathcal{J}_{r,h}\hat{\Gamma}_{h,h'}(0)\mathcal{J}_{h',r'}$ as we scale up the synaptic amplitude J_0 and observe progressively smaller fractions of neurons f.

The numerical results are shown as solid curves in Fig. (3), for both strong coupling and weak coupling. There are two striking results. First, deviations are nearly negligible ($\mathcal{O}(1/\sqrt{pN})$) for $1/N$ scaling of connections. Thus, for large Erdős-Réyni networks with synapses that scale with the system size, vast numbers of hidden neurons combine to have negligible effect on effective couplings. This is in marked contrast to the case when coupling is strong ($1/\sqrt{N}$ scaling), when hidden neurons have a pronounced impact ($\mathcal{O}(1)$). This is particularly the case when $f \ll 1$—as in almost every experiment in neuroscience, where the hidden neurons outnumber observed ones by orders of magnitude—or when $J_0 \lesssim 1.0$, when typical deviations become half the magnitude of the true couplings themselves (upper blue line). For $J_0 \gtrsim 1.0$, the network activity is unstable for an exponential nonlinearity.

To gain insight into these numerical results, we use our path series expansion to calculate the standard deviation $\sigma[\mathcal{J}^{\rm eff}_{r,r'} - \mathcal{J}_{r,r'}]$, normalized by $\sigma[\mathcal{J}_{r,r'}]$, up to contributions from paths up to length-3. We find

$$\frac{\sigma[\mathcal{J}^{\rm eff}_{r,r'} - \mathcal{J}_{r,r'}]}{\sigma[\mathcal{J}_{r,r'}]} \approx \lambda_0 J_0 e^{\mu_0}\sqrt{1-f} \times \left(1 + \frac{3}{2}(\lambda_0 J_0 e^{\mu_0})^2(1-f)\right), \quad (3)$$

corresponding to the black dashed curves in Fig. 3. Eq. (3) is a truncation of a series in powers of $\lambda_0 J_0 e^{\mu_0}\sqrt{1-f}$, where $f = N_{\rm rec}/N$ is the fraction of recorded neurons. The most important feature of this series is the fact that it only depends on the *fraction* of recorded neurons f, not the absolute number, N. Contributions from long paths remain finite, even as $N \to \infty$. Importantly, this is not the case for weak $1/N$ coupling, in which the series is in powers of $\lambda_0 J_0 e^{\mu_0}\sqrt{(1-f)/(pN)}$, so that contributions from long paths are negligible in large networks $N \gg 1$. (See [56] for derivation and results for $N = 100$.) Deviations of Eq. (3) from the numerical solutions in Fig. 3 indicate that contributions from truncated terms are not negligible when $f \ll 1$. As these terms correspond to paths of length-4 or more, this shows that long chains through the network contribute significantly to shaping effective interactions.

Discussion — We have derived a quantitative relationship between "ground-truth" synaptic interactions and the effective interactions that unobserved neurons generate between subsets of observed neurons. This provides the tools to determine the conditions under which unobserved neurons substantially reshape observed neural dynamics. For the commonly assumed Erdős-Réyni case with independent synaptic weights scaling with $N^{-1/2}$ such hidden units yield qualitative changes to the magnitude and dynamics neural interactions. Together with theoretical and experimental evidence for this scaling in cortex [32–36], this suggests that neural coupling filters inferred from cortical activity data may differ markedly from the true connectivity.

Our work offers a way to probe the architecture of the unobserved network through measured $J^{\rm eff}_{r,r'}(t)$. While inverting Eq. (2) to obtain a unique set of true couplings is generally an ill-posed problem, future work could use the path length expansion combined with physiological constraints to make inferences about the patterns of true connectivities consistent with the measured effective ones. Intriguingly, this same approach may yield insight into the set of circuits that can perform specific computations that are expressed in terms of effective interactions among a subset of cells. This can be particularly useful when the interactions required cor a computation violate Dale's principle or have architectures inconsistent with anatomical measurements [57, 58]. If these regarded as effective interactions, however, then we may be able to use Eq. (2) to generate patterns of biologically plausible "hidden" circuit architectures consistent with the effec-

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

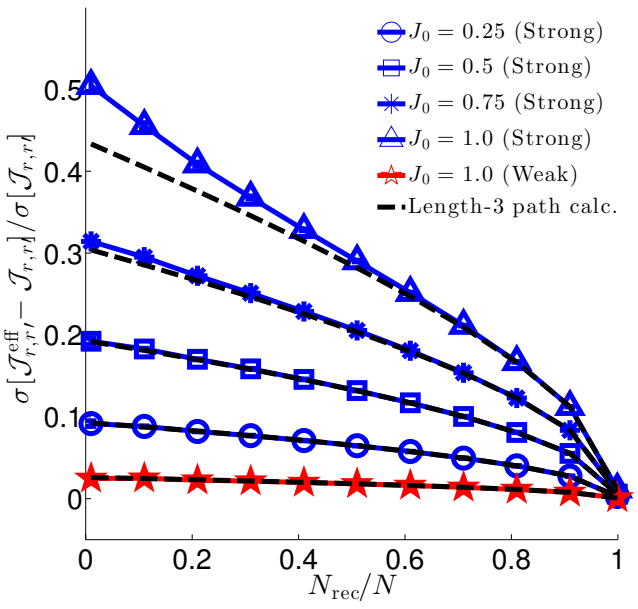

FIG. 3. Solid lines: Numerical estimates of the standard deviation of the difference between effective coupling weights $\mathcal{J}^{\rm eff}_{r,r'}$ and true coupling weights $\mathcal{J}_{r,r'}$ between neurons $r \neq r'$, normalized by the standard deviation of $\mathcal{J}_{r,r'}$. These estimates account for all paths through hidden neurons. For weak $1/N$ coupling (red), the ratio of standard deviations is $\mathcal{O}(1/\sqrt{N})$. For strong $1/\sqrt{N}$ coupling (blue) the ratio is $\mathcal{O}(1)$ and grows in strength as the fraction of recorded neurons $N_{\rm rec}/N$ decreases or the typical synaptic strength J_0 increases. Dashed black lines show theoretical estimates of the relative standard deviations accounting only for hidden paths of length-3 connecting recorded neurons. Deviations from the length-3 prediction at small f and large J_0 indicate that contributions from circuit paths involving many hidden neurons are significant in these regimes. Parameter values are given in Table SI in the Supplemental Material (SM).

tive interactions of the designed circuit.

Similar strategies may even be employed by the nervous system itself. For instance, why do many principal neurons—those which project from one circuit to another—not make direct reciprocal connections to one another, instead being linked by intermediary neurons? One hypothesis is that direct synaptic interactions $J_{r,r'}(t)$ are limited by physiological constraints, and "hidden" interneurons are necessary to reshape the spatiotemporal dynamics of neural interactions into forms that can perform required computations, such as by mixing excitatory and inhibitory interactions to generate a richer range of possible synaptic calculations (cf. Fig. 1). The explicit link we draw between effective interactions and their origins in biological circuitry is thus a new lens for interpreting emerging data on neural connectivity.

Acknowledgments—We thank Gabe Ocker for providing the GLM network simulation code that we modified to perform the full network simulations in this work, Tyler Kekona for work on an early version of a related project, and Ben Lansdell and Christof Koch for helpful feedback. Support provided by the Sackler Scholar Program in Integrative Biophysics (BAWB), CRCNS grant DMS-1208027 (ESB, FR), NIH grant EY11850 and HHMI (FR). This work was partially based on work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 (MAB). MAB and ESB wish to thank the Allen Institute for Brain Science founders, Paul G. Allen and Jody Allen, for their vision, encouragement, and support.

[1] D. S. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg, Journal of Neuroscience **28**, 9239 (2008), http://www.jneurosci.org/content/28/37/9239.full.pdf.
[2] M. A. Kramer, E. D. Kolaczyk, and H. E. Kirsch, Epilepsy Research **79**, 173 (2008).
[3] K. Supekar, V. Menon, D. Rubin, M. Musen, and M. D. Greicius, PLOS Computational Biology **4**, 1 (2008).
[4] C.-Y. Lo, P.-N. Wang, K.-H. Chou, J. Wang, Y. He, and C.-P. Lin, Journal of Neuroscience **30**, 16876 (2010), http://www.jneurosci.org/content/30/50/16876.full.pdf.
[5] M. Chavez, M. Valencia, V. Navarro, V. Latora, and J. Martinerie, Phys. Rev. Lett. **104**, 118701 (2010).
[6] L. Douw, M. de Groot, E. van Dellen, J. J. Heimans, H. E. Ronner, C. J. Stam, and J. C. Reijneveld, PLOS ONE **5**, 1 (2010).
[7] E. van Diessen, J. I. Hanemaaijer, W. M. Otte, R. Zelmann, J. Jacobs, F. E. Jansen, F. Dubeau, C. J. Stam, J. Gotman, and M. Zijlmans, NeuroImage **82**, 564 (2013).
[8] Y. D. Reijmer, A. Leemans, K. Caeyenberghs, S. M. Heringa, H. L. Koek, G. J. Biessels, and O. behalf of the Utrecht Vascular Cognitive Impairment Study Group, Neurology **80**, 1370 (2013), http://www.neurology.org/content/80/15/1370.full.pdf+html.
[9] E. H. Seo, D. Y. Lee, J.-M. Lee, J.-S. Park, B. K. Sohn, D. S. Lee, Y. M. Choe, and J. I. Woo, PLOS ONE **8**, 1 (2013).
[10] C. J. Stam, Nat Rev Neurosci **15**, 683 (2014).
[11] D. E. Warren, J. D. Power, J. Bruss, N. L. Denburg, E. J. Waldron, H. Sun, S. E. Petersen, and D. Tranel, Proceedings of the National Academy of Sciences **111**, 14247 (2014), http://www.pnas.org/content/111/39/14247.full.pdf.
[12] K. T. E. Olde Dubbelink, A. Hillebrand, D. Stoffers, J. B. Deijen, J. W. R. Twisk, C. J. Stam, and H. W. Berendse, Brain **137**, 197 (2014), http://brain.oxfordjournals.org/content/137/1/197.full.pdf.
[13] B. C. Bernhardt, L. Bonilha, and D. W. Gross, Epilepsy & Behavior **50**, 162 (2015).
[14] J. D. Medaglia and D. S. Bassett, ArXiv e-prints (2017), arXiv:1701.01101 [q-bio.NC].
[15] C. J. Honey, J.-P. Thivierge, and O. Sporns, NeuroImage **52**, 766 (2010), computational Models of the Brain.
[16] E. Simoncelli, L. Paninski, J. Pillow, and O. Schwartz, in *The Cognitive Neurosciences*, edited by M. Gazzaniga (MIT Press, 2004) 3rd ed., pp. 327–338.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[17] L. Paninski, Network: Computation in Neural Systems **15**, 243 (2004).

[18] J. Pillow, L. Paninski, V. J. Uzzell, E. Simoncelli, and E. J. Chichilnisky, The Journal of neuroscience : the official journal of the Society for Neuroscience **25**, 11003 (2005).

[19] J. E. Kulkarni and L. Paninski, Network: Computation in Neural Systems **18**, 375 (2007), pMID: 17943613, http://dx.doi.org/10.1080/09548980701625173.

[20] J. W. Pillow and P. E. Latham, in *Advances in Neural Information Processing Systems 20*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (MIT Press, Cambridge, MA, 2007) pp. 1161–1168.

[21] J. Pillow, J. Shlens, L. Paninski, A. Sher, A. Litke, E. Chichilnisky, and E. Simoncelli, Nature **454**, 995 (2008).

[22] G. Field, J. L. Gauthier, A. Sher, M. Greschner, T. a. Machado, L. H. Jepson, J. Shlens, D. E. Gunning, K. Mathieson, W. Dabrowski, L. Paninski, A. Litke, and E. J. Chichilnisky, Nature **467**, 673 (2010).

[23] M. Vidne, Y. Ahmadian, J. Shlens, J. Pillow, J. Kulkarni, A. Litke, E. Chichilnisky, E. Simoncelli, and L. Paninski, J Comput Neurosci **33**, 97 (2012).

[24] L. Paninski, **6536** (2015), 10.1088/0954-898X.

[25] R. Dahlhaus, Metrika **51**, 157 (2000).

[26] M. Eichler, R. Dahlhaus, and J. Sandkühler, Biological Cybernetics **89**, 289 (2003).

[27] I. H. Stevenson, J. M. Rebesco, L. E. Miller, and K. P. Körding, Current Opinion in Neurobiology **18**, 582 (2008).

[28] I. H. Stevenson, B. M. London, E. R. Oby, N. A. Sachs, J. Reimer, B. Englitz, S. V. David, S. A. Shamma, T. J. Blanche, K. Mizuseki, A. Zandvakili, N. G. Hatsopoulos, L. E. Miller, and K. P. Körding, PLOS Computational Biology **8**, 1 (2012).

[29] R. Liégeois, B. Mishra, M. Zorzi, and R. Sepulchre, in *2015 54th IEEE Conference on Decision and Control (CDC)* (2015) pp. 3965–3970.

[30] J. Peters, P. Bühlmann, and N. Meinshausen, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78**, 947 (2016).

[31] N. J. Foti, R. Nadkarni, A. K. Lee, and E. B. Fox, in *2nd KDD Workshop on Mining and Learning from Time Series* (2016).

[32] C. van Vreeswijk and H. Sompolinsky, Science **274**, 1724 (1996), http://science.sciencemag.org/content/274/5293/1724.full.pdf.

[33] A. Litwin-Kumar and B. Doiron, Nat Neurosci **15**, 1498 (2012).

[34] R. Rosenbaum and B. Doiron, Phys. Rev. X **4**, 021039 (2014).

[35] S. Deneve and C. K. Machens, Nat Neurosci **19**, 375 (2016).

[36] J. Barral and A. D Reyes, Nat Neurosci **19**, 1690 (2016).

[37] G. Koch Ocker, K. Josić, E. Shea-Brown, and M. A. Buice, ArXiv e-prints (2016), arXiv:1610.03828 [q-bio.NC].

[38] E. S. Chornoboy, L. P. Schramm, and A. F. Karr, Biological Cybernetics **59**, 265 (1988).

[39] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From single neurons to networks and models of cognition* (Cambridge University Press, Cambridge, U.K., 2014).

[40] S. Ostojic and N. Brunel, PLOS Computational Biology **7**, 1 (2011).

[41] B. Dunn and Y. Roudi, Phys. Rev. E **87**, 022127 (2013).

[42] J. Tyrcha and J. Hertz, Mathematical Biosciences and Engineering **11**, 149 (2014).

[43] B. Dunn and C. Battistin, ArXiv e-prints (2016), arXiv:1612.06185 [cond-mat.dis-nn].

[44] B. Bravi and P. Sollich, ArXiv e-prints (2016), arXiv:1611.09849 [physics.chem-ph].

[45] B. Bravi and P. Sollich, ArXiv e-prints (2016), arXiv:1612.01976 [cond-mat.dis-nn].

[46] B. Bravi, M. Opper, and P. Sollich, Phys. Rev. E **95**, 012122 (2017).

[47] This differs from the linear response function to perturbations in the *output*, $\hat{\Delta}_{h,h'}(\omega)$. However, the two are related by $\hat{\Gamma}_{h,h'}(\omega) = \hat{\Delta}_{h,h'}(\omega)\gamma_{h'}$, where $\gamma_h = \lambda_0\phi'(\mu_h + \sum_{h'} \mathcal{J}_{h,h'}\nu_{h'}(t'))$ is the gain.

[48] The series expansion is also more practical in the general case in which the baselines $\mu_h(t)$ are time-dependent and we cannot solve the problem in the frequency domain.

[49] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter, PLOS Computational Biology **7**, 1 (2011).

[50] Y. Hu, J. Trousdale, K. Josi, and E. Shea-Brown, Journal of Statistical Mechanics: Theory and Experiment **2013**, P03012 (2013).

[51] Y. Hu, J. Trousdale, K. Josić, and E. Shea-Brown, Phys. Rev. E **89**, 032802 (2014).

[52] D. Q. Nykamp, SIAM Journal on Applied Mathematics **65**, 2005 (2005), http://dx.doi.org/10.1137/S0036139903437072.

[53] D. Q. Nykamp, Mathematical Biosciences **205**, 204 (2007).

[54] D. Q. Nykamp, SIAM Journal on Applied Mathematics **68**, 354 (2007), http://dx.doi.org/10.1137/070683350.

[55] D. Q. Nykamp, Phys. Rev. E **78**, 021902 (2008).

