PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Predicting how and when hidden neurons skew measured synaptic interactions

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

1 Department of Applied Mathematics, University of Washington, Seattle, Washington, United States of America, **2** Department of Physiology and Biophysics, University of Washington, Seattle, Washington, United States of America, **3** Graduate Program in Neuroscience, University of Washington, Seattle, Washington, United States of America, **4** Allen Institute for Brain Science, Seattle, Washington, United States of America

¤ Current address: Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, New York, United States of America
* braden.brinkman@stonybrook.edu

Abstract

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 and determine how biological structure gives rise to neural circuit function, 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 pre-synaptic 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 particular example of interest, we derive a formula for the impact of hidden units in random networks with "strong" coupling—connection weights that scale with $1/\sqrt{N}$, where N is the network size, precisely the scaling observed in recent experiments. With this quantitative relationship between measured and true interactions, we can study how network properties shape effective interactions, which properties are relevant for neural computations, and how to manipulate effective interactions.

Author summary

No experiment in neuroscience can record from more than a tiny fraction of the total number of neurons present in a circuit. This severely complicates measurement of a network's true properties, as unobserved neurons skew measurements away from what would be measured if all neurons were observed. For example, the measured post-synaptic response of a neuron to a spike from a particular pre-synaptic neuron incorporates direct

Citation: Brinkman BAW, Rieke F, Shea-Brown E, Buice MA (2018) Predicting how and when hidden neurons skew measured synaptic interactions. PLoS Comput Biol 14(10): e1006490. https://doi.org/10.1371/journal.pcbi.1006490

Editor: Jeff Beck, Duke University, UNITED STATES

Received: December 7, 2017

Accepted: September 5, 2018

Published: October 22, 2018

Data Availability Statement: Scripts and data files used to produce figures are available online at https://github.com/bradenbrinkman/hiddenneurons.

Funding: Support provided by the Sackler Scholar Program in Integrative Biophysics (BAWB), Collaborative Research in Computational Neuroscience (CRCNS, https://crcns.org/) grant DMS-1208027 (ESB, FR), National Institute of Health (NIH, https://www.nih.gov/) grant EY11850 and the Howard Hughes Medical Institute (http://www.hhmi.org) (FR). This work was partially based on work supported by the Center for Brains, Minds and Machines (CBMM, https://cbmm.mit.edu/), funded by National Science Foundation (NSF, https://www.nsf.gov) 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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

connections between the two neurons as well as the effect of any number of indirect connections, including through unobserved neurons. To understand how measured quantities are distorted by unobserved neurons, we calculate a general relationship between measured "effective" synaptic interactions and the ground-truth interactions in the network. This allows us to identify conditions under which hidden neurons substantially alter measured interactions. Moreover, it provides a foundation for future work on manipulating effective interactions between neurons to better understand and potentially alter circuit function—or dysfunction.

Introduction

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 contribute significantly to circuit dysfunction [1–15].

A major obstacle to uncovering structure-function relationships is the fact that most experiments can only directly observe small fractions of an active network. State-of-the-art methods for determining connections between neurons in living networks infer them by fitting statistical models to neural spiking data [16–25]. However, the fact that we cannot observe all neurons in a network means that the statistically inferred connections are "effective" connections, representing some dynamical relationship between the activity of nodes but not necessarily a true physical connection [24–33]. Intuitively, reverberations through the network must contribute to these effective interactions; our goal in this work is to formalize this intuition and establish a quantitative relationship between measured effective interactions and the true synaptic interactions between neurons. With such a relationship in hand we can study the effective interactions generated by different choices of synaptic properties and circuit architectures, allowing us to not only improve interpretation of experimental measurements but also probe how circuit structure is tied to function.

The intuitive relationship between measured and effective interactions is demonstrated schematically in Fig 1. Fig 1A demonstrates that in a fully-sampled network the directed interactions between neurons—here, the change in membrane potential of the post-synaptic neuron after it receives a spike from the pre-synaptic neuron—can be measured directly, as observation of the complete population means different inputs to a neuron are not conflated. However, as shown in Fig 1B, the vastly more realistic scenario is that the recorded neurons are part of a larger network in which many neurons are unobserved or "hidden." The response of the post-synaptic neuron 2 to a spike from pre-synaptic neuron 1 is a combination of both the direct response to neuron 1's input as well as input from the hidden network driven by neuron 1's spiking. Thus, the measured membrane response of neuron 2 due to a spike fired by neuron 1—which we term the "effective interaction" from neuron 1 to 2—may be quite different from the true interaction. It is well-known that circuit connections between recorded neurons, as drawn in Fig 1C, are at best effective circuits that encapsulate the effects of unobserved neurons, but are not necessarily indicative of the true circuit architecture. The formalized relationship we will establish in the Results is given in Fig 2.

Even once we establish a relationship between the effective and true connections, we will in general not be able to use measurements of effective interactions to extrapolate back to a unique set of true connections; at best, we may be able to characterize some of the statistical properties of the full network. The obstacle is that several different networks—different both

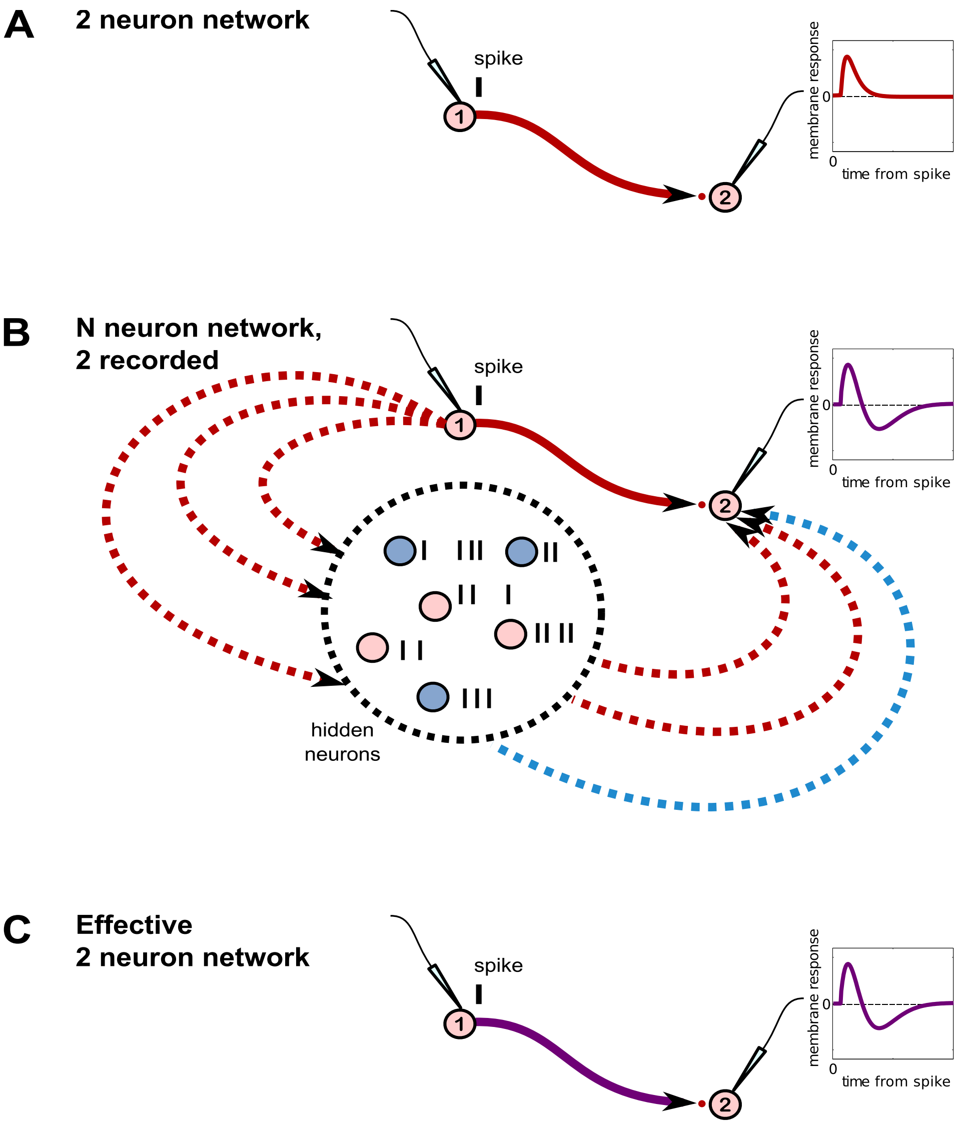

Fig 1. The hidden unit problem. **A**. In a hypothetical circuit consisting of just two recorded neurons (no hidden neurons), we can measure the strength and time course of the directed interactions between neurons by measuring the response of the post-synaptic neuron's membrane potential to a spike from the pre-synaptic neuron. **B**. Realistically, there are many more neurons in the network that are unrecorded and hence "hidden." In this schematic, only two neurons are observed. The hidden neurons are driven by input from the presynaptic neuron labeled 1, and provide input to the recorded post-synaptic neuron labeled 2. Because the activity of the hidden neurons is not controlled, the membrane response reflects a combination of neuron 1's direct influence on neuron 2 and its indirect influence through the hidden network. **C**. The "effective" 2 neuron network observed experimentally.