[56] *See Supplemental Material.*

[57] P. D. King, J. Zylberberg, and M. R. DeWeese, Journal of Neuroscience **33**, 5475 (2013), http://www.jneurosci.org/content/33/13/5475.full.pdf.

[58] H. F. Song, G. R. Yang, and X.-J. Wang, PLOS Computational Biology **12**, 1 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SUPPLEMENTAL MATERIAL

A. Model definition: nonlinear Hawkes model (Generalized linear point process model)

The firing rate of a neuron i in the full network is given by

$$\lambda_i(t) = \lambda_0 \phi\left(\mu_i + \sum_j \int_{-\infty}^{\infty} dt' J_{ij}(t-t')\dot{n}_j(t') + \mathcal{E}_i(t)\right), \tag{S.1}$$

where λ_0 is a characteristic rate, $\phi(x) \geq 0$ is a nonlinear function, μ_i (potentially a function of some external stimulus θ) is a time-independent tonic drive that sets the baseline firing rate of the neuron in the absence of input from other neurons, $J_{ij}(t-t')$ is a coupling filter that filters spikes $\dot{n}_j(t')$ fired by presynaptic neuron j at time t', incident on post-synaptic neuron i, and $\mathcal{E}_i(t)$ is an external input current. We will take $\mathcal{E}_i(t) = 0$ for simplicity in this work, focusing on the spontaneous activity of the network. We need not attach a mechanistic interpretation to these filters, but a convenient interpretation is that the GLM model is like a soft-threshold integrate-and-fire network model, such that we can interpret the spike filtering as coming from some synaptic dynamics [1, 2]. As such, we take the actual rate to be

$$\lambda_i(t) = \lambda_0 \phi\left(\mu_i + \sum_j \mathcal{J}_{ij} s_j(t)\right),$$

where $s_j(t)$ is the synaptic activity of pre-synaptic neuron j at time t. In general,

$$s_j(t) = \int_{-\infty}^{t} dt'\ g_j(t-t')\dot{n}_j(t'),$$

where $g_j(t)$ is the temporal waveform of the post-synaptic currents evoked by pre-synaptic neuron j, normalized to integrate to 1.

This interpretation of the model is useful for simulating the network activity, and we use such a method for our simulations verifying our analytic calculations (see "Simulations of network activity"). Otherwise, we will simply use the form of the model as written in Eq. (S.1).

B. Model network architectures

Our main result, Eq. (2), is valid for general network architectures with arbitrary weighted synaptic connections, so long as the hidden subset of the network has stable dynamics when the recorded neurons are removed. An example for which our method must be modified would be a network in which all or the majority of the hidden neurons are excitatory, as the hidden network is unlikely to be stable when the recorded neurons are disconnected. Similarly, we find that synaptic weight distributions with undefined moments will generally cause the network activity to be unstable. For example, $\mathcal{J}_{i,j}$ drawn from a Cauchy distribution generally yield unstable network dynamics unless the weights are scaled inversely with a large power of the network size N.

To specify a concrete network architecture to study, we choose a directed Erdős-Réyni random network with sparsity p—i.e., each directed connection between neurons is assigned independently with probability p that the weight is non-zero. The weights of the non-zero connections are then drawn from a separate distribution chosen to have mean 0 and variance $J_0^2/(pN)^{2a}$. The choice of exponent a determines whether the coupling is weak ($a = 1$) or strong ($a = 1/2$). In most of our analytical results we only need the mean and variances of the weights, so we do not need to specify the exact distribution. In simulations, we use a normal distribution, but have also checked (symmetrized) lognormal distributions, which yield similar results (not shown). The reason for scaling the weights as $1/(pN)^a$, as opposed to just $1/N^a$, is that the mean incoming degree of connections is $p(N-1) \approx pN$ for large networks; this scaling thus controls for the typical magnitude of incoming spikes.

For strongly coupled networks, the combined effect of sparsity and synaptic weight distribution yields an overall standard deviation of $\sqrt{p}J_0/\sqrt{pN} = J_0/\sqrt{N}$. Because the sparsity parameter p cancels out, it does not matter if we consider p to be fixed or $k_0 = pN$ to be fixed—both cases are equivalent. However, this is not the case if we scale $\mathcal{J}_{i,j}$ by $1/k_0$, as the overall standard deviation would then be $\sqrt{p}J_0/k_0$, which only corresponds to the weak-coupling limit if p is fixed. If k_0 is fixed, the standard deviation would scale as $1/\sqrt{N}$.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

It is worth noting that the determination of "weak" versus "strong" coupling depends not only on the power of N with which synaptic weights scale, but also on the network architecture and correlation structure of the weights $\mathcal{J}_{i,j}$. For example, for an all-to-all connected matrix with symmetric rank-1 synaptic weights of the form $\mathcal{J}_{i,j} = \zeta_i \zeta_j$, where the ζ_i are independently distributed normal random variates, the standard deviation of *each* ζ must scale as $1/\sqrt{N}$ in order for hidden paths to generate $\mathcal{O}(1)$ contributions to effective interactions, such that $\mathcal{J}_{i,j}$ scales as $1/N$ but the coupling is still strong.

Real neurons appear to split into two broad separate classes, "excitatory" and "inhibitory," a dichotomy know as Dale's principle. Neurons in a network that obeys this principle will have coupling filters $J_{i,j}(t)$ that are strictly positive for excitatory neurons and strictly negative for inhibitory neurons. We do not impose this restriction on the model used to generate the results presented in the main text, but have tested the consequence of imposing Dale's principle on the network. We find similar results, shown in Fig. (S1) below (compare to Fig. (3) in main text). The trends are the same as in networks that do not obey Dale's principle, with the resulting ratios being slightly reduced. Because our analysis requires calculation of the mean field firing rates of the hidden network in absence of the recorded neurons, random sampling of the network may, by chance, yield hidden networks with an imbalance of excitatory neurons, for which the mean field firing rates of the hidden network may diverge for our choice of exponential nonlinearity. This is the origin of the relatively larger error bars in Fig. (S1): less random samplings for which the hidden network was stable were available to perform the computation. This is merely an artifact that arises due to the choice of reference state of network activity around which we perform the expansion discussed in Sec. D. This artifact can be cured either by choosing a nonlinearity that saturates, such as $\phi(x) = c/(1 + \exp(-x))$, which prevents the mean-field firing rates from diverging and yields stable network activity (see Fig. S4 in Sec. C), or by choosing a different reference state, as discussed in Sec. L.

Finally, although Erdős-Réyni networks are relatively easy to analyze analytically, and are ubiquitous in many influential computational and theoretical studies, real world networks typically have more structure. As preliminary tests of our results in networks with more realistic features, we have also simulated networks with Watts-Strogatz network architectures. A Watts-Strogatz network is generated by starting with a K-nearest neighbor network (such that fraction of non-zero connections each neuron makes is $p = K/(N-1)$) and rewiring a fraction β of those connections. The limit $\beta = 0$ remains a K-nearest neighbor network, while $\beta \to 1$ yields an Erdős-Réyni network. We have simulated networks on Watts-Strogatz architectures, finding similar results as for the Erdős-Réyni network, as seen in Fig. (S2). We generated the adjacency matrices of the Watts-Strogatz networks using code available in [3].

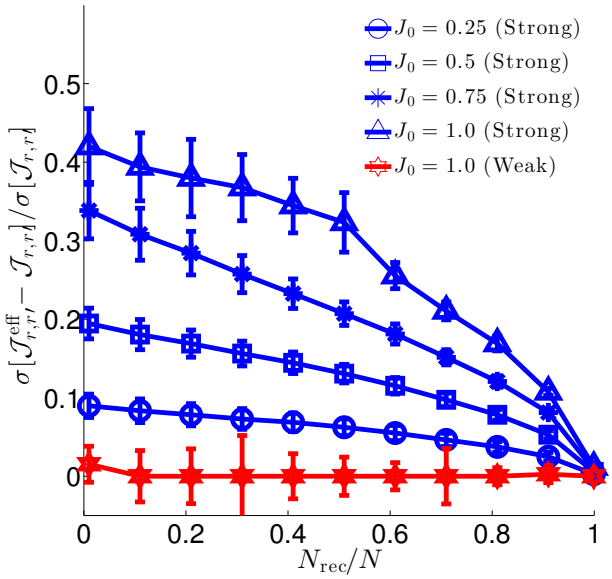

FIG. S1. Same as Fig. (3) in main text, but for a Erdős-Réyni network architecture with Dale's principle imposed; i.e., all connections neuron j makes are positive if the neuron is excitatory ($\mathcal{J}_{i,j} > 0$) or negative if the neuron is inhibitory ($\mathcal{J}_{i,j} < 0$). The fraction of excitatory to inhibitory neurons is chosen to be 50% on average. The resulting ratios are smaller than the corresponding ratios for Erdős-Réyni networks in which we do not impose Dale's principle, but the trends otherwise hold. The greater uncertainties on estimates are because random sampling of the full network may generate hidden networks with an unstable mean field theory (see main text for explanation). Theoretical estimates of the relative standard deviations were not calculated. Parameter values are given in Table SI.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

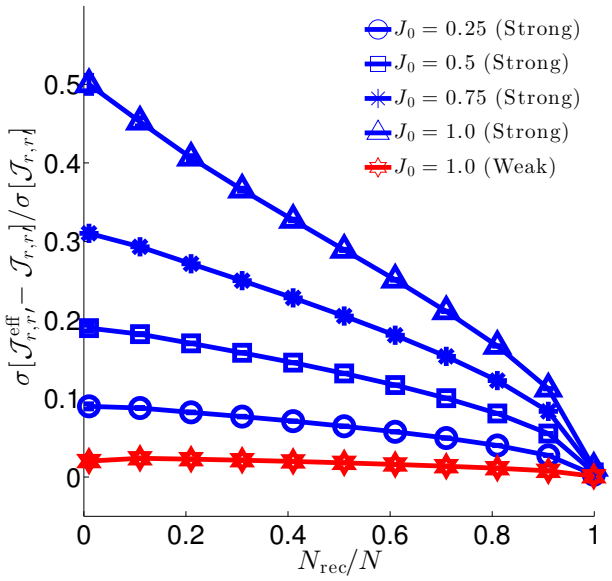

FIG. S2. Same as Fig. (3) in main text (Erdős-Réyni network, Dale's principle not imposed), but for a Watts-Strogatz network architecture. The resulting curves are only slightly smaller than for an Erdős-Réyni network. Theoretical estimates of the relative standard deviations were not calculated. Parameter values are given in Table SI.

Parameter values used to generate our networks are given in Table SI.

TABLE SI. Network connectivity parameter values.

Number of neurons N	1000
Number of hidden neurons N_{hid}	See individual figure captions.
Number of recorded neurons N_{rec}	$N - N_{\text{hid}}$
Baselines μ_i	-1.0, $\forall i$
Sparsity p	0.2
Coupling weights $\mathcal{J}_{ij}$	$\mathcal{N}(0, J_0^2/(pN)^{2a})$
Coupling regime a	1 (weak coupling) or 1/2 (strong coupling)
Rewiring probability β (Watts-Strogatz only)	0.3
Characteristic synaptic weight J_0	0.25, 0.5, 0.75, 1.0
Firing frequency λ_0	1.0

C. Choice of nonlinearity $\phi(x)$

The nonlinear function $\phi(x)$ sets the instantaneous firing rate for the neurons in our model. The "canonical" choice of nonlinearity for our network model is an exponential, $\phi(x) = \exp(x)$ [4–7]. The exponential has particularly nice theoretical properties, but is also convenient for fitting this model to data, as the log-likelihood function of the model will be convex for the exponential (and some similar families of nonlinearity).

The fact that the exponential is unbounded is necessary to *guarantee* that a neuron spikes given enough input. A bounded nonlinearity imposes a maximum instantaneous firing rate, such that it is possible that the instantaneous rate saturates but does not guarantee the neuron will spike. The downside of an unbounded nonlinearity is that it is possible for the average firing rates to diverge, and the network never reaches a steady state. For example, in a purely excitatory network (all $\mathcal{J}_{i,j} \geq 0$) with an exponential nonlinearity, neural firing will run away without a sufficiently strong self-refractory coupling to suppress the firing rate. This will not occur with a bounded nonlinearity, as excitation can only drive neurons to fire at some maximum but finite rate.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This is a problem in simulations of networks obeying Dale's principle. For the exponential nonlinearity, the mean field theory for the hidden network occasionally does not exist due to an imbalance of excitatory and inhibitory neurons caused by our random recorded of neurons. However, the Dale's law network is stable if the nonlinearity is bounded. We demonstrate this below in Figs. S3 and S4, comparing simulations of Erdős-Réyni networks with and without Dale's principle for a sigmoidal nonlinearity $\phi(x) = 2/(1 + e^{-x})$.

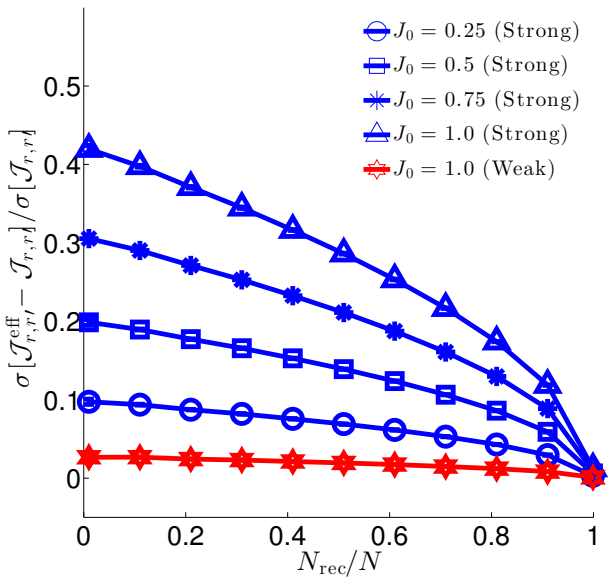

FIG. S3. Same as Fig. (3) in main text, but for a sigmoidal nonlinearity $\phi(x) = 2/(1 + e^{-x})$.

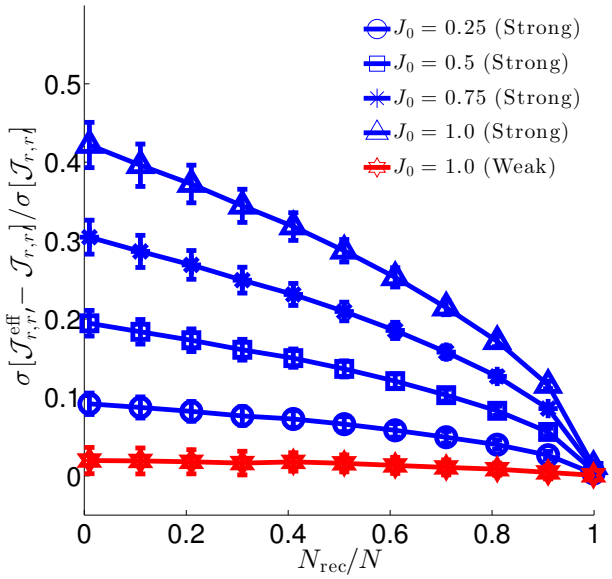

FIG. S4. Same as Fig. (S1) above, but for a sigmoidal nonlinearity $\phi(x) = 2/(1 + e^{-x})$. Because the sigmoid is bounded the mean field solution cannot diverge, yielding better results.

D. Derivation of effective baselines and coupling filters

To study how hidden neurons affect the inferred properties of recorded neurons, we partition the network into "recorded" neurons, labeled by indices r (with sub- or superscripts to differentiate different recorded neurons, e.g., r

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and r' or r_1 and r_2) and "hidden" neurons labeled by indices h (with sub- or superscripts). The rates of these two groups are thus

$$\lambda_r(t) = \lambda_0 \exp\left(\mu_r + \sum_{r'} J_{r,r'} * \dot{n}_{r'} + \sum_h J_{r,h} * \dot{n}_h\right),$$

$$\lambda_h(t) = \lambda_0 \exp\left(\mu_h + \sum_r J_{h,r} * \dot{n}_r + \sum_{h'} J_{h,h'} * \dot{n}_{h'}\right).$$

Because the set hidden neurons is generally much larger than the set of recorded neurons, we expect that we can approximate the input to the recorded neurons with the mean input, conditioned on the activity of the recorded neurons. That is, we can split the hidden input to the recorded neurons up into two terms, the mean plus fluctuations around the mean:

$$\sum_h J_{r,h} * \dot{n}_h(t) = \sum_h J_{r,h} * \mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r\}\right] + \xi_r(t),$$

where $\mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r\}\right]$ denotes the mean activity of the hidden neurons conditioned on the activity of the recorded units, and $\xi_r(t)$ are fluctuations around this mean, i.e., $\xi_r(t) \equiv \sum_h J_{r,h} * (\dot{n}_h - \mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r\}\right])$. Note that $\xi_r(t)$ is also conditional on the activity of the recorded units.