https://doi.org/10.1371/journal.pcbi.1006490.g001

in terms of architecture and intrinsic neural properties—may give rise to the same network behaviors, a theme of much focus in the neuroscience literature [34–39]. That is, inferring the connections and intrinsic neural properties in a full network from activity recordings from a subset of neurons is in general an ill-posed problem, possessing several degenerate solutions. Several statistical inference methods have been constructed to attempt to infer the presence of, and connections to, hidden neurons [28, 40–42]; the subset of the degenerate solutions that each of these methods finds will depend on the particular assumptions of the inference method (e.g., the regularization penalties applied). As an example, we demonstrate two small circuit motifs that give rise to nearly identical effective interactions, despite crucial differences between the circuits (Figs 3 and 4).

Understanding the effect of hidden neurons on small circuit motifs is only a piece of the hidden neuron puzzle, and a full understanding necessitates scaling up to large circuits containing many different motifs. Having an analytic relationship between true and effective interactions greatly facilitates such analyses by directly studying the structure of the relationship itself, rather than trying to extract insight indirectly through simulations. In particular, in going to large networks we focus on the degree to which hidden neurons skew measured interactions (Fig 5), and how we can predict the features of effective interactions we expect to measure when recording from only a subset of neurons in a network with hypothesized true interactions (Fig 6).

Establishing a theoretical relationship between measured and "true" interactions will thus enable us to study how one can alter the network properties to reshape the effective interactions, and will be of immediate importance not only for interpreting experimental measurements of synaptic interactions, but for elucidating their role in neural coding. Moreover, understanding how to shape effective interactions between neurons may yield new avenues for altering, in a principled way, the computations performed by a network, which could have applications for treating neurological diseases caused in part by pathological synaptic interactions.

Results

Overview

Our goal is to derive a relationship between the effective synaptic interactions between recorded neurons and the true synaptic interactions that would be obtained if the network were fully observed. This makes explicit how the synaptic interactions between neurons are modified by unobserved neurons in the network, and under what conditions these modifications are—or are not—significant. We derive this result first, using a probabilistic model of network activity in which all properties are known. We then build intuition by applying our result to two simple networks: a 3-neuron feedforward-inhibition circuit in which we are able to qualitatively reproduce measurements by Pouille and Scanziani [43], and a 4-neuron circuit that demonstrates how degeneracies in hidden networks are handled within our framework.

To extend our intuition to larger networks, we then study the effective interactions that would be observed in sparse random networks with N cells and strong synaptic weights that scale as $1/\sqrt{N}$ [44–47], as has been recently observed experimentally [48]. We show how unobserved neurons significantly reshape the effective synaptic interactions away from the ground-truth interactions. This is not the case with "classical" synaptic scaling, in which synaptic strengths are inversely proportional to the number of inputs they receive (assumed $\mathcal{O}(N)$), as we will also show. (The case of classical scaling has also been studied previously using a different approach in [49–52]).

Model

We model the full network of N neurons as a nonlinear Hawkes process [53], often referred to as a "Generalized linear (point process) model" in neuroscience, and broadly used to fit neural activity data [16–23, 54]. Here we use it as a generative model for network activity, as it approximates common spiking models such as leaky integrate and fire systems driven by noisy inputs [55, 56], and is equivalent to current-based leaky integrate-and-fire models with soft-threshold (stochastic) spiking dynamics (see Methods).

To derive an approximate model for an observed subset of the network, we partition the network into recorded neurons (labeled by indices r) and hidden neurons (labeled by indices h). Each recorded neuron has an instantaneous firing rate $\lambda_r(t)$ such that the probability that the neuron fires within a small time window $[t, t + dt]$ is $\lambda_r(t)dt$, when conditioned on the inputs to the neuron. The instantaneous firing rate in our model is

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

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 synaptic interaction (or "spike filter") $J_{i,j}(t)$ with spike train $\dot{n}_j(t)$ *from* pre-synaptic neuron j *to* post-synaptic neuron i, for neural indices i and j that may be either recorded or hidden. In this work we take the tonic drive to be constant in time, and focus on the steady-state network activity in response to this drive. We consider interactions 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. We include self-couplings $J_{i,i}(t)$ not to represent autapses, but to account for intrinsic neural properties such as refractory periods ($\mathcal{J}_{i,i} < 0$) or burstiness ($\mathcal{J}_{i,i} > 0$). 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 calculating the effective membrane response of the recorded neurons by averaging over the activity of the hidden neurons *conditioned on the activity of the recorded neurons*. In practice this is intractable to perform exactly [57–60]. Here, we use a mean field approximation to calculate the mean input from the hidden neurons (again, conditioned on the activity of the recorded neurons). The value of deriving such a relationship analytically, as opposed to simply numerically determining the effective interactions, is that the resulting expression will give us insight into how the effective interactions decompose into contributions of different network features, how tuning particular features shapes the effective interactions, and conditions under which we expect hidden units to skew our measurements of connectivity in large partially observed networks.

As shown in detail in the Methods, the instantaneous firing rates of the recorded neurons can then be approximated as

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

The effective baselines $\mu_r^{\text{eff}} = \mu_r + \sum_h \mathcal{J}_{r,h} \nu_h$, are simply modulated by the net tonic input to the neuron, so we do not focus on them here. The $\xi_r(t)$ are effective noise sources arising from fluctuation input from the hidden network. At the level of our mean field approximation these fluctuations vanish; corrections to the mean field approximation are straightforward and yield non-zero noise correlations, but will not impact our calculation of the effective interactions (see the Methods and SI), so as with the effective baselines we will not focus on the effective noise here.

The effective coupling filters are given in the frequency domain by

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

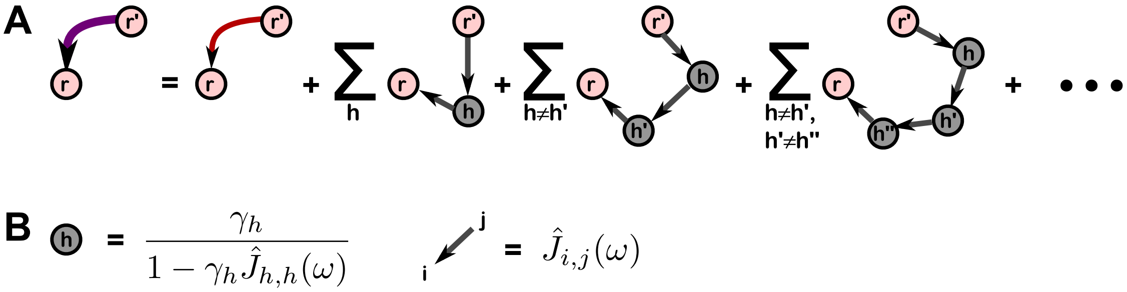
A
r'
r
=
r'
r
+ $\sum_{h}$
r'
r
h
+ $\sum_{h \neq h'}$
r'
r
h
h'
+ $\sum_{h \neq h', h' \neq h''}$
r'
r
h
h'
h''
+ • • •
B
h
= $\dfrac{\gamma_h}{1-\gamma_h \hat{J}_{h,h}(\omega)}$
j
i
= $\hat{J}_{i,j}(\omega)$

Fig 2. Expansion of effective interactions into contributions from hidden paths. A. Graphical representation of Eq (2). The linear response of the hidden network, $\hat{\Gamma}_{h,h'}(\omega)$, has been expanded as a series (corresponding to the grey hidden nodes and links between them), such that each term in the overall series can be interpreted as a contribution from a path through which the pre-synaptic neuron r' is able to send a signal to post-synaptic neuron r via 1, 2, etc. hidden neurons. This expression holds for any pair of neurons in the recorded subset. **B**. Quantitative expressions for each diagram in the series can be read off by assigning the shown factors for each hidden neuron node and each link between neurons, recorded or hidden, and multiplying them together. (No factor is assigned to the recorded neuron nodes). γ_h is the gain of neuron h and $\hat{J}_{i,j}(\omega)$ is the true interaction from j to i in the frequency domain.

https://doi.org/10.1371/journal.pcbi.1006490.g002

These results hold for any pair of recorded neurons r' and r, and any choice of network parameters for which the mean field steady state of the hidden network exists. 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*. That is, $\Gamma_{h,h'}(t-t')$ is the linear response of hidden neuron h at time t due to a perturbation to the input of neuron h' at time t', and incorporates the effects of h' propagating its signal to h through other hidden neurons, as demonstrated graphically in Fig 2. Both ν_h and $\hat{\Gamma}_{h,h'}(\omega)$ are calculated *in the absence of the recorded neurons*. 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. For details on the derivations and justification of approximations, see the Methods and Supporting Information (SI).

The effective coupling filters are what we would—in principle—measure experimentally if we observe only a subset of a network, for example by pairwise recordings shown schematically in Fig 1. For larger sets of recorded neurons, interactions between neurons are typically inferred using statistical methods, an extremely nontrivial task [16–23, 28, 40, 41], and details of the fitting procedure could potentially further skew the inferred interactions away from what would be measured by controlled pairwise recordings. We will put aside these complications here, and assume we have access to an inference procedure that allows us to measure $J^{\mathrm{eff}}_{r,r'}(t)$ 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 $\hat{J}_{r,r'}(\omega)$ (as defined in Eq (1)) 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 through which neuron r' is able to send a signal to neuron r *via hidden neurons only*.