We can calculate the cross-correlation of the fluctuations,

$$\mathbb{E}\left[\xi_r(t)\xi_{r'}(t')\right] = \int_{-\infty}^{\infty} dt_1 dt_2 \sum_{h_1,h_2} J_{r,h_1}(t-t_1) J_{r',h_2}(t'-t_2) C_{h_1,h_2}(t_1,t_2),$$

where $C_{h_1,h_2}(t_1,t_2)$ is the cross-correlation between hidden neurons h_1 and h_2. If the autocorrelation of the fluctuations ($r = r'$) is small compared to the mean input to the recorded neurons ($\sum_h J_{r,h} * \mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r\}\right]$), or if $J_{r,h}$ is small, then we may neglect these fluctuations and focus only on the effects that the mean input has on the recorded subnetwork. At the level of the mean field theory approximation we make in this work, the spike-train correlations are zero. One can calculate corrections to mean field theory (see below), but we leave investigation of the properties of these fluctuations for the focus of future studies. This will not affect our overall analysis, as non-negligible noise can be treated as a separate input to the recorded neurons, and hence will not alter the form of the effective couplings between neurons. Averaging out the effective noise, however, will generate new interactions between neurons; we leave investigation of this issue for future work.

In order to calculate how hidden input shapes the activity of recorded neurons, we need to calculate the mean $\mathbb{E}\left[\dot{n}_h | \{\dot{n}_r\}\right]$. This mean input is difficult to calculate in general, especially when conditioned on the activity of the recorded neurons. In principle, the mean can be calculated as

$$\mathbb{E}\left[\dot{n}_h | \{\dot{n}_r\}\right] = \mathbb{E}\left[\lambda_0 \exp\left(\mu_h + \sum_r J_{h,r} * \dot{n}_r + \sum_{h'} J_{h,h'} * \dot{n}_{h'}\right) \middle| \{\dot{n}_r\}\right].$$

This is not a tractable calculation. Taylor series expanding the exponential reveals that the mean will depend on *all* higher cumulants of the hidden unit spike trains, which cannot in general be calculated explicitly. Instead, we again appeal to the fact that in a large, sufficiently connected network, we expect fluctuations to be small, as long as the network is not near a critical point. In this case, we may make a mean field approximation, which amounts to solving the self-consistent equation

$$\mathbb{E}\left[\dot{n}_h | \{\dot{n}_r\}\right] = \lambda_0 \exp\left(\mu_h + \sum_r J_{h,r} * \dot{n}_r + \sum_{h'} J_{h,h'} * \mathbb{E}\left[\dot{n}_{h'} | \{\dot{n}_r\}\right]\right). \tag{S.2}$$

In general, this equation must be solved numerically. Unfortunately, the conditional dependence on the activity of the recorded neurons presents a problem, as in principle we must solve this equation for *all possible patterns of recorded unit activity*. Instead, we note that the mean hidden neuron firing rate is a *functional* of the filtered recorded input $I_h(t) \equiv \sum_r J_{h,r} * \dot{n}_r$, so we can expand it as a functional Taylor series (sometimes known as a Volterra series) around

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

some reference filtered activity $I_h^0(t) = \sum_r J_{h,r} * \dot{n}_r^0$,

$$\mathbb{E}\left[\dot{n}_h(t)|\left\{I_h(t)\right\}\right] = \mathbb{E}\left[\dot{n}_h|\left\{I_h^0(t)\right\}\right] + \int dt_1 \sum_{h_1} \frac{\delta\mathbb{E}\left[\dot{n}_h(t)|\left\{I_h^0(t)\right\}\right]}{\delta I_{h_1}(t_1)}(I_{h_1}(t_1) - I_{h_1}^0(t_1))$$
$$+ \frac{1}{2}\int dt_1 dt_2 \sum_{h_1,h_2} \frac{\delta^2\mathbb{E}\left[\dot{n}_h(t)|\left\{I_h^0(t)\right\}\right]}{\delta I_{h_2}(t_2)\delta I_{h_1}(t_1)}(I_{h_2}(t_2) - I_{h_2}^0(t_2))(I_{h_1}(t_1) - I_{h_1}^0(t_1)) + \ldots$$

Within our mean field approximation, the Taylor coefficients are simply the response functions of the network — i.e., the zeroth order coefficient is the mean firing rate of the neurons in the reference state $I_h^0(t)$, the first order coefficient is the linear response function of the network, the second order coefficient is a nonlinear response function, and so on.

There are two natural choices for the reference state $I_h^0(t)$. The first is simply the state of zero recorded unit activity, while the second is the mean activity of the recorded neurons. The zero-activity case conforms to the choice of GLM models used in practice. Choosing the mean activity as the reference state may be more appropriate if the recorded neurons have high firing rates, but requires adjusting the form of the GLM model so that firing rates are modulated by filtering deviations of spikes from the mean firing rate, rather than filtering the spikes themselves. In the main text, we focus on the zero-activity reference state. We present the formulation for the mean field reference state at the end of the SI.

For the zero-activity reference state $I_h^0(t) = 0$, the conditional mean is

$$\mathbb{E}\left[\dot{n}_h(t)|\left\{I_h(t)\right\}\right] = \mathbb{E}\left[\dot{n}_h|0\right] + \int dt_1 \sum_{h_1} \frac{\delta\mathbb{E}\left[\dot{n}_h(t)|0\right]}{\delta I_{h_1}(t_1)} I_{h_1}(t_1) + \frac{1}{2}\int dt_1 dt_2 \sum_{h_1,h_2} \frac{\delta^2\mathbb{E}\left[\dot{n}_h(t)|0\right]}{\delta I_{h_2}(t_2)\delta I_{h_1}(t_1)} I_{h_2}(t_2) I_{h_1}(t_1) + \ldots.$$

The mean inputs $\mathbb{E}\left[\dot{n}_h|0\right]$ are the mean field approximations to the firing rates of the hidden neurons in the absence of the recorded neurons. Defining $\nu_h \equiv \mathbb{E}\left[\dot{n}_h|0\right]$, these firing rates are given by

$$\nu_h = \lambda_0 \phi\left(\mu_h + \sum_{h'} \mathcal{J}_{h,h'}\nu_{h'}\right);$$

in writing this equation we have used the fact that the steady-state mean field firing rates will be time-independent, and hence the convolution $J_{h,h'} * \nu_{h'} = \mathcal{J}_{h,h'}\nu_{h'}$, where $\mathcal{J}_{h,h'} = \int_0^\infty dt\, J_{h,h'}(t)$. The mean field equations for the ν_h are a system of transcendental equations that in general cannot be solved exactly. In practice we will solve the equations numerically, but we can develop a series expansion for the solutions (see below).

The next term in the series expansion is the linear response function of the hidden unit network, $\Gamma_{h,h'}(t-t') \equiv \frac{\delta\mathbb{E}[\dot{n}_h(t)|0]}{\delta I_{h'}(t')}$, given by

$$\Gamma_{h,h'}(t-t') = \gamma_h\left(\delta_{h,h'}\delta(t-t') + \sum_{h''}\int_0^\infty dt'' J_{h,h''}(t-t'')\Gamma_{h'',h'}(t''-t')\right).$$

The "gain" γ_h is defined by

$$\gamma_h \equiv \lambda_0 \phi'\left(\mu_h + \sum_{h'} \mathcal{J}_{h,h'}\nu_{h'}\right),$$

where $\phi'(x)$ is the derivative of the nonlinearity with respect to its argument.

We may solve for $\Gamma_{h,h'}(t-t')$ by first converting to the frequency domain and performing a matrix inverse:

$$\hat{\Gamma}_{h,h'}(\omega) = \left[\mathbb{I} - \hat{\mathbf{V}}(\omega)\right]^{-1}_{h,h'} \gamma_{h'},$$

where $V_{h,h'}(\omega) = \gamma_h J_{h,h'}(\omega)$ and $\left[\mathbb{I} - \hat{\mathbf{V}}(\omega)\right]^{-1} \equiv \hat{\mathbf{\Delta}}(\omega)$ is the output-linear response function.

If $||\hat{\mathbf{V}}(\omega)|| < 1$ for some matrix norm $||\cdot||$, then the matrix $\left[\mathbb{I} - \mathbf{V}(\omega)\right]^{-1}$ admits a convergent series expansion [8]

$$\left[\mathbb{I} - \mathbf{V}(\omega)\right]^{-1} = \sum_{\ell=0} \hat{\mathbf{V}}(\omega)^\ell,$$

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where $\hat{\mathbf{V}}(\omega)^\ell$ is a matrix product and $\hat{\mathbf{V}}(\omega)^0 \equiv \mathbb{I}$. We can write an element of the matrix product out as

$$\left[\hat{\mathbf{V}}(\omega)^\ell\right]_{h,h'} = \sum_{h_1,\ldots,h_\ell} \hat{V}_{h,h_1}(\omega)\hat{V}_{h_1,h_2}(\omega)\ldots\hat{V}_{h_{\ell-1},h_\ell}(\omega)\hat{V}_{h_\ell,h'}(\omega);$$

inserting $\hat{V}_{h_i,h_j}(\omega) = \gamma_{h_i}\hat{J}_{h_i,h_j}(\omega)$ yields

$$\left[\hat{\mathbf{V}}(\omega)^\ell\right]_{h,h'} = \sum_{h_1,\ldots,h_\ell} \gamma_h\hat{J}_{h,h_1}(\omega)\gamma_{h_1}\hat{J}_{h_1,h_2}(\omega)\ldots\gamma_{h_{\ell-1}}\hat{J}_{h_{\ell-1},h_\ell}(\omega)\gamma_{h_\ell}\hat{J}_{h_\ell,h'}(\omega).$$

It is this expression that we interpret in terms of summing over paths through network of hidden neurons that join two observed neurons: the $\hat{J}_{h_i,h_j}(\omega)$ are represented by edges from neuron h_j to h_i, and the γ_{h_i} are represented by the nodes. In this expansion, we allow edges from one neuron back to itself, meaning we include paths in which signals loop back around to the same neuron arbitrarily many times before the signal is propagated further. Such loops can be easily factored, contributing a factor $\sum_{m=0}^{\infty}(\gamma_h\hat{J}_{h,h}(\omega))^m = 1/(1-\gamma_h\hat{J}_{h,h}(\omega))$. We can thus remove the need to consider self-loops in our rules for calculating effective coupling filters by assigning a factor $\gamma_h/(1-\gamma_h J_{h,h}(\omega))$ to each node, rather than just γ_h.

This result can be derived directly. Let us decompose the matrix $\hat{\mathbf{V}}(\omega)$ in a diagonal and off-diagonal piece, $\hat{\mathbf{V}}(\omega) = \hat{\mathbf{V}}_{\text{diag}}(\omega) + \hat{\mathbf{V}}_{\text{off}}(\omega)$. Then,

$$\begin{aligned}
\left[\mathbb{I} - \hat{\mathbf{V}}(\omega)\right]^{-1} &= \left[\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega) - \hat{\mathbf{V}}_{\text{off}}(\omega)\right]^{-1} \\
&= \left[\left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)\left(\mathbb{I} - \left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right)\right]^{-1} \\
&= \left[\mathbb{I} - \left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right]^{-1}\left[\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right]^{-1}
\end{aligned}$$

We assumed that $\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)$ is invertible, which requires that there is no element for which $1 - \gamma_h\hat{J}_{h,h}(\omega) = 0$ for all ω. Assuming this is the case, the inverse of the matrix is trivial to calculate, as it is diagonal:

$$\left[\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right]^{-1}_{h,h'} = \frac{1}{1-\gamma_h J_{h,h}(\omega)}\delta_{h,h'}.$$

The matrix $\left[\mathbb{I} - \left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right]^{-1}$ can be expressed as a series, as before:

$$\begin{aligned}
\left[\mathbb{I} - \left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right]^{-1}_{h,h''} &= \sum_{\ell=0}\left[\left[\left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right]^\ell\right]_{h,h''} \\
&= \sum_{\ell=0}\sum_{h_1,\ldots,h_\ell}\left[\left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right]_{h,h_1}\ldots\left[\left(\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right)^{-1}\hat{\mathbf{V}}_{\text{off}}(\omega)\right]_{h_\ell,h''} \\
&= \sum_{\ell=0}\sum_{h_1,\ldots,h_\ell;h_i\neq h_{i+1}}\frac{\gamma_h}{1-\gamma_h\hat{J}_{h,h}(\omega)}\hat{J}_{h,h_1}(\omega)\ldots\frac{\gamma_{h_\ell}}{1-\gamma_{h_\ell}\hat{J}_{h_\ell,h_\ell}(\omega)}J_{h_\ell,h''}(\omega)
\end{aligned}$$

Hence, inserting the contribution from the factor $\left[\mathbb{I} - \hat{\mathbf{V}}_{\text{diag}}(\omega)\right]^{-1}$ that we pulled out, and the factor $\gamma_{h'}$ that left-multiplies $\left[\mathbb{I} - \hat{\mathbf{V}}(\omega)\right]^{-1}$ to give $\hat{\Gamma}_{h,h'}(\omega)$, we have

$$\hat{\Gamma}_{h,h'}(\omega) = \sum_{\ell=0}\sum_{h_1,\ldots,h_\ell;h_i\neq h_{i+1}}\frac{\gamma_h}{1-\gamma_h\hat{J}_{h,h}(\omega)}\hat{J}_{h,h_1}(\omega)\ldots\frac{\gamma_{h_\ell}}{1-\gamma_{h_\ell}\hat{J}_{h_\ell,h_\ell}(\omega)}\hat{J}_{h_\ell,h'}(\omega)\frac{\gamma_{h'}}{1-\gamma_{h'}\hat{J}_{h,h'}(\omega)}$$

This is the same as our previous expression, with $\gamma_h \to \gamma_h/(1-\gamma_h\hat{J}_{h,h}(\omega))$ and restricting the sum over hidden units such that self-loops are removed ($h_i \neq h_{i+1}$), proving the result described in the main text. We note again that this puts restrictions on the allowed size of self-interactions, as the zeros of $1-\gamma_h\hat{J}_{h,h}(\omega)$ must be in the upper-half plane of the complex ω plane in order for the filters to be causal and physically meaningful.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The complete expression for the correction term $\sum_{h,h'} \hat{J}_{r,h}(\omega)\hat{\Gamma}_{h,h'}(\omega)\hat{J}_{h',r'}(\omega)$ is thus

$$\sum_{h,h'} \hat{J}_{r,h}(\omega)\hat{\Gamma}_{h,h'}(\omega)\hat{J}_{h',r'}(\omega) = \sum_{\ell=0} \sum_{h,h_1,\ldots,h_\ell,h';h_i\neq h_{i+1}} \hat{J}_{r,h}(\omega)\frac{\gamma_h}{1-\gamma_h\hat{J}_{h,h}(\omega)}\hat{J}_{h,h_1}(\omega)\ldots\frac{\gamma_{h_\ell}}{1-\gamma_{h_\ell}\hat{J}_{h_\ell,h_\ell}(\omega)}\hat{J}_{h_\ell,h'}(\omega)\frac{\gamma_{h'}}{1-\gamma_{h'}\hat{J}_{h,h'}(\omega)}\hat{J}_{h',r'}(\omega).$$

This is the exact mathematical expression underlying the rules given in the main text.

We have now determined the zero and first order Taylor series coefficients in our expansion of $\mathbb{E}[\dot{n}_h(t)|\{\dot{n}_r\}]$. If these are the dominant terms, i.e., if we may neglect higher order terms in this expansion, we may approximate the instantaneous firing rates of the recorded neurons by

$$\lambda_r(t) \approx \lambda_0\phi\left(\mu_r^{\text{eff}} + \sum_{r'} J_{r,r'}^{\text{eff}} * \dot{n}_{r'}(t)\right),$$

where

$$\mu_r^{\text{eff}} = \mu_r + \sum_h \mathcal{J}_{r,h}\nu_h$$

are the effective baselines of the recorded neurons and

$$\hat{J}_{r,r'}^{\text{eff}}(\omega) = \hat{J}_{r,r'}(\omega) + \sum_{h,h'} \hat{J}_{r,h}(\omega)\hat{\Gamma}_{h,h'}(\omega)\hat{J}_{h',r'}(\omega)$$

are the effective coupling filters in the frequency domain, as given in the main text. In addition to neglecting the higher order spike filtering terms here, we have also neglected fluctuations around the mean input from the hidden network. These fluctuations are zero within our mean field approximation, but we could in principle calculate corrections to the mean field predictions using the results of [9]. Below, we use these methods to calculate the tree-level corrections to the correlation functions of the spikes and estimate the size of these fluctuations.