We may write down a set of "Feynmanesque" graphical rules for explicitly calculating terms in this series [53]. First, we define the input-output gain of a hidden neuron h, $\gamma_h \equiv \lambda_0\phi'(\mu_h + \sum_{h'}\mathcal{J}_{h,h'}\nu_{h'})$, calculated in the absence of recorded neurons. The contribution of each term can then be written down using the following rules, shown graphically in Fig 2: *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.

The graphical expansion is reminiscent of recent works expanding correlation functions of linear models of network spiking in terms of network "motifs" [61–63]. Computationally, this expression is practical for calculating the effective interactions in small networks involving only a few hidden neurons (as in the next section), but is generally unwieldy for large networks. In practice, for moderately large networks the linear response matrix $\hat{\Gamma}_{h,h'}(\omega)$ can be calculated directly by numerical matrix inversion and an inverse Fourier transform back into the time domain. 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. For example, one immediate insight the path decomposition offers is that neurons only develop effective interactions between one another if there is a path by which one neuron can send a signal to the other.

Feedforward inhibition and degeneracy of hidden networks in small circuits

Effective interactions & emergent timescales in a small circuit. To build intuition for our result and compare to a well-known circuit phenomenon, we apply our Eq (2) to a 3-neuron circuit implementing feedforward inhibition, like that studied by Pouille and Scanziani [43]. Feedforward inhibition can sharpen the temporal precision of neural coding by narrowing the "window of opportunity" in which a neuron is likely to fire. For example, in the circuit shown in Fig 3A, excitatory neuron 1 projects to both neurons 2 and 3, and 3 projects to 2. Neuron 1 drives both 2 and 3 to fire more, while neuron 3 is inhibitory and will counteract the drive neuron 2 receives from 1. The window of opportunity can be understood by looking at the effective interaction between neurons 1 and 2, treating neuron 3 as hidden. We use our path expansion (Fig 2) to quickly write down the effective interaction we expect to measure in the frequency domain,

$$\hat{J}^{\text{eff}}_{2,1}(\omega) = \hat{J}_{2,1}(\omega) + \frac{\hat{J}_{2,3}(\omega)\gamma_3\hat{J}_{3,1}(\omega)}{1-\gamma_3\hat{J}_{3,3}(\omega)}. \tag{3}$$

The corresponding true synaptic interactions and resulting effective interaction are shown in Fig 3B. The effective interaction matches qualitatively the observed changes measured by Pouille and Scanziani [43], and shows a narrow window after neuron 2 receives a spike in which the change in membrane potential is depolarized and neuron 2 is more likely to fire. Following this brief window, the membrane potential is hyperpolarized and the cell is less likely to fire until it receives more excitatory input.

The effective interaction from neuron 1 to 2 in this simple circuit also displays several features that emerge in more complex circuits. Firstly, although the true interactions are either excitatory (positive) or inhibitory (negative), the effective interaction has a mixed character, being initially excitatory (due to excitatory inputs from neuron 1 arriving first through the monosynaptic pathway), but then becoming inhibitory (due to inhibitory input arriving from the disynaptic pathway).

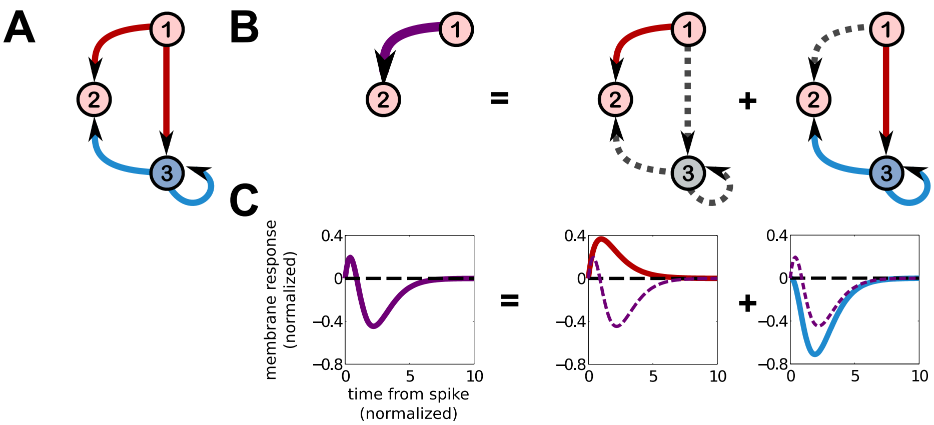

Fig 3. 3 neuron feedforward inhibition circuit. A: A 3-neuron circuit displaying feedforward inhibition. Neuron 1 provides excitatory input to neurons 2 and 3, while neuron 3 provides inhibitory input to neuron 2. Neuron 3 also has a self-history coupling, denoted by an autaptic loop, which implements a refractory period in this circuit model. **B**: Leftmost, the effective interaction from neuron 1 to 2 when neuron 3 is unobserved. Subsequent plots decompose this interaction into contributions from neuron 1's direct input to neuron 2, and its indirect input through neuron 3. The indirect input through neuron 3 also takes account of neuron 3's self-history interaction. **C**. Leftmost, the effective interaction (membrane response) from neuron 1 to 2, subsequently decomposed into contributions from the direct interaction and the indirect interaction from 1 to 2.

https://doi.org/10.1371/journal.pcbi.1006490.g003

Secondly, emergent timescales develop due to reverberations between hidden neurons with bi-directional connections, represented as loops between neurons in our circuit schematics (e.g., between neurons 3 and 4 in Fig 4). This includes self-history interactions such as refractoriness, schematically represented by loops like the 3 → 3 loop shown in Fig 3, corresponding to the factor $1/\left(1-\gamma_3\hat{J}_{3,3}(\omega)\right)$. In the particular example shown in Fig 3, in which we use a self-history interaction $J_{33}(\tau) = \mathcal{J}_{33}\beta_{33}\exp(-\beta_{33}\tau)$, a new timescale $\beta_{33}^{-1}(1-\gamma_3\mathcal{J}_{33})^{-1}$ develops. Other choices of interactions can generate more complicated emergent timescales and temporal dynamics, including oscillations. For example, in the 4-neuron circuit discussed below (Fig 4), the choice $J_{3,4}(\tau) = J_{4,3}(\tau) = -|\mathcal{J}|\alpha^2\tau e^{-\alpha\tau}$ yields effective interactions with new decay and oscillatory timescales equal to $\left(\alpha\left(1-\lambda_0|\mathcal{J}|\right)\right)^{-1}$ and $\left(\alpha\lambda_0|\mathcal{J}|\right)^{-1}$. In the larger networks we consider in the next section, inter-neuron interactions must scale with network size in order to maintain network stability. Because emergent timescales depend on the synaptic strengths of hidden neurons, we typically expect emergent timescales generated by loops between hidden neurons to be negligible in large random networks. However, because the magnitudes of the self-history interaction strengths need not scale with network size, they may generate emergent timescales large enough to be detected.

It is worth noting explicitly that only the interaction from neuron 1 to 2 has been modified by the presence of the hidden neuron 3, for the particular wiring diagram shown in Fig 3. The self-history interactions of both neurons 1 and 2, as well as the interaction from neuron 2 to 1 (zero in this case) are unmodified. The reason the hidden neuron did not modify these interactions is that the only link neuron 3 makes is from 1 to 2. There is no path by which neuron 1 can send a signal back to itself, hence its self-interaction is unmodified, nor is there a path that neuron 2 can send signals to neuron 3 or on to neuron 1, and hence neuron 2's self-history interaction and its interaction to neuron 1 are unmodified.

Degeneracy of hidden networks giving rise to effective interactions. It is well known that different networks may produce the same observed circuit phenomena [34–39]. To