E. Specific choices of network properties used to generate Figures 2 and 3

To generate the results in Fig. 2 in the main text, we choose the coupling filters to be $J_{i,j}(t) = \mathcal{J}_{i,j}\alpha^2 te^{-\alpha t}$, which has Fourier transform

$$\hat{J}_{i,j}(\omega) = \frac{\mathcal{J}_{i,j}\alpha^2}{(\alpha+i\omega)^2},$$

using the Fourier convention

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} dt\ e^{-i\omega t} f(t).$$

The weight matrix $\mathcal{J}$ is generated as described in "Model network architectures," choosing $J_0 = 1.0$. We partition this network up into recorded and hidden subsets. For a network of N neurons, we choose neurons 1 to N_{rec} to be recorded, and the remainder to be hidden, hence we define (using an index notation starting at 1; indices should be subtracted by 1 for 0-based index counting)

$$\mathcal{J}^{\text{RR}} = \mathcal{J}[1:N_{\text{rec}}, 1:N_{\text{rec}}],$$

$$\mathcal{J}^{\text{RH}} = \mathcal{J}[1:N_{\text{rec}}, (N_{\text{rec}}+1):N],$$

$$\mathcal{J}^{\text{HR}} = \mathcal{J}[(N_{\text{rec}}+1):N, 1:N_{\text{rec}}],$$

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and

$$\mathcal{J}^{\mathrm{HH}} = \mathcal{J}[(N_{\mathrm{rec}}+1):N,(N_{\mathrm{rec}}+1):N].$$

We numerically calculate the linear response matrix $\hat{\mathbf{\Gamma}}(\omega)$ by evaluating

$$\hat{\mathbf{\Gamma}}(\omega) = \left[\mathbb{I} - \hat{\mathbf{V}}^{\mathrm{HH}}(\omega)\right]^{-1} \mathrm{diag}(\vec{\gamma}),$$

where $\hat{V}^{\mathrm{HH}}_{h,h'}(\omega) = \gamma_h \mathcal{J}_{h,h'}(\omega)$ and $\mathrm{diag}(\vec{\gamma})$ is an $N_{\mathrm{hid}} \times N_{\mathrm{hid}}$ diagonal matrix with elements γ_h.

The effective coupling filter in the frequency domain can then be evaluated pointwise at a desired set of frequencies ω by matrix multiplication,

$$\hat{\mathbf{J}}^{\mathrm{eff}}(\omega) = \frac{\alpha^2}{(\alpha+i\omega)^2}\mathcal{J}^{\mathrm{RR}} + \left(\frac{\alpha^2}{(\alpha+i\omega)^2}\right)^2 \mathcal{J}^{\mathrm{RH}}\hat{\mathbf{\Gamma}}(\omega)\mathcal{J}^{\mathrm{HR}}.$$

We then return to the time domain by inverse Fourier transforming the result, achieved by treating $\hat{\mathbf{J}}^{\mathrm{eff}}(\omega)$ as an $N_{\mathrm{rec}} \times N_{\mathrm{rec}} \times N_{\mathrm{freq}}$ array (where N_{freq} is the number of frequencies at which we evaluate the effective coupling) and multiplying along the frequency dimension by an $N_{\mathrm{freq}} \times N_{\mathrm{time}}$ matrix $\mathbf{E}$ with elements $E_{\omega,t} = \exp(i\omega t)\Delta t/(2\pi)$, for N_{time} sufficiently small time bins of size $\delta t = 0.1/\alpha$, for $\alpha = 10$, as listed in Table SII.

To generate Fig. 3, we focus on the zero-frequency component of $\hat{J}^{\mathrm{eff}}(\omega)$, which is also equal to the time integral of $\mathbf{J}^{\mathrm{eff}}(t)$. As in the main text, we label the elements of this component $\mathcal{J}^{\mathrm{eff}}_{r,r'} = \hat{J}^{\mathrm{eff}}_{r,r'}(\omega=0)$, which is equal to

$$\mathcal{J}^{\mathrm{eff}}_{r,r'} = \mathcal{J}_{r,r'} + \sum_{h,h'} \mathcal{J}_{r,h}\hat{\Gamma}_{h,h'}(0)\mathcal{J}_{h',r'}.$$

We do not need to simulate the full network to study the statistics of $\mathcal{J}^{\mathrm{eff}}_{r,r'}$. We only need to generate samples of the matrix $\mathcal{J}$ and evaluate $\hat{\mathbf{\Gamma}}(0)$. This is where the choice of an Erdős-Réyni network that is not restricted to obey Dale's principle becomes convenient. Because the weights $\mathcal{J}_{i,j}$ are *i.i.d.* and the sign of the weight is random, population averages will be equivalent to expected values. i.e., the sample mean

$$\tilde{\mathcal{J}}_{\mathrm{mean}} = \frac{1}{N_{\mathrm{rec}}(N_{\mathrm{rec}}-1)} \sum_{r\neq r'} \mathcal{J}^{\mathrm{eff}}_{r,r'}$$

and sample variance

$$\tilde{\mathcal{J}}_{\mathrm{var}} = \frac{1}{N_{\mathrm{rec}}(N_{\mathrm{rec}}-1)-1} \sum_{r\neq r'} \left(\mathcal{J}^{\mathrm{eff}}_{r,r'} - \tilde{\mathcal{J}}_{\mathrm{mean}}\right)^2$$

will tend to the expected values $\mathbb{E}[\mathcal{J}^{\mathrm{eff}}_{r,r'}]$ and $\mathrm{var}[\mathcal{J}^{\mathrm{eff}}_{r,r'}]$ for large networks. We have explicitly removed the diagonal elements from these averages because these elements will have slightly different statistics from the off-diagonal elements due to the fact that all ground-truth self-couplings are set to zero, $\mathcal{J}_{r,r} = 0$. This allows us to compare the population variance, plotted in Fig. 3 (after normalization by the population variance of the true off-diagonal weights), to the expected variance calculated analytically below.

The error bars in Fig. 3 are generated by first drawing a single sample of true weights $\mathcal{J}$, and then taking 100 random subsets of $N_{\mathrm{rec}} = \{10, 110, 210, 310, 410, 510, 610, 710, 810, 910, 999\}$ recorded neurons. For this analysis, random subsets were generated by permuting the indices of the full weight matrix $\mathcal{J}$ and taking the last N_{rec} neurons to be recorded. For each random subset of the network we calculate the population statistics. The standard error of, for example, the population variance $\tilde{\mathcal{J}}_{\mathrm{var}}$ across subsets gives an estimate of the error. However, if we only use a single sample of the network architecture and weights $\mathcal{J}_{i,j}$, this estimate may depend on the particular instantiation of the network. To average over the effects of global network architecture, we draw a total of 10 network architecture samples, and average a second time over these samples to obtain our final estimates of the population variance of $\mathcal{J}^{\mathrm{eff}}_{r,r'}$. We note that for an Erdős-Réyni network, this second stage of averaging is probabilistically unnecessary: for a large enough network random subsets of a single large network are statistically identical to random subsets drawn from several samples of full Erdős-Réyni networks. However, this will not be true for networks with more structure, such as Watts-Strogatz networks or networks that obey Dale's principle, and this second stage of averaging over the global network architecture is necessary in these cases.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

F. Derivation of a series approximation for the mean field firing rates (exponential nonlinearity)

The mean field firing rates for the hidden neurons are given by

$$\nu_h = \lambda_0 \exp\left(\mu_h + \sum_{h'} \mathcal{J}_{h,h'} \nu_{h'}\right),$$

where we focus specifically on the case of exponential nonlinearity $\phi(x) = \exp(x)$. For this choice of nonlinearity, $\gamma_h = \nu_h$, so our series for the mean field firing rates is also automatically a series for the gains.

This system of transcendental equations generally cannot be solved analytically. However, for small $\exp(\mu_h) \ll 1$ we can derive, recursively, a series expansion for the firing rates. We first consider the case of $\mu_h = \mu_0$ for all hidden neurons h. Let $\epsilon = \exp(\mu_0)$. We may then write

$$\nu_h = \lambda_0 \epsilon \sum_{\ell=0}^{\infty} a_h^{(\ell)} (\lambda_0 \epsilon)^\ell.$$

Plugging this into the mean field equation,

$$\begin{aligned}
\sum_{\ell=0}^{\infty} a_h^{(\ell)} (\lambda_0 \epsilon)^\ell &= \exp\left(\sum_{h'} \mathcal{J}_{h,h'} \sum_{\ell=0}^{\infty} a_{h'}^{(\ell)} (\lambda_0 \epsilon)^{\ell+1}\right) \\
&= 1 + \sum_{m=1}^{\infty} \frac{1}{m!} \left(\sum_{h'} \mathcal{J}_{h,h'} \sum_{\ell=0}^{\infty} a_{h'}^{(\ell)} (\lambda_0 \epsilon)^{\ell+1}\right)^m \\
&= 1 + \sum_{m=1}^{\infty} \frac{1}{m!} \sum_{\ell_1,\ldots,\ell_m,h'_1,\ldots,h'_m} \mathcal{J}_{h,h'_1} a_{h'_1}^{(\ell_1)} \ldots \mathcal{J}_{h,h'_m} a_{h'_m}^{(\ell_m)} (\lambda_0 \epsilon)^{\ell_1+\cdots+\ell_m+m} \\
&= 1 + \sum_{\ell=1}^{\infty} \left\{ \sum_{m=1}^{\infty} \frac{1}{m!} \sum_{\ell_1,\ldots,\ell_m,h'_1,\ldots,h'_m} \mathcal{J}_{h,h'_1} a_{h'_1}^{(\ell_1)} \ldots \mathcal{J}_{h,h'_m} a_{h'_m}^{(\ell_m)} \delta_{\ell,\ell_1+\cdots+\ell_m+m} \right\} (\lambda_0 \epsilon)^\ell.
\end{aligned}$$

Thus, matching powers of $\lambda_0 \epsilon$ on the left and right hand sides, we find $a_h^{(0)} = 1$ and

$$a_h^{(\ell)} = \sum_{m=1}^{\infty} \frac{1}{m!} \sum_{\ell_1,\ldots,\ell_m,h'_1,\ldots,h'_m} \mathcal{J}_{h,h'_1} a_{h'_1}^{(\ell_1)} \ldots \mathcal{J}_{h,h'_m} a_{h'_m}^{(\ell_m)} \delta_{\ell,\ell_1+\cdots+\ell_m+m}$$

for $\ell > 0$.

For $\ell = 1$, the sum in m truncates at $m = 1$ (as $\delta_{\ell,\ell_1+\cdots+\ell_m+m}$ is zero for $m > \ell$, as all indices are positive). Thus,

$$\begin{aligned}
a_h^{(1)} &= \sum_{\ell_1=0}^{\infty} \sum_{h'_1} \mathcal{J}_{h,h'_1} a_{h'_1}^{(\ell_1)} \delta_{1,\ell_1+1} \\
&= \sum_{h'_1} \mathcal{J}_{h,h'_1} a_{h'_1}^{(0)} \\
&= \sum_{h'_1} \mathcal{J}_{h,h'_1}.
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For $\ell = 2$,

$$\begin{aligned}
a_h^{(2)} &= \sum_{\ell_1=0}^{\infty} \sum_{h_1'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(\ell_1)} \delta_{2,\ell_1+1} + \frac{1}{2} \sum_{\ell_1,\ell_2} \sum_{h_1',h_2'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(\ell_1)} \mathcal{J}_{h,h_2'} a_{h_2'}^{(\ell_2)} \delta_{2,\ell_1+\ell_2+2} \\
&= \sum_{h_1'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(1)} + \frac{1}{2} \sum_{h_1',h_2'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(0)} \mathcal{J}_{h,h_2'} a_{h_2'}^{(0)} \\
&= \sum_{h_1',h_2'} \mathcal{J}_{h,h_1'} \mathcal{J}_{h_1',h_2'} + \frac{1}{2} \left(\sum_{h'} \mathcal{J}_{h,h'} \right)^2 \\
&= \sum_{h_1',h_2'} \left\{ \mathcal{J}_{h,h_1'} \mathcal{J}_{h_1',h_2'} + \frac{1}{2} \mathcal{J}_{h,h_1'} \mathcal{J}_{h,h_2'} \right\}.
\end{aligned}$$

For $\ell = 3$,

$$\begin{aligned}
a_h^{(3)} &= \sum_{\ell_1=0}^{\infty} \sum_{h_1'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(\ell_1)} \delta_{3,\ell_1+1} + \frac{1}{2} \sum_{\ell_1,\ell_2} \sum_{h_1',h_2'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(\ell_1)} \mathcal{J}_{h,h_2'} a_{h_2'}^{(\ell_2)} \delta_{3,\ell_1+\ell_2+2} \\
&\quad + \frac{1}{3!} \sum_{\ell_1,\ldots,\ell_3} \sum_{h_1',\ldots,h_3'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(\ell_1)} \ldots \mathcal{J}_{h,h_3'} a_{h_3'}^{(\ell_3)} \delta_{3,\ell_1+\cdots+\ell_3+3} \\
&= \sum_{h_1'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(2)} + \frac{1}{2} \sum_{h_1',h_2'} \left\{ \mathcal{J}_{h,h_1'} a_{h_1'}^{(1)} \mathcal{J}_{h,h_2'} a_{h_2'}^{(0)} + \mathcal{J}_{h,h_1'} a_{h_1'}^{(0)} \mathcal{J}_{h,h_2'} a_{h_2'}^{(1)} \right\} \\
&\quad + \frac{1}{3!} \sum_{h_1',\ldots,h_3'} \mathcal{J}_{h,h_1'} a_{h_1'}^{(0)} \ldots \mathcal{J}_{h,h_3'} a_{h_3'}^{(0)} \\
&= \sum_{h_1'} \mathcal{J}_{h,h_1'} \left\{ \sum_{h_2',h_3'} \mathcal{J}_{h_1',h_2'} \mathcal{J}_{h_2',h_3'} + \frac{1}{2} \left(\sum_{h'} \mathcal{J}_{h_1',h'} \right)^2 \right\} + \frac{1}{2} \sum_{h_1',h_2',h_3'} \left\{ \mathcal{J}_{h,h_1'} \mathcal{J}_{h_1',h_3'} \mathcal{J}_{h,h_2'} + \mathcal{J}_{h,h_1'} \mathcal{J}_{h,h_2'} \mathcal{J}_{h_2',h_3'} \right\} \\
&\quad + \frac{1}{3!} \left(\sum_{h'} \mathcal{J}_{h,h'} \right)^3 \\
&= \sum_{h_1'} \mathcal{J}_{h,h_1'} \left\{ \sum_{h_2',h_3'} \mathcal{J}_{h_1',h_2'} \mathcal{J}_{h_2',h_3'} + \frac{1}{2} \left(\sum_{h'} \mathcal{J}_{h_1',h'} \right)^2 \right\} + \sum_{h_1',h_2',h_3'} \mathcal{J}_{h,h_1'} \mathcal{J}_{h,h_2'} \mathcal{J}_{h_2',h_3'} + \frac{1}{3!} \left(\sum_{h'} \mathcal{J}_{h,h'} \right)^3 \\
&= \sum_{h_1',h_2',h_3'} \left\{ \mathcal{J}_{h,h_1'} \mathcal{J}_{h_1',h_2'} \mathcal{J}_{h_2',h_3'} + \frac{1}{2} J_{h,h_1'} \mathcal{J}_{h_1',h_2'} \mathcal{J}_{h_1',h_3'} + \mathcal{J}_{h,h_1'} \mathcal{J}_{h,h_2'} \mathcal{J}_{h_2',h_3'} + \frac{1}{3!} \mathcal{J}_{h,h_1'} \mathcal{J}_{h,h_2'} \mathcal{J}_{h,h_3'} \right\}.
\end{aligned}$$

With this we have calculated the firing rates to $\mathcal{O}(\epsilon^4)$.