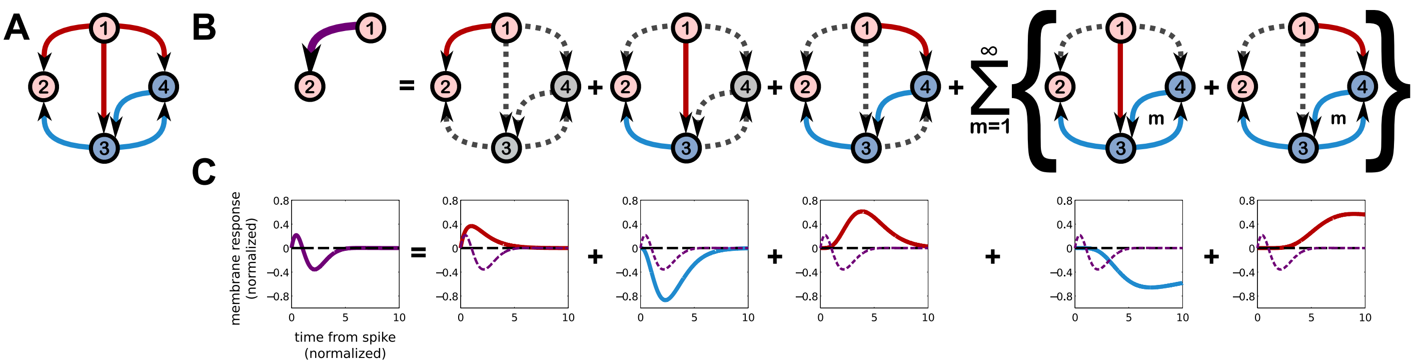

Fig 4. Different complete circuits may underly similar effective circuits. **A**: A circuit very similar to that in Fig 3, except that neuron 1 also provides excitatory input to neuron 4, which in turn provides inhibitory input to neuron 3. The self-history coupling of neuron 3 to itself has also been removed in this example. **B**: Leftmost, the effective interaction from neuron 1 to 2, which is qualitatively and quantitatively similar to the effective interaction shown in Fig 3. Subsequent plots indicate each path through the circuit that neuron 1 can send a signal to neuron 2 through the hidden neurons 3 and 4. **C**. Leftmost, the effective interaction from neuron 1 to 2. Subsequent plots decompose this interaction into contributions from the paths shown above in **B**.

https://doi.org/10.1371/journal.pcbi.1006490.g004

illustrate that our approach may be used to identify degenerate solutions in which more than one network underlies observed effective interactions, we construct a 4-neuron circuit that produces a quantitatively similar effective interaction between the recorded neurons 1 and 2, shown in Fig 4. Specifically, in this circuit we have removed neuron 3's self-history interaction and introduced a second inhibitory hidden network that receives excitatory input from neuron 1 and provides inhibitory input to neuron 3. By tuning the interaction strengths we are able to produce the desired effective interaction. This demonstrates that intrinsic neural properties such as refractoriness can trade off against inputs from other hidden neurons, making it difficult to distinguish the two cases from one another (or from a potential infinity of other circuits that could have produced this interaction; for example, a qualitatively similar interaction is produced in the $N = 1000$ network in which only three neurons are recorded, shown below in Fig 6). Statistical inference methods may favor one of the possible underlying choices of complete network consistent with a measured set of effective interactions, suggesting there may be some sense of a "best" solution. However, the particular "best" network will depend on many factors, including the amount and fidelity of data recorded, regularization choices, and how well the fitted model generalizes to new data (i.e., how "close" the fitted model is to the generative model). Potentially, if these conditions were met, with enough data the slight quantitative differences between the effective interactions produced by different hidden networks (including higher order effective interactions, which we assume to be negligible here; see SI), could help distinguish different hidden networks. However, the amount of data required to perform this discrimination and validate the result may be impractically large [36, 64–66]. It is thus worth studying the structure of the observed effective interactions directly in search of possible signatures that elucidate the statistical properties of the complete network.

Strongly coupled large networks

Constructing networks that produce particular effective interactions is tractable for small circuits, but much more difficult for larger circuits composed of many circuit motifs. Not only can combinations of different circuit motifs interact in unexpected ways, one must also take care to ensure the resulting network is both active and stable—i.e., that firing will neither die out nor skyrocket to the maximum rate. Stability in networks is often implemented by either building networks with classical (or "weak") synapses whose strength scales inversely with the number of inputs they receive, assumed here to be proportional to network size, and hence $\mathcal{J}_{i,j} \sim 1/N$, or by building balanced networks in which excitatory and inhibitory synaptic strengths balance out, on average, and scale as $\mathcal{J}_{i,j} \sim 1/\sqrt{N}$ [44, 48] (but note the distinction that we use a "soft threshold" firing model with nonlinearity that is fixed as N varies, whereas previous work has typically used hard threshold models). In both cases the synapses tend to be small in value in large networks, but are compensated for by large numbers of incoming connections. In the case of $1/N$ scaling, neurons are driven primarily by the mean of their inputs, while in "strong" balanced $1/\sqrt{N}$ networks neurons are driven primarily by fluctuations in their inputs.

Our goal is to understand how the interplay between the presence of hidden neurons and different synaptic scaling or network architectures shapes effective interactions. Previous work has studied the hidden-neuron problem in the weak coupling limit [49–52] using a different approach to relate inferred synaptic parameters to true parameters; here we use our approach to study the $1/\sqrt{N}$ strong coupling limit, theoretically predicted to be an important feature that supports computations in networks in a balanced regime [44–47]. Moreover, experiments in cultured neural tissue have been found to be more consistent with the $1/\sqrt{N}$ scaling than $1/N$ [48], indicating that it may have intrinsic physiological importance.