The analysis can be straightforwardly extended to the case of heterogeneous μ_h, though it becomes more tedious to compute terms in the (now multivariate) series. Assuming $\epsilon_h \equiv \exp(\mu_h) \ll 1$ for all h, to $\mathcal{O}(\epsilon^3)$ we find

$$\nu_h = \lambda_0 \epsilon_h \left(1 + \sum_{h'} \mathcal{J}_{h,h'} \lambda_0 \epsilon_{h'} + \sum_{h_1',h_2'} \left\{ \mathcal{J}_{h,h_1'} \mathcal{J}_{h_1',h_2'} + \frac{1}{2} \mathcal{J}_{h,h_1'} \mathcal{J}_{h,h_2'} \right\} \lambda_0 \epsilon_{h_1'} \lambda_0 \epsilon_{h_2'} + \ldots \right).$$

G. Variance of the effective coupling to second order in N_{rec}/N & fourth order in $\lambda_0 J_0 e^{\mu_0}$ (exponential nonlinearity)

To estimate the strength of the hidden paths, we would like to calculate the variance of the effective coupling $\mathcal{J}_{r,r'}^{\text{eff}}$ and compare its strength to the variance of the direct couplings $\mathcal{J}_{r,r'}$, where $\mathcal{J}_{r,r'}^{\text{eff}} \equiv \int_0^\infty dt\, J_{r,r'}^{\text{eff}}(t)$ and $\mathcal{J}_{r,r'} \equiv \int_0^\infty dt\, J_{r,r'}^{\text{eff}}(t)$, as in the main text.

We assume that the synaptic weights $\mathcal{J}_{i,j}$ are independently and identically distributed with zero mean and variance

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\mathrm{var}(\mathcal{J}) = p\frac{J_0^2}{(pN)^{2a}}$ for $i \neq j$, where $a = 1$ corresponds to weak coupling and $a = 1/2$ corresponds to strong coupling. We assume no self-couplings, $\mathcal{J}_{i,i} = 0$ for all neurons i. The overall factor of p in $\mathrm{var}[\mathcal{J}]$ comes from the sparsity of the network. For example, for normally distributed non-zero weights with variance J_0^2/N^{2a}, the total probability for every connection in the network is

$$\rho_{ER\times J}(\mathcal{J}) = (1-p)\delta(\mathcal{J}) + p\frac{\exp\left(-\frac{N^{2a}}{2}\frac{\mathcal{J}^2}{J_0^2}\right)}{\sqrt{2\pi J_0^2/N^{2a}}}.$$

Because the $\mathcal{J}_{i,j}$ are *i.i.d.*, the mean of $\mathcal{J}^{\mathrm{eff}}_{r,r'}$:

$$\begin{aligned}\overline{\mathcal{J}^{\mathrm{eff}}_{r,r'}} &= \overline{\mathcal{J}_{r,r'}} + \sum_{h,h'} \overline{\mathcal{J}_{r,h}\hat{\Gamma}_{h,h'}\mathcal{J}_{h',r'}} \\ &= 0 + \sum_{h,h'} \overline{\mathcal{J}_{r,h}}\ \overline{\hat{\Gamma}_{h,h'}}\ \overline{\mathcal{J}_{h',r'}} \\ &= 0,\end{aligned}$$

where we used the fact that $\hat{\Gamma}_{h,h'} \equiv \hat{\Gamma}_{h,h'}(0)$ depends only on the hidden neuron couplings $\mathcal{J}_{h,h'}$, which are independent of the couplings to the recorded neurons, $\mathcal{J}_{r,h}$ and $\mathcal{J}_{h',r'}$. This holds for any pair of neurons (r, r'), including $r = r'$ because of the assumption of no self-coupling.

The variance of $\mathcal{J}^{\mathrm{eff}}_{r,r'}$ is thus equal to the mean of its square, for $r \neq r'$,

$$\begin{aligned}\mathrm{var}\left[\mathcal{J}^{\mathrm{eff}}_{r,r'}\right] &= \overline{\left(\mathcal{J}^{\mathrm{eff}}_{r,r'}\right)^2} \\ &= \overline{(\mathcal{J}_{r,r'})^2} + \overline{\left(\sum_{h,h'} \mathcal{J}_{r,h}\hat{\Gamma}_{h,h'}\mathcal{J}_{h',r'}\right)^2} \\ &= \mathrm{var}[\mathcal{J}] + \sum_{h_1,h_1',h_2,h_2'} \overline{\mathcal{J}_{r,h_1}\hat{\Gamma}_{h_1,h_1'}\mathcal{J}_{h_1',r'}\mathcal{J}_{r,h_2}\Gamma_{h_2,h_2'}\mathcal{J}_{h_2',r'}} \\ &= \mathrm{var}[\mathcal{J}] + \sum_{h,h'} \overline{\mathcal{J}^2_{r,h}}\ \overline{\hat{\Gamma}^2_{h,h'}}\ \overline{\mathcal{J}^2_{h',r'}} \\ &= \mathrm{var}[\mathcal{J}] + \mathrm{var}[\mathcal{J}]^2 \sum_{h,h'} \overline{\hat{\Gamma}^2_{h,h'}}\end{aligned}$$

In this derivation, we used the fact that $\overline{\mathcal{J}_{r,h_1}\mathcal{J}_{r,h_2}} = \overline{\mathcal{J}^2_{r,h_1}}\delta_{h_1,h_2}$ due to the fact that the synaptic weights are uncorrelated. We now need to compute $\overline{\hat{\Gamma}^2_{h,h'}}$. This is intractable in general, so we will resort to calculating this in a series expansion in powers of $\epsilon \equiv \exp(\mu_0)$ for the exponential nonlinearity model. Our result will also turn out to be an expansion in powers of J_0 and $1 - f \equiv N_{\mathrm{hid}}/N$. Because we are explicitly using the exponential nonlinearity, $\gamma_h = \nu_h$, so we do not need to derive a series for γ_h in powers of ϵ in order to perform this calculation.

The lowest order approximation is obtained by the approximation $\nu_h \approx \lambda_0\epsilon$ and $\Gamma_{h,h'} \approx \nu_h\delta_{h,h'}$, yielding

$$\begin{aligned}\frac{\mathrm{var}\left[\mathcal{J}^{\mathrm{eff}}_{r,r'}\right]}{\mathrm{var}[\mathcal{J}]} &= 1 + (\lambda_0\epsilon)^2 N_{\mathrm{hid}}\mathrm{var}[\mathcal{J}] \\ &= 1 + (\lambda_0 J_0\epsilon)^2(1-f)\frac{1}{(pN)^{2a-1}}.\end{aligned} \tag{S.3}$$

This result varies linearly with f, while numerical evaluation of the variance shows obvious curvature for $f \ll 1$ and $J_0 \lesssim 1$, so we need to go to higher order. This becomes tedious very quickly, so we will only work to $\mathcal{O}(\epsilon^4)$ (it turns out $\mathcal{O}(\epsilon^3)$ corrections vanish).

We calculate $\overline{\hat{\Gamma}^2_{h,h'}}$ using a recursive strategy, rather than the formal series solution. That is, we begin with the

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expression

$$\hat{\Gamma}_{h,h'} = \nu_h \delta_{h,h'} + \sum_{h''} \nu_h \mathcal{J}_{h,h''} \hat{\Gamma}_{h'',h'}$$

and plug it into itself until we obtain an expression to a desired order in ϵ. In doing so, we note that $\nu_h \sim \mathcal{O}(\epsilon)$, so we will first work to fourth order in ν_h, and then plug in the series for ν_h in powers of ϵ.

We begin with

$$\begin{aligned}
\hat{\Gamma}^2_{h,h'} &= \nu_h^2 \delta_{h,h'} + 2\delta_{h,h'} \sum_{h''} \nu_h^2 \mathcal{J}_{h,h''} \hat{\Gamma}_{h'',h'} + \left(\sum_{h''} \nu_h \mathcal{J}_{h,h''} \hat{\Gamma}_{h'',h'} \right)^2 \\
&= \nu_h^2 \delta_{h,h'} + 2\delta_{h,h'} \sum_{h''} \nu_h^2 \mathcal{J}_{h,h''} \hat{\Gamma}_{h'',h'} + \sum_{h_1,h_2} \nu_h^2 \mathcal{J}_{h,h_1} \mathcal{J}_{h,h_2} \hat{\Gamma}_{h_1,h'} \hat{\Gamma}_{h_2,h'} \\
&\approx \nu_h^2 \delta_{h,h'} + 2\delta_{h,h'} \sum_{h''} \nu_h^2 \mathcal{J}_{h,h''} \left\{ \nu_{h''} \delta_{h'',h'} + \sum_{h_2} \nu_{h''} \mathcal{J}_{h'',h_2} \nu_{h_2} \delta_{h_2,h'} \right\} + \sum_{h_1,h_2} \nu_h^2 \nu_{h'}^2 \mathcal{J}_{h,h_1} \mathcal{J}_{h,h_2} \delta_{h_1,h'} \delta_{h_2,h'} \\
&= \nu_h^2 \delta_{h,h'} + 2\delta_{h,h'} \left\{ \nu_h^2 \nu_{h'} \mathcal{J}_{h,h'} + \sum_{h''} \nu_h^2 \nu_{h''} \mathcal{J}_{h,h''} \mathcal{J}_{h'',h'} \nu_{h'} \right\} + \nu_h^2 \nu_{h'}^2 \mathcal{J}^2_{h,h'} \\
&= \left\{ \nu_h^2 + 2\nu_h^2 \nu_{h'} \mathcal{J}_{h,h'} + 2 \sum_{h''} \nu_h^2 \nu_{h''} \mathcal{J}_{h,h''} \mathcal{J}_{h'',h'} \nu_{h'} \right\} \delta_{h,h'} + \nu_h^2 \nu_{h'}^2 \mathcal{J}^2_{h,h'} \\
&= \left\{ \nu_h^2 + 2 \sum_{h''} \nu_h^3 \nu_{h''} \mathcal{J}_{h,h''} \mathcal{J}_{h'',h} \right\} \delta_{h,h'} + \nu_h^2 \nu_{h'}^2 \mathcal{J}^2_{h,h'}
\end{aligned}$$

The third order term $\nu_h^3 \mathcal{J}_{h,h'} \delta_{h,h'}$ vanished because we assume no self-couplings. We have obtained $\hat{\Gamma}^2_{h,h'}$ to fourth order in ν_h; now we need to plug in the series expression for ν_h to obtain the series in powers of $\lambda_0 \epsilon$. We will do this order by order in ν_h. The easiest terms are the fourth order terms, as

$$\nu_h^2 \nu_{h'}^2 \approx (\lambda_0 \epsilon)^4 \text{ and } \nu_h^3 \nu_{h''} \approx (\lambda_0 \epsilon)^4.$$

The second order term is

$$\begin{aligned}
\nu_h^2 &\approx (\lambda_0 \epsilon)^2 \left(1 + \sum_{h_1} \mathcal{J}_{h,h_1} \lambda_0 \epsilon + \sum_{h_1,h_2} a^{(2)}_{h,h_1,h_2} (\lambda_0 \epsilon)^2 \right) \left(1 + \sum_{h_1'} \mathcal{J}_{h,h_1'} \lambda_0 \epsilon + \sum_{h_1',h_2'} a^{(2)}_{h,h_1',h_2'} (\lambda_0 \epsilon)^2 \right) \\
&\approx (\lambda_0 \epsilon)^2 \left(1 + 2 \left(\sum_{h_1} \mathcal{J}_{h,h_1} \lambda_0 \epsilon + \sum_{h_1,h_2} a^{(2)}_{h,h_1,h_2} (\lambda_0 \epsilon)^2 \right) + \left(\sum_{h_1} \mathcal{J}_{h,h_1} \lambda_0 \epsilon \right)^2 \right) \\
&= (\lambda_0 \epsilon)^2 \left(1 + 2 \sum_{h_1} \mathcal{J}_{h,h_1} \lambda_0 \epsilon + \sum_{h_1,h_2} \left\{ 2a^{(2)}_{h,h_1,h_2} + \mathcal{J}_{h,h_1} \mathcal{J}_{h,h_2} \right\} (\lambda_0 \epsilon)^2 \right),
\end{aligned}$$

where $a^{(2)}_{h,h_1,h_2} = \mathcal{J}_{h,h_1} \mathcal{J}_{h_1,h_2} + \frac{1}{2} \mathcal{J}_{h,h_1} \mathcal{J}_{h,h_2}$. We need the average $\overline{\nu_h^2}$. The third-order term will vanish upon averaging, and $\overline{2a^{(2)}_{h,h_1,h_2} + \mathcal{J}_{h,h_1} \mathcal{J}_{h,h_2}} = \overline{2\mathcal{J}_{h,h_1} \mathcal{J}_{h_1,h_2} + 2\mathcal{J}_{h,h_1} \mathcal{J}_{h,h_2}} = 2\text{var}[\mathcal{J}] \delta_{h_1,h_2} (1 - \delta_{h,h_1})$, using the fact that synaptic weights are independent (giving the δ_{h_1,h_2} factor) and self-couplings are zero (giving the $1 - \delta_{h,h_1}$ factor). We thus obtain

$$\overline{\nu_h^2} = (\lambda_0 \epsilon)^2 + 2(\lambda_0 \epsilon)^4 (N_{\text{hid}} - 1) \text{var}[\mathcal{J}].$$

The first fourth order term in $\hat{\Gamma}^2_{h,h'}$, $2 \sum_{h''} \nu_h^3 \nu_{h''} \mathcal{J}_{h,h''} \mathcal{J}_{h'',h} \delta_{h,h'}$, will vanish upon averaging because matching indices requires $h'' = h = h'$ and we assume no self-couplings. The second fourth order term is $\mathcal{J}^2_{h,h'}$, which averages to $\text{var}[\mathcal{J}](1 - \delta_{h,h'})$, where the factor of $(1 - \delta_{h,h'})$ again accounts for the fact that this term does not contribute when

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$h = h'$ due to no self-couplings. We thus arrive at

$$\begin{aligned}\overline{\hat{\Gamma}^2_{h,h'}} &= \left((\lambda_0\epsilon)^2 + 2(\lambda_0\epsilon)^4(N_{\text{hid}}-1)\text{var}[\mathcal{J}]\right)\delta_{h,h'} + (\lambda_0\epsilon)^4\text{var}[\mathcal{J}](1-\delta_{h,h'}) \\ &= \left((\lambda_0\epsilon)^2 + (\lambda_0\epsilon)^4(2N_{\text{hid}}-3)\text{var}[\mathcal{J}]\right)\delta_{h,h'} + (\lambda_0\epsilon)^4\text{var}[\mathcal{J}];\end{aligned}$$

Putting everything together,

$$\begin{aligned}\frac{\text{var}\left[\mathcal{J}^{\text{eff}}_{r,r'}\right]}{\text{var}[\mathcal{J}]} &= 1 + \text{var}[\mathcal{J}]\sum_{h,h'}\overline{\hat{\Gamma}^2_{h,h'}} \\ &= 1 + \text{var}[\mathcal{J}]\left[\sum_h\left\{(\lambda_0\epsilon)^2 + (\lambda_0\epsilon)^4(2N_{\text{hid}}-3)\text{var}[\mathcal{J}]\right\} + \sum_{h,h'}(\lambda_0\epsilon)^4\text{var}[\mathcal{J}]\right] \\ &= 1 + \text{var}[\mathcal{J}]\left[N_{\text{hid}}\left\{(\lambda_0\epsilon)^2 + (\lambda_0\epsilon)^4(2N_{\text{hid}}-3)\text{var}[\mathcal{J}]\right\} + N^2_{\text{hid}}(\lambda_0\epsilon)^4\text{var}[\mathcal{J}]\right] \\ &= 1 + N_{\text{hid}}\text{var}[\mathcal{J}]\left[(\lambda_0\epsilon)^2 + (\lambda_0\epsilon)^4\left(2N_{\text{hid}}-3\right)\text{var}[\mathcal{J}] + N_{\text{hid}}(\lambda_0\epsilon)^4\text{var}[\mathcal{J}]\right] \\ &= 1 + N_{\text{hid}}\text{var}[\mathcal{J}]\left[(\lambda_0\epsilon)^2 + (\lambda_0\epsilon)^4\left(3 - \frac{3}{N_{\text{hid}}}\right)N_{\text{hid}}\text{var}[\mathcal{J}]\right]\end{aligned}$$

For weak coupling, this tends to 1 in the $N \gg 1$ limit, as $N_{\text{hid}}\text{var}[\mathcal{J}] = (1-f)J_0^2/N \to 0$, for fixed fraction of observed neurons $f = N_{\text{rec}}/N$. For strong coupling, $N_{\text{hid}}\text{var}[\mathcal{J}] = (1-f)J_0^2$, which is constant as $N \to \infty$, and hence

$$\frac{\text{var}\left[\mathcal{J}^{\text{eff}}_{r,r'}\right]}{\text{var}[\mathcal{J}]} = 1 + (\lambda_0 J_0\epsilon)^2(1-f) + 3(\lambda_0 J_0\epsilon)^4(1-f)^2 + o\left((\lambda_0 J_0\epsilon)^4(1-f)^2\right), \tag{S.4}$$

where we have used little-o notation to denote that there are higher order terms dominated by $(\lambda_0 J_0\epsilon)^4(1-f)^2$. With this expression, we have improved on our approximation of the relative variance of the effective coupling to the true coupling; however, the neglected higher order terms still become significant as $f \to 0$ and $J_0 \to 1$, indicating that hidden paths have a significant impact when synaptic strengths are moderately strong and only a small fraction of the neurons have been observed.

Because the synaptic weights $\mathcal{J}_{i,j}$ are independent, we may rewrite Eq. (S.4) as

$$\frac{\text{var}\left[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}\right]}{\text{var}[\mathcal{J}]} \approx (\lambda_0 J_0\epsilon)^2(1-f) + 3(\lambda_0 J_0\epsilon)^4(1-f)^2;$$

or, in terms of the ratio of standard deviations,

$$\frac{\sigma\left[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}\right]}{\sigma[\mathcal{J}]} \approx (\lambda_0 J_0\epsilon)\sqrt{1-f}\left(1 + \frac{3}{2}(\lambda_0 J_0\epsilon)^2(1-f)\right),$$

where we used the approximation $\sqrt{1+x} \approx 1 + x/2$ for x small.

In the main text, we plotted results for $N = 1000$ total neurons (Fig. 3). For strongly coupled networks, the results should only depend on the fraction of observed neurons, $f = N_{\text{rec}}/N$, while for weak coupling the results do depend on the absolute number N. To demonstrate this, in Fig. S5 we remake Fig. 3 for $N = 100$ neurons. We see that the strongly coupled results have not been altered, whereas the weakly coupled results yield stronger deviations (as the deviations are $\mathcal{O}(1/\sqrt{N})$).

H. Estimating the error from neglecting higher order spike filtering (exponential nonlinearity)

In the main text we calculate corrections to the baseline and linear spike filters, neglecting higher-order spike filtering and fluctuations around the mean input to the recorded neurons. We would like to know when these approximations are valid. We will do so within mean field theory (meaning the noise fluctuations contribute zero error as they do not contribute to the mean field approximation). Specifically, we will assume that the quadratic spike filtering term is small, and calculate the corresponding correction to our mean field approximation of the firing rates when this term

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

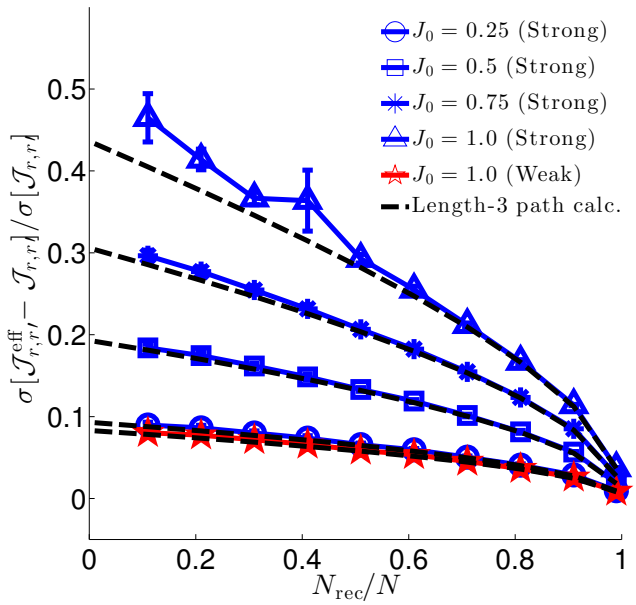

FIG. S5. Same as Fig. 3, but for $N = 100$ neurons and $N_{\rm rec} = \{1, 11, 21, 31, 41, 51, 61, 71, 81, 91, 99\}$ recorded neurons. Because we plot the relative deviations of the coupling strength against the fraction of observed neurons, the curves for the strongly coupled case are the same as for $N = 1000$, as expected, while the weakly coupled case yields stronger deviations.

is completely neglected. If we take as our approximation of the recorded neuron firing instantaneous firing rates

$$\lambda_r(t) \approx \lambda_0 \exp\left(\mu_r^{\rm eff} + \sum_{r_1} \int dt_1 J^{\rm eff}_{r,r'}(t-t_1)\dot{n}_{r_1}(t_1) + b \sum_{r_1,r_2} \int dt_1 dt_2 \; A_{r,r_1,r_2}(t,t_1,t_2)\dot{n}_{r_1}(t_1)\dot{n}_{r_2}(t_2)\right),$$

then the mean field approximation of the firing rates is

$$\langle \dot{n}_r \rangle \approx \lambda_0 \exp\left(\mu_r^{\rm eff} + \sum_{r_1} \mathcal{J}^{\rm eff}_{r,r'} \langle \dot{n}_{r_1} \rangle + b \sum_{r_1,r_2} \mathcal{A}_{r,r_1,r_2} \langle \dot{n}_{r_1} \rangle \langle \dot{n}_{r_2} \rangle\right),$$

where we have used the fact that the average firing rates are independent of time, and replaced $J^{\rm eff}_{r,r'}(t-t_1)$ and $A_{r,r_1,r_2}(t,t_1,t_2)$ with their time integrals, denoted by $\mathcal{J}^{\rm eff}_{r,r'}$ and $\mathcal{A}_{r,r_1,r_2}$. The parameter b is just a book-keeping parameter.

To calculate the lowest order correction to the linear filtering approximation ($b \to 0$), we write $\langle \dot{n}_r \rangle = \nu_r^{\rm sub} + b\tilde{\nu}_r$, treating b formally as a small parameter. The linear firing rate $\nu_r^{\rm sub}$ is given by

$$\nu_r^{\rm sub} = \lambda_0 \exp\left(\mu_r^{\rm eff} + \sum_{r'} \mathcal{J}^{\rm eff}_{r,r'} \nu^{\rm sub}_{r'}\right).$$

For the quadratically-modified firing rates, keeping terms only to linear order in b,

$$\begin{aligned}
\nu_r^{\rm sub} + b\tilde{\nu}_r &\approx \lambda_0 \exp\left(\mu_r^{\rm eff} + \sum_{r_1} \mathcal{J}^{\rm eff}_{r,r'} \nu^{\rm sub}_{r'} + b \sum_{r'} \mathcal{J}^{\rm eff}_{r,r'} \tilde{\nu}_{r'} + b \sum_{r_1,r_2} \mathcal{A}_{r,r_1,r_2} \nu^{\rm sub}_{r_1} \nu^{\rm sub}_{r_2}\right) \\
&= \nu_r^{\rm sub} \exp\left(b \sum_{r'} \mathcal{J}^{\rm eff}_{r,r'} \tilde{\nu}_{r'} + b \sum_{r_1,r_2} \mathcal{A}_{r,r_1,r_2} \nu^{\rm sub}_{r_1} \nu^{\rm sub}_{r_2}\right) \\
&\approx \nu_r^{\rm sub} \left\{1 + b \sum_{r'} \mathcal{J}^{\rm eff}_{r,r'} \tilde{\nu}_{r'} + b \sum_{r_1,r_2} \mathcal{A}_{r,r_1,r_2} \nu^{\rm sub}_{r_1} \nu^{\rm sub}_{r_2}\right\}.
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Collecting on b and rearranging,

$$\sum_{r'} \left[\delta_{r,r'} - \nu_r^{\text{sub}} \mathcal{J}_{r,r'}^{\text{eff}}\right] \tilde{\nu}_{r'} = \nu_r^{\text{sub}} \sum_{r_1,r_2} \mathcal{A}_{r,r_1,r_2} \nu_{r_1}^{\text{sub}} \nu_{r_2}^{\text{sub}}.$$

Because $\nu_r^{\text{sub}} \propto \exp(\mu_r^{\text{eff}}) \propto \exp(\mu_r) = \epsilon_r$, the expansion parameters we have been using, the lowest order approximation for $\tilde{\nu}_r$ is

$$\tilde{\nu}_r \approx \nu_r^{\text{sub}} \sum_{r_1,r_2} \mathcal{A}_{r,r_1,r_2} \nu_{r_1}^{\text{sub}} \nu_{r_2}^{\text{sub}}.$$

The coefficient $\mathcal{A}_{r,r_1,r_2}$ is the amplitude of the quadratic spike filter. The expression, which we calculate later in the SI, is

$$A_{r,r_1,2}(t,t_1,t_2) = (\lambda_0\epsilon) \sum_h \int dt' \; J_{r,h}(t-t') J_{h,r_1}(t'-t_1) J_{h,r_2}(t'-t_2),$$

and hence

$$\mathcal{A}_{r,r_1,r_2} \equiv \int dt_1 dt_2 \; A_{r,r_1,2}(t,t_1,t_2) = (\lambda_0\epsilon) \sum_h \mathcal{J}_{r,h} \mathcal{J}_{h,r_1} \mathcal{J}_{h,r_2}.$$

To lowest order the error term $\tilde{\nu}_r$ is

$$\tilde{\nu_r} = (\lambda_0\epsilon)^4 \sum_{h,r_1,r_2} \mathcal{J}_{r,h} \mathcal{J}_{h,r_1} \mathcal{J}_{h,r_2}.$$

For $\mathcal{J}_{i,j}$ *i.i.d.*, the population average should converge to the expected value, which is zero because the $\mathcal{J}_{i,j}$ have mean zero. We can calculate the root-mean-squared-error by looking at the variance:

$$\begin{aligned} \text{var}(\tilde{\nu}_r) = \text{var}\left(\sum_{h,r_1,r_2} J_{r,h} J_{h,r_1} J_{h,r_2}\right) &= \overline{\left(\sum_{h,r_1,r_2} \mathcal{J}_{r,h} \mathcal{J}_{h,r_1} \mathcal{J}_{h,r_2}\right)^2} \\ &= \sum_{h,r_1,r_2} \overline{\mathcal{J}_{r,h}^2} \; \overline{\mathcal{J}_{h,r_1}^2 \mathcal{J}_{h,r_2}^2} \end{aligned}$$

In principle, we should take care to separate out the $r_1 \neq r_2$ and $r_1 = r_2$ terms from the sum, as the latter will contribute a factor $\overline{\mathcal{J}_{h,r_1}^4}$, which we have not specified yet (though one could calculate for specific choices, such as the normal distribution that we use in practice). However, both $\overline{\mathcal{J}_{h,r}^2}^2$ and $\overline{\mathcal{J}_{h,r_1}^4}$ will scale as $(J_0^2/(pN)^{2a})^2$, so we will neglect constant factors and simply use this scaling to arrive at the result

$$\text{var}(\tilde{\nu}_r) \sim (\lambda_0\epsilon)^8 N_{\text{rec}}^2 N_{\text{hid}} \frac{J_0^6}{(pN)^{6a}}.$$

If we take $N \to \infty$ with $N_{\text{rec}} = fN$ and $N_{\text{hid}} = (1-f)N$ for f fixed, the root mean squared error (RMSE) scales as

$$\tilde{\nu}_{\text{RMSE}} \sim (\lambda_0\epsilon)^4 f\sqrt{1-f} \frac{J_0^3}{(pN)^{3a-3/2}}.$$

For $a = 1$ (weak coupling), the error falls off quite quickly as $N^{3/2}$, while it is independent of N for $a = 1/2$ (strong coupling). However, the error does still scale with the fraction of observed neurons, as $f\sqrt{1-f}$. This demonstrates that the typical error that arises from neglecting the nonlinear filtering is small both when most neurons have been observed ($f \lesssim 1$) and when very few neurons have been observed ($f \gtrsim 0$). While it may at first seem surprising that the error is small when very few neurons have been observed, the result does make intuitive sense: when a very small fraction of the network is observed, we can treat the unobserved portion of the network as a "reservoir" or "bath." Feedback from the observed neurons into the reservoir has a comparatively small effect, so we can get away with neglecting feedback between the observed and unobserved partitions of the network. However, when the number of observed neurons is comparable to the number of unobserved neurons, neither can be treated as a reservoir, and

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

feedback between the two partitions is substantial. Neglecting the higher order spike filter terms may be inaccurate in this case. We check this numerically below.

I. Validating the mean field approximation and linear conditional rate approximation via direct simulations of network activity (exponential nonlinearity)

The results presented in the main text are based on analytical calculations or numerical analyses using analytically derived formulas. For example, the statistics of $\mathcal{J}^{\text{eff}}_{r,r'}$ are calculated based on our expression $\mathcal{J}^{\text{eff}}_{r,r'} = \mathcal{J}_{r,r'} + \sum_{h,h'} \mathcal{J}_{r,h}\hat{\Gamma}_{h,h'}(0)\mathcal{J}_{h',r'}$, where $\hat{\Gamma}_{h,h'}(0)$ can be calculated by solving the matrix equation

$$\hat{\Gamma}_{h,h'}(0) = \delta_{h,h'} + \sum_{h''} \nu_h \mathcal{J}_{h,h''}\hat{\Gamma}_{h'',h'}(0).$$

Generating these results does not require a simulating the full network, so we check here that our approximations indeed agree with the results of full network simulations.

We check the validity of two main results: 1) that mean field theory is an accurate approximation for the parameters we consider, and 2) that our truncation of the conditional hidden firing rates $\mathbb{E}[\dot{n}_h(t)|\{\dot{n}_r\}]$ at linear order in $\dot{n}_r(t)$ is valid.

We first discuss some basic details of the simulation. The simulation code we use is a modification of the code used in [9], written by Gabe Ocker; refer to this paper for full details of the simulation.

The main changes we made are considering exponential nonlinearities and synaptic weights drawn from normal or lognormal distributions.

As in [9], we choose the coupling filters to follow an alpha function

$$g_j(t) = \alpha^2 t e^{-\alpha t}\Theta(t), \ \forall j.$$

The Heaviside step function $\Theta(t)$ enforces causality of the filter, using the convention $\Theta(0) = 0$. All neurons have the same time constant $1/\alpha$.

To efficiently simulate this network the code computes the synaptic variable $s_j(t) = \int dt' g(t-t')\dot{n}_j(t')$ not by direct convolution but by solving the inhomogeneous system of differential equations, setting $x(t) = s(t)$ and $y(t) = \dot{s}(t)$,

$$\begin{aligned}\dot{x}_j(t) &= y_j(t)\\ \dot{y}_j(t) &= -2\alpha_j y_j(t) - \alpha_j^2 x_j(t) + \alpha_j^2 \dot{n}_j(t),\end{aligned}$$

The instantaneous firing rates of the neurons can in this way be quickly computed in time steps of a specified size Δt. The number of spikes n_i that neuron i fires in the tth time bin is drawn from a Poisson distribution with probability $(\lambda_i(t)\Delta t)^{n_i}\exp(-\lambda_i(t)\Delta t)/(n_i)!$. An initial transient period of spiking before the network achieves a steady state is discarded.

The parameters we use in our simulations of the full network are given in Table SII.

TABLE SII. Network activity simulation parameter values.

Network connectivity parameters	See Table SI.
Alpha function decay time $\tau \equiv 1/\alpha$	10
Time bin width Δt	0.01τ
Transient time window	5τ
Simulation stopping time	4000τ + transient

1. *Verifying the mean field approximation*

To confirm that the mean field approximation is valid, we seek to compare the empirically measured spike rates measured from simulations of the network activity to the calculated mean field rates. The empirical rates are measured

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

as

$$\langle \dot{n}_i \rangle^{\text{emp}} = \frac{\text{number of spikes emitted by neuron } i}{\text{length of spike train window}},$$

calculated after discarding the initial transient period of firing, for any neuron i (recorded or hidden).

The steady-state mean field firing rates are the solutions of the transcendental equation

$$\langle \dot{n}_i \rangle^{\text{full MFT}} = \lambda_0 \exp\left(\mu_i + \sum_j \mathcal{J}_{i,j} \langle \dot{n}_j \rangle^{\text{full MFT}}\right).$$

The only difference between this equation and the equation for ν_h is that the neuron indices are not restricted to hidden units. i.e., the ν_h are the mean field rates for the hidden neurons *only* (recorded neurons removed entirely), whereas the $\langle \dot{n}_i \rangle^{\text{full MFT}}$ are the mean field rates for the entire network. If the mean field approximation is valid, the empirical rates should be approximately equal to the mean field rates, so a scatter plot of $\langle \dot{n}_i \rangle^{\text{MFT}}$ versus $\langle \dot{n}_i \rangle^{\text{emp}}$ should roughly lie along the identity line. We test this for a network in the strong coupling limit ($\sqrt{\text{var}(\mathcal{J})} = J_0/\sqrt{N}$) for four values of J_0, $J_0 = 0.25, 0.5, 0.75$, and 1.0. We expect $J_0 = 1.0$ to be close to the stability threshold of the model based on a linearized analysis [10, 11]; i.e., for $J_0 \gtrsim 1.0$ there may not be a steady state, so this may be where we expect the mean field approximation to break down. As seen in Fig. S6, the mean field approximation appears to hold well even up to $J_0 = 1.0$, though there are some slight deviations for neurons with large rates.

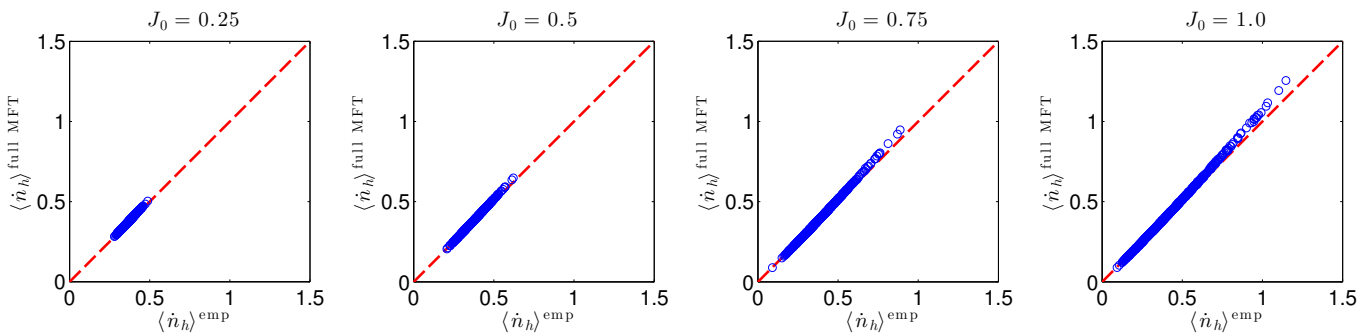

FIG. S6. Empirical estimates of average neuron firing rates from simulations plotted against mean firing rates predicted by mean field theory. The fact that the data lies along the identity line demonstrates validity of MFT up to $J_0 = 1.0$.

Verifying the linearized conditional mean approximation

Having verified that the mean field approximation is valid, we now seek to check our linearized approximation of the firing rates of the hidden neurons *conditioned on the activity of the recorded neurons*, $\mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r(t)\}\right]$. That is, we calculated above that

$$\mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r(t)\}\right] \approx \nu_h + \sum_{h',r} [\Gamma_{h,h'} * J_{h',r} * \dot{n}_r](t) + \ldots;$$

the $\ldots$ correspond to higher order spike filtering terms that we have neglected in our analyses, assuming them to be small. In an earlier calculation above, we estimated that the error incurred by neglecting higher order spike filtering is of the order $(\lambda_0 \exp(\mu_0))^4 f\sqrt{1-f} J_0^3$, but we would like to confirm the negligibility of the higher order coupling through simulations.

To do so, we compare the empirical firing rates of the designated "hidden" neurons obtained from simulations of the full network models with the approximation of the firing rates of the hidden neurons conditioned on the recorded neurons using the linear expansion, averaged over recorded neuron activity to give

$$\langle \dot{n}_h \rangle^{\text{approx}} \approx \nu_h + \sum_{h',r} \hat{\Gamma}_{h,h'}(0) \mathcal{J}_{h',r} \langle \dot{n}_r \rangle^{\text{emp}},$$

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where as usual the zero-frequency component of the linear response function $\hat{\Gamma}_{h,h'}(0)$ of the hidden neurons is calculated in the absence of recorded neurons.

If we make a scatter plot of this against the empirical estimates of the hidden neurons, $\langle \dot{n}_h \rangle^{\text{emp}}$, the data points will lie along the identity line if our approximation is valid. If the data deviates from the identity line, it indicates that the neglected higher-order filtering terms contribute substantially to the firing rates of the neurons. It is possible that the zeroth order rate approximation, ν_h, would be sufficient to describe the data, so we compare the empirical rates to both ν_h and $\langle \dot{n}_h \rangle^{\text{approx}}$.

As in the mean field approximation test, we focused on a strongly coupled network with $J_0 = 0.25, 0.5, 0.75$, and 1.0. Because our analytical estimate of the error suggests small error for small fractions of recorded neurons and larger error when $N_{\text{rec}} \sim N_{\text{hid}}$, we check both $N_{\text{rec}} = 100$ neurons out of $N = 1000$ neurons ($f = 0.1$) in Fig. S7 and $N_{\text{rec}} = 500$ neurons out of $N = 1000$ ($f = 0.5$) in Fig. S8.

For each value of J_0, we present two plots: the empirical rates versus the mean field rates ν_h in the absence of recorded neurons (the zeroth order approximation; Figs. S7 and S8, top row), and the empirical rates versus the linear approximation $\langle \dot{n}_h \rangle^{\text{approx}}$ (the first order approximation; Figs. S7 and S8, bottom row). We find that in both cases the data is centered around the identity line, but the spread of data grows with J_0 for the zeroth order approximation, while it is quite tight for the first order approximation up to $J_0 = 1.0$, validating our neglect of the higher order spike filtering terms. We also confirm that $N_{\text{rec}} = 500$ offers worse agreement than $N_{\text{rec}} = 100$, though the agreement between $\langle \dot{n}_h \rangle^{\text{emp}}$ and $\langle \dot{n}_h \rangle^{\text{approx}}$ is still not too bad.

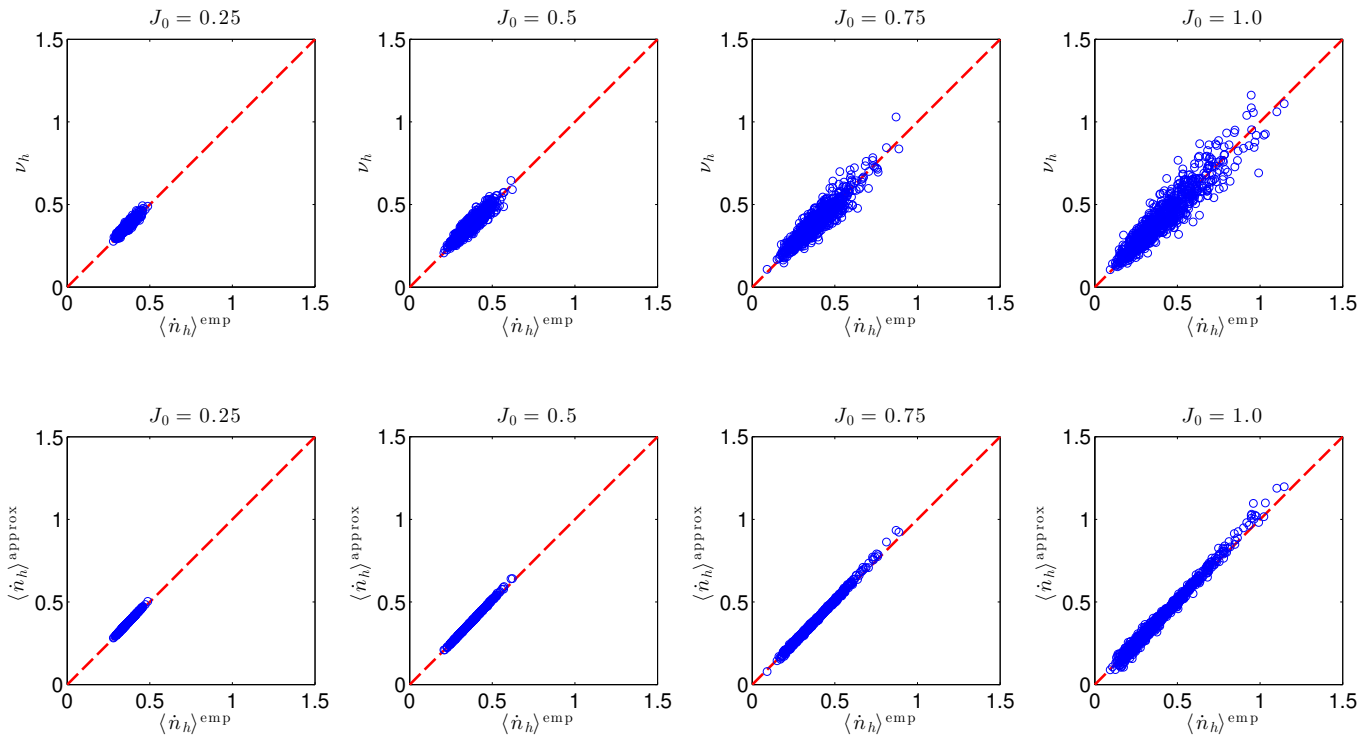

FIG. S7. **Top row:** scatter plot comparing ν_h, the mean field firing rates of the hidden neurons in the absence of recorded neurons, to empirically estimated firing rates in simulations of the full network, for four different values of typical synaptic strength, $J_0 = 0.25, 0.5, 0.75$, and 1.0. The data lie along the identity line, demonstrating a strong correlation between ν_h and the empirical data. However, the spread of data around the identity line indicates that deviations of the mean firing rates away from ν_h, caused by coupling to the recorded neurons, is significant. **Bottom row:** Comparison of the first order approximation of the firing rates of hidden neurons, which accounts for the effects of recorded neurons, to the empirical rates. The data lie tightly along the identity with very little dispersion, demonstrating that higher order spike filtering is unnecessary even up to $J_0 = 1.0$, for $N_{\text{rec}} = 100$.

J. Second order nonlinear response function

Higher order terms in the series expansion represent nonlinear response functions. We do not focus on these terms in this work, assuming instead that we can truncate this series expansion at linear order. We will, however, estimate

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

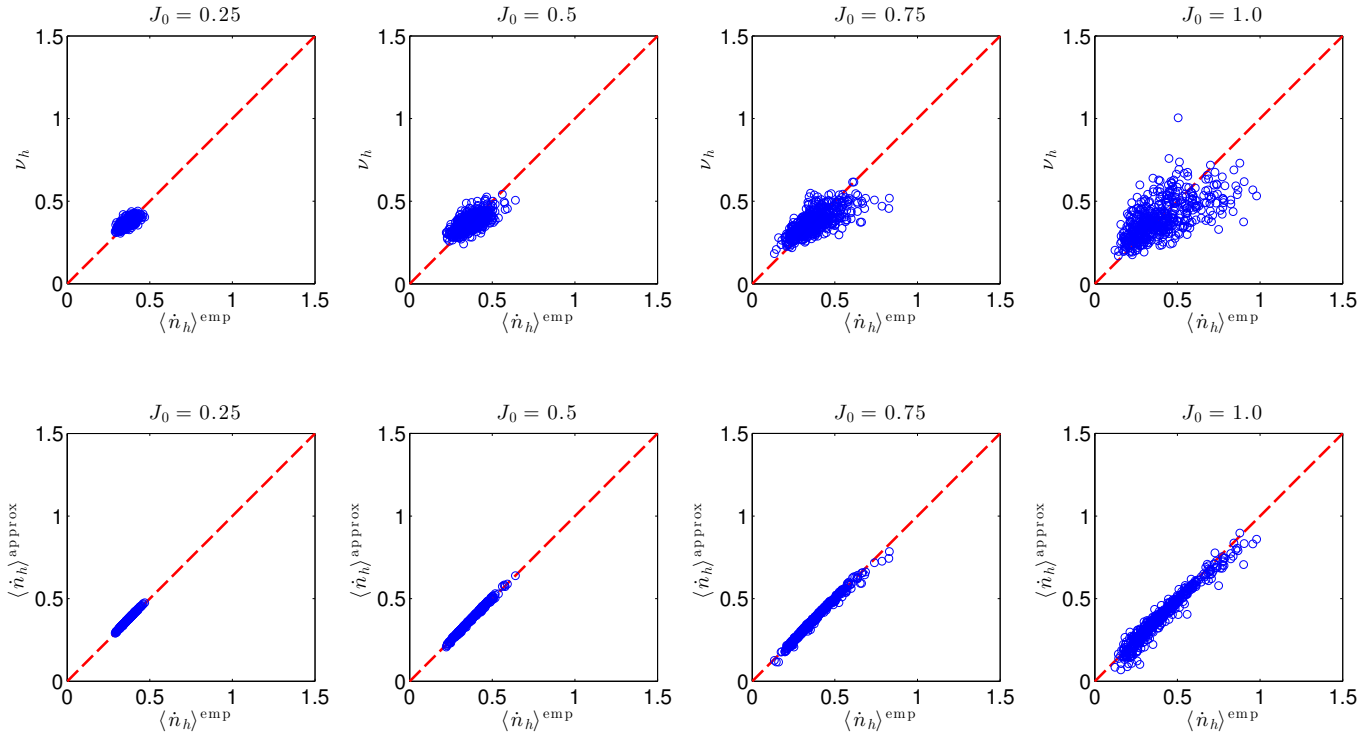

FIG. S8. Same as Fig. S7 but for $N_{\text{rec}} = 500$ recorded neurons out of a total of $N = 1000$. Demonstrates validity of linear approximation (neglecting higher order spike filtering) up to $J_0 = 1.0$, for $N_{\text{rec}} = 500$. The zeroth order approximation (top row) is quite poor, indicating the necessity of accounting for feedback from the recorded neurons. This first order approximation (bottom row) lies tightly along the identity line, indicating that even when the recorded and hidden populations are of comparable size, higher order spike filtering may not be significant. However, there appears to be some deviation for $J_0 = 1.0$, indicating that accounting for higher order spike filtering may be beneficial in this parameter regime.

the error incurred by this truncation. To do so will we need the next order term, so we now go to second order. Rather than differentiate our formal solution for the linear response, we differentiate the implicit form, yielding

$$\begin{aligned}
\Gamma^{(2)}_{h,h_1,h_2}(t,t_1,t_2) &\equiv \left.\frac{\delta^2 \mathbb{E}[\dot{n}_h|\,\{I_h(t)\}]}{\delta I_{h_2}(t_2)\delta I_{h_1}(t_1)}\right|_{I_h=0} \\
&= \gamma_h \left[\delta_{h,h_2}\delta(t-t_2) + \sum_{h'} J_{h,h'} * \Gamma_{h',h_2}\right]\left[\delta_{h-h_1}\delta(t-t_1) + \sum_{h'} J_{h,h'} * \Gamma_{h',h_1}\right] \\
&\quad + \gamma_h \left[\sum_{h'} \int_{-\infty}^{\infty} dt'\; J_{h,h'}(t-t')\Gamma^{(2)}_{h,h_1,h_2}(t',t_1,t_2)\right]
\end{aligned}$$

Rearranging,

$$\begin{aligned}
&\int dt' \sum_{h'} \left[\delta_{h,h'}\delta(t-t') - \gamma_h J_{h,h'}(t-t')\right]\Gamma^{(2)}_{h',h_1,h_2}(t',t_1,t_2) \\
&\qquad = \gamma_h \left[\delta_{h-h_2}\delta(t-t_2) + \sum_{h'} J_{h,h'} * \Gamma_{h',h_2}\right]\left[\delta_{h-h_1}\delta(t-t_1) + \sum_{h'} J_{h,h'} * \Gamma_{h',h_1}\right].
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Inverting the operator on the left hand side yields the input linear response function (when combined with the factor of γ_h on the right hand side), giving the solution

$$\Gamma^{(2)}_{h,h_1,h_2}(t,t_1,t_2) = \int_{-\infty}^{\infty} dt' \sum_{h'} \Gamma_{h,h'}(t-t') \left[\delta_{h'-h_2}\delta(t'-t_2) + \sum_{h''}\int_{-\infty}^{\infty} dt''\ J_{h',h''}(t'-t'')\Gamma_{h'',h_2}(t''-t_2)\right] \times \left[\delta_{h'-h_1}\delta(t'-t_1) + \sum_{h''}\int_{-\infty}^{\infty} dt''\ J_{h',h''}(t'-t'')\Gamma_{h'',h_1}(t''-t_1)\right]$$

Because $\Gamma_{h,h'}(t-t')$ is proportional to γ_h, the second order nonlinear response function is also proportional to γ_h.

The effective quadratic spike filtering that enters in the instantaneous rate of the recorded neurons is thus

$$\int dt' dt_1 dt_2 dt_1' dt_2' \sum_{h,h_1,h_2,r_1',r_2'} J_{r,h}(t-t')\Gamma^{(2)}_{h,h_1,h_2}(t',t_1,t_2) J_{h_1,r_1'}(t_1-t_1') J_{h_2,r_2'}(t_2-t_2')\dot{n}_{r_1'}(t_1')\dot{n}_{r_2'}(t_2'),$$

where we define the quadratic spike filter to be

$$A_{r,r_1',r_2'}(t,t_1',t_2') = \int dt' dt_1 dt_2 \sum_{h,h_1,h_2} J_{r,h}(t-t')\Gamma^{(2)}_{h,h_1,h_2}(t',t_1,t_2) J_{h_1,r_1'}(t_1-t_1') J_{h_2,r_2'}(t_2-t_2') \tag{S.5}$$

We would like to estimate the typical size of this term to leading order so that we may estimate the error we make by neglecting it. For the exponential nonlinearity, $\gamma_h = \nu_h$ and to leading order in $\nu_h \sim \lambda_0\epsilon$, $\Gamma^{(2)}_{h,h_1,h_2}(t,t_1,t_2) \approx \lambda_0\epsilon\delta_{h,h_1}\delta_{h,h_2}\delta(t-t_1)\delta(t-t_2)$, and hence

$$A_{r,r_1,r_2}(t,t_1,t_2) \approx \lambda_0\epsilon \int dt' \sum_h J_{r,h}(t-t') J_{h,r_1'}(t_1-t') J_{h,r_2'}(t_2-t').$$

This is the result we used in our earlier calculation estimating the error we make in the mean firing rates of the recorded neurons by neglecting higher order spike filtering.

K. Tree-level calculation of the effective noise correlations

In our approximation of the model for the recorded neurons, we also neglected fluctuations from the mean input around the hidden neuron input. We should therefore check how strong this noise is. At the level of a mean-field approximation we may neglect it, so we will need to go to a tree-level approximation to calculate it.

The noise is defined by

$$\xi_r(t) = \sum_h \int_{-\infty}^{\infty} dt'\ J_{r,h}(t-t')\Big(\dot{n}_h(t') - \mathbb{E}[\dot{n}_h|\{\dot{n}_r\}]\Big).$$

It has zero mean (by construction), conditioned on the activity of the recorded units — i.e., the "noise" receives feedback from the recorded neurons. We can evaluate the cross-correlation function of this noise, conditioned on the recorded unit activity. This is given by

$$\mathbb{E}[\xi_r(t)\xi_{r'}(t')|\{\dot{n}_r\}]^c = \sum_{h_1,h_2}\int_{-\infty}^{\infty} dt_1 dt_2\ J_{r,h_1}(t-t_1) J_{r',h_2}(t'-t_2)\mathbb{E}[\dot{n}_{h_1}(t_1)\dot{n}_{h_2}(t_2)|\{\dot{n}_r\}]^c,$$

where $\mathbb{E}[\dot{n}_{h_1}(t_1)\dot{n}_{h_2}(t_2)|\{\dot{n}_r\}]^c = \mathbb{E}[\dot{n}_{h_1}(t_1)\dot{n}_{h_2}(t_2)|\{\dot{n}_r\}] - \mathbb{E}[\dot{n}_{h_1}(t_1)|\{\dot{n}_r\}]\mathbb{E}[\dot{n}_{h_2}(t_2)|\{\dot{n}_r\}]$ is the cross-correlation function of the spikes (the superscript c denotes 'cumulant' or 'connected' correlation function to distinguish it from the moments without the superscript). At the level of mean field theory $\mathbb{E}[\dot{n}_{h_1}(t_1)\dot{n}_{h_2}(t_2)|\{\dot{n}_r\}] \approx \mathbb{E}[\dot{n}_{h_1}(t_1)|\{\dot{n}_r\}]\mathbb{E}[\dot{n}_{h_2}(t_2)|\{\dot{n}_r\}]$, and thus the cross-correlation function is zero. We can go beyond mean field theory and calculate the tree-level contribution to the correlation functions using the field theory diagrammatic techniques of [9]. We will do so for the reference state of zero-recorded unit activity, $\{\dot{n}_r\} = \{0\}$, as we expect this to be the leading order contribution to the correlation function. As we are interested primarily in the typical magnitude of the noise compared to the interaction

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

terms, we will work only to leading order in $\epsilon = \exp(\mu_0)$ for the exponential nonlinearity network. We find

$$\begin{aligned}\mathbb{E}[\dot{n}_{h_1}(t_1)\dot{n}_{h_2}(t_2)|0]^c_{\text{tree}} &= \int_{-\infty}^{\infty} dt' \sum_{h'} \Delta_{h_1,h'}(t_1-t')\Delta_{h_2,h'}(t_2-t')\nu_{h'} \\ &\approx \lambda_0 \epsilon \delta_{h_1,h_2}\delta(t_1-t_2),\end{aligned}$$

where $\Delta_{h,h'}(\omega) \approx \delta_{h,h'} + \mathcal{O}(\epsilon)$ is the linear response to perturbations to the *output* of a neuron's rate. It is related to $\Gamma_{h,h'}(\omega)$ by $\Gamma_{h,h'}(\omega) = \Delta_{h,h'}(\omega)\gamma_{h'}$, where $\gamma_h = \nu_h$ for $\phi(x) = e^x$. The overall noise cross-correlation function is then approximately

$$\mathbb{E}[\xi_r(t)\xi_{r'}(t')|0]^c = \lambda_0 \epsilon \sum_h \int_{-\infty}^{\infty} dt_1 \; J_{r,h}(t-t_1)J_{r',h}(t'-t_1).$$

If $r \neq r'$, the expected noise cross-correlation, averaged over the synaptic weights $\mathcal{J}_{i,j}$, is zero. If $r = r'$, the expected value is non-zero. The expected noise auto-correlation function is then

$$\begin{aligned}\overline{\mathbb{E}[\xi_r(t)\xi_r(t')|0]^c} &= \lambda_0 \epsilon N_{\text{hid}} \text{var}(\mathcal{J}) \int_{-\infty}^{\infty} dt_1 \; g(t-t_1)g(t'-t_1) \\ &= \lambda_0 \epsilon (1-f) J_0^2 \frac{1}{(pN)^{2a-1}} \int_{-\infty}^{\infty} dt_1 \; g(t-t_1)g(t'-t_1).\end{aligned}$$

For the specific case of $g(t) = \alpha^2 t e^{-\alpha t}\Theta(t)$, we have

$$\overline{\mathbb{E}[\xi_r(t)\xi_r(t')|0]^c} = \frac{1}{4}\lambda_0 \epsilon (1-f) J_0^2 \frac{1}{(pN)^{2a-1}} \alpha e^{-\alpha|t-t'|}\Big(1 + \alpha|t-t'|\Big).$$

For weak coupling ($a = 1$), the expected autocorrelation function falls off with network size as $1/N$, while for strong coupling ($a = 1/2$), it scales with the fraction of observed neurons f, but is independent of the absolute network size. The overall $\lambda_0\epsilon$ scaling puts the magnitude of the autocorrelation function on par with contributions from hidden paths through a single hidden neuron that contributes a factor of $\lambda_0\epsilon$ to the correction to the coupling filters. Based on our results shown in Fig. 3, which suggest that contributions from long paths through hidden neurons are significant when the fraction of neurons f is small and $J_0 \lesssim 1$, we expect that network noise will also be significant in these regimes. This will not modify the results presented in the main paper, however. It simply means that this noise should be retained in the rate of our approximate model,

$$\lambda_r(t) \approx \lambda_0 \exp\left(\mu_r^{\text{eff}} + \sum_{r'} J_{r,r'}^{\text{eff}} * \dot{n}_{r'}(t) + \xi_r(t)\right).$$

Lastly, we note that here we only calculated the leading order contribution to the noise correlation functions around the reference state of no recorded unit activity. Much as we expanded $\mathbb{E}[\dot{n}_h(t)|\{\dot{n}_r\}]$ in a functional Taylor series around $\{\dot{n}_r\} = 0$, we could do so for the correlation functions as well to study how these effective noise correlations depend on the activity of the recorded neurons. We leave such a calculation for future investigations.

L. Full mean-field reference state

For most of our analyses, we have been expanding the conditional firing rates of the hidden neurons around a reference state of zero activity of the recorded neurons. The quantities ν_h, γ_h, $\hat{\Gamma}_{h,h'}(\omega)$, and so on, are thus calculated using a network in which the recorded neurons have been removed. We have demonstrated that this approximation is valid for the networks considered in this paper. However, this approximation may break down in networks in which the recorded neurons spike at high rates. In this case, we may need another reference state to expand the conditional rates around. A natural choice of reference state $\dot{n}_r^{(0)}(t)$ in this case would be the mean firing rates of the neurons. We will set up this expansion here.

The mean firing rates of the neurons are intractable to calculate exactly, so we will estimate them by the mean field rates, an approximation that we expect to be accurate in the high-firing rate regime.

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The mean field equations for the full network are

$$\langle \dot{n}_i \rangle = \lambda_0 \phi \left(\mu_i + \sum_j J_{ij} * \langle \dot{n}_j \rangle \right).$$

We can then expand $\mathbb{E}\left[\dot{n}_h | \{\dot{n}_r\}\right]$ around $\dot{n}_r = \langle \dot{n}_r \rangle$, truncating at linear order to obtain

$$\mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r\}\right] \approx \langle \dot{n}_h \rangle + \sum_{h',r} \int_{-\infty}^{\infty} dt' dt'' \; \Gamma^{\text{full}}_{h,h'}(t-t') J_{h',r}(t'-t'')(\dot{n}_r(t'') - \langle \dot{n}_r \rangle),$$

where $\Gamma^{\text{full}}_{h,h'}(t-t')$ is the input linear response of the *full network*, including the recorded neurons.

We can then approximate the instantaneous firing rates of the recorded neurons by

$$\begin{aligned}
\lambda_r(t) &\approx \lambda_0 \phi \left(\left\{ \mu_r + \sum_{r'} J_{r,r'} * \langle \dot{n}_r \rangle \right\} + \sum_{r'} J_{r,r'} * (\dot{n}_r - \langle \dot{n}_r \rangle) + \sum_h J_{r,h} * \mathbb{E}\left[\dot{n}_h(t) | \{\dot{n}_r\}\right] \right) \\
&\approx \lambda_0 \phi \left(\left\{ \mu_r + \sum_{r'} J_{r,r'} * \langle \dot{n}_r \rangle + \sum_h J_{r,h} * \langle \dot{n}_h \rangle \right\} + \sum_{r'} \left\{ J_{r,r'} + \sum_{h,h'} J_{r,h} * \Gamma^{\text{full}}_{h,h'} * J_{h',r'} \right\} * (\dot{n}_r - \langle \dot{n}_r \rangle) \right);
\end{aligned}$$

note that we introduced $0 = \sum_{r'} \langle \dot{n}_{r'} \rangle - \sum_{r'} \langle \dot{n}_{r'} \rangle$ so that we could write the instantaneous firing not as a function of filtered spike trains but as a functioned of filtered deviations from the mean firing rate. Importantly, although it looks like only the baseline is different from the zero-activity reference state case but the coupling is the same, the linear response function $\Gamma^{\text{full}}_{h,h'}(\tau)$ is *not* the same as the zero-reference state case, and hence the correction to the coupling is slightly different. The solutions look similar, but the linear response functions now incorporate the effects of the recorded units as well. In particular, $\Gamma^{\text{full}}_{ij}(t-t')$ satisfies the equation

$$\int_{-\infty}^{\infty} dt'' \left[\delta_{ik} - \gamma^{\text{full}}_i J_{ik}(t-t'') \right] \Gamma^{\text{full}}_{kj}(t''-t') = \gamma^{\text{full}}_i \delta_{ij} \delta(t-t'),$$

where γ^{full}_i is the gain of neuron i accounting for the entire network,

$$\gamma^{\text{full}}_i = \lambda_0 \phi' \left(\mu_i + \sum_j J_{ij} * \langle \dot{n}_j \rangle \right).$$

Thus, in Fourier space

$$\begin{aligned}
\hat{\Gamma}^{\text{full}}_{ij}(\omega) &= \left[\mathbb{I} - \hat{\mathbf{V}}^{\text{full}}(\omega) \right]^{-1}_{ij} \langle \dot{n}_j \rangle \\
&= \sum_{\ell=0}^{\infty} \left[\hat{\mathbf{V}}^{\text{full}}(\omega)) \right]^{\ell}_{ij} \langle \dot{n}_j \rangle,
\end{aligned}$$

where $\hat{V}^{\text{full}}_{i,j}(\omega) = \gamma^{\text{full}}_i \hat{J}_{i,j}(\omega)$ is an $N \times N$ matrix – i.e., it contains the couplings and firing rates of *all* neurons, recorded and hidden. Hence, while this looks formally similar to the result we obtained in the zero activity state, the inclusion of recorded neurons modifies our rules for calculating contributions to the effective coupling filters. In particular, $\hat{J}^{\text{eff}}_{r,r'}(\omega) - \hat{J}_{r,r'}(\omega)$ involves contributions from paths through both hidden and recorded neurons, unlike the zero-activity reference case, which involved contributions only from paths through hidden neurons. The reason for this, of course, is that the reference state depends on the entire network, not just the hidden neurons. The difference between the cases matters only at higher orders in our expansion — paths of length $\ell = 4$ or greater. We can see this

bioRxiv preprint doi: https://doi.org/10.1101/105510; this version posted February 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by writing out the first few terms in the path length expansion of the effective coupling,

$$\hat{J}^{\text{eff}}_{r,r'}(\omega) = \hat{J}_{r,r'}(\omega) + \sum_h \hat{J}_{r,h}(\omega)\gamma^{\text{full}}_h \hat{J}_{h,r}(\omega) + \sum_{h,h'} \hat{J}_{r,h}(\omega)\gamma^{\text{full}}_h \hat{J}_{h,h'}(\omega)\gamma^{\text{full}}_{h'} \hat{J}_{h',r'}(\omega)$$
$$+ \sum_{h,h',j} \hat{J}_{r,h}(\omega)\gamma^{\text{full}}_h \hat{J}_{h,j}(\omega)\gamma^{\text{full}}_j \hat{J}_{j,h'}(\omega)\gamma^{\text{full}}_{h'} \hat{J}_{h',r'}(\omega) + \ldots;$$

for conciseness, we have assumed zero-self coupling ($\hat{J}_{i,i}(\omega) = 0$), but this could be restored by setting $\gamma^{\text{full}}_i \to \gamma^{\text{full}}_i/(1 - \gamma^{\text{full}}_i \hat{J}_{i,i}(\omega))$.

We see that the first few terms of the expansion are the same as the zero-activity reference case, with the exception that the γ^{full}_h are the gains for the entire network, not just the hidden network absent the recorded neurons. It is only the $\ell = 4$ term at which contributions to the linear response functions involving paths through any neuron j, recorded or hidden, start to appear. Because we typically expect the amplitude of these terms to be small, we anticipate expanding around the mean field reference state will yield similar results to the expansion around the zero-activity reference state presented in the main paper.

[1] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From single neurons to networks and models of cognition* (Cambridge University Press, Cambridge, U.K., 2014).
[2] S. Ostojic and N. Brunel, PLOS Computational Biology **7**, 1 (2011).
[3] H. F. Song and X.-J. Wang, Phys. Rev. E **90**, 062801 (2014).
[4] E. Simoncelli, L. Paninski, J. Pillow, and O. Schwartz, in *The Cognitive Neurosciences*, edited by M. Gazzaniga (MIT Press, 2004) 3rd ed., pp. 327–338.
[5] L. Paninski, Network: Computation in Neural Systems **15**, 243 (2004).
[6] J. Pillow, L. Paninski, V. J. Uzzell, E. Simoncelli, and E. J. Chichilnisky, The Journal of neuroscience : the official journal of the Society for Neuroscience **25**, 11003 (2005).
[7] J. Pillow, J. Shlens, L. Paninski, A. Sher, A. Litke, E. Chichilnisky, and E. Simoncelli, Nature **454**, 995 (2008).
[8] R. A. Horn and C. R. Johnson, eds., *Matrix Analysis* (Cambridge University Press, New York, NY, USA, 1986).
[9] G. Koch Ocker, K. Josić, E. Shea-Brown, and M. A. Buice, ArXiv e-prints (2016), arXiv:1610.03828 [q-bio.NC].
[10] A. G. HAWKES, Biometrika **58**, 83 (1971).
[11] P. Brémaud and L. Massoulié, Ann. Probab. **24**, 1563 (1996).