We analytically determine how significantly effective interaction strengths are skewed away from the true interaction strengths as a function of both the number of observed neurons and typical synaptic strength. We consider several simple networks ubiquitous in neural modeling: first, an Erdős-Réyni (ER) network with "mixed synapses" (i.e., a neuron may have both positive and negative synaptic weights), a balanced ER network with Dale's law imposed (a neuron's synapses are all the same sign), and a Watts-Strogatz (WS) small world network with mixed synapses. Each network has N neurons and connection sparsity p (only $100p\%$ of connections are non-zero). Connections in ER networks are chosen randomly and independently, while connections in the WS network are determined by randomly rewiring a fraction β of the connections in a $(pN)^{\text{th}}$-nearest-neighbor ring network. such that the overall network has a backbone of local synaptic connections with a web of sparse long-range connections. In each network N_{rec} neurons are recorded randomly.

For simplicity we take the baselines of all neurons to be equal, $\mu_i = \mu_0$ (such that in the absence of synaptic input the probability that a neuron fires in a short time window Δt is $\lambda_0 \Delta t \exp(\mu_0)$). We choose the rate nonlinearity to be exponential, $\phi(x) = e^x$; this is the "canonical" choice of nonlinearity often used when fitting this model to data [16–18, 20, 67]. 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 $\mathcal{J}_{i,j}$ are independently drawn from a normal distribution with zero mean and standard deviation $J_0/(pN)^a$, where J_0 controls the overall strength of the weights and $a = 1$ or $1/2$, corresponding to "weak" and "strong" coupling. For simplicity, we do not consider intrinsic self-coupling effects in this part of the analysis, i.e., we take $\mathcal{J}_{i,i} = 0$ for all neurons i. For the Dale's law network, the overall distribution of synaptic weights follows the same normal distribution as the mixed synapse networks, but the signs of the weights correspond to whether the pre-synaptic neuron is excitatory or inhibitory. Neurons are randomly chosen to be excitatory and inhibitory, the average number of each type being equal so that the network is balanced. Numerical values of all parameters are given in Table 1.

We seek to assess how the presence of hidden neurons can shape measured network interactions. We first focus on the typical strength of the effective interactions as a function of both the fraction of neurons recorded, $f = N_{\text{rec}}/N$, and the strength of the synaptic weights J_0. We quantify the strength of the effective interactions by defining the effective synaptic weights $\mathcal{J}^{\text{eff}}_{r,r'} \equiv \int_0^\infty d\tau\ J^{\text{eff}}_{r,r'}(\tau) = \hat{J}^{\text{eff}}_{r,r'}(\omega = 0)$; c.f. $\mathcal{J}_{r,r'} = \int_0^\infty d\tau\ J_{r,r'}(\tau)$ for the true synaptic weights. We then study the sample statistics of the difference, $\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}$, averaged across both subsets of recorded neurons and network instantiations, to estimate the typical contribution of hidden neurons to the measured interactions. The mean of the synaptic weights is near zero (because the weights are normally distributed with zero mean in the mixed synapse networks and due to balance of excitatory and inhibitory neurons in the Dale's law network), so we focus on the root-mean-square of $\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}$. This measure is a conservative estimate of changes in strength, as $J^{\text{eff}}_{r,r'}(\tau)$ may have both positive and negative components that partially cancel when integrated over time, unlike $J_{r,r'}(\tau)$. An alternative measure we could have chosen that avoids potential cancellations is $\int_0^\infty d\tau\ |J^{\text{eff}}_{r,r'}(\tau) - J_{r,r'}(\tau)|$, i.e., the integrated absolute difference between effective and true interactions. However, this will depend on our specific choices of waveform $g(\tau)$ in our definition $J_{i,j}(\tau) = \mathcal{J}_{i,j} g(\tau)$, whereas $\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}$ does not due to our normalization $\int_0^\infty d\tau\ g(\tau) = 1$. As $|\int d\tau f(\tau)| \leq \int d\tau\ |f(\tau)|$, for any $f(\tau)$, we can consider our choice of $\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}$ as a lower bound on the strength that would be quantified by $\int_0^\infty d\tau\ |J^{\text{eff}}_{r,r'}(\tau) - J_{r,r'}(\tau)|$.

Numerical evaluations of the population statistics for all three network types are shown as solid curves in Fig (5), for both strong coupling and weak coupling. All three networks yield qualitatively similar results. The vertical axes measure the root-mean-square deviations between the statistically expected true synaptic $\mathcal{J}_{r,r'}$ and the corresponding effective synaptic weight $\mathcal{J}^{\text{eff}}_{r,r'}$, normalized by the true root mean square of $\mathcal{J}_{r,r'}$. Thus, a ratio of 0.5 corresponds to a 50% root-mean-square difference in effective versus true synaptic strength. We measure these ratios as a function of both the fraction of neurons recorded (horizontal axis) and the parameter J_0 (labeled curves).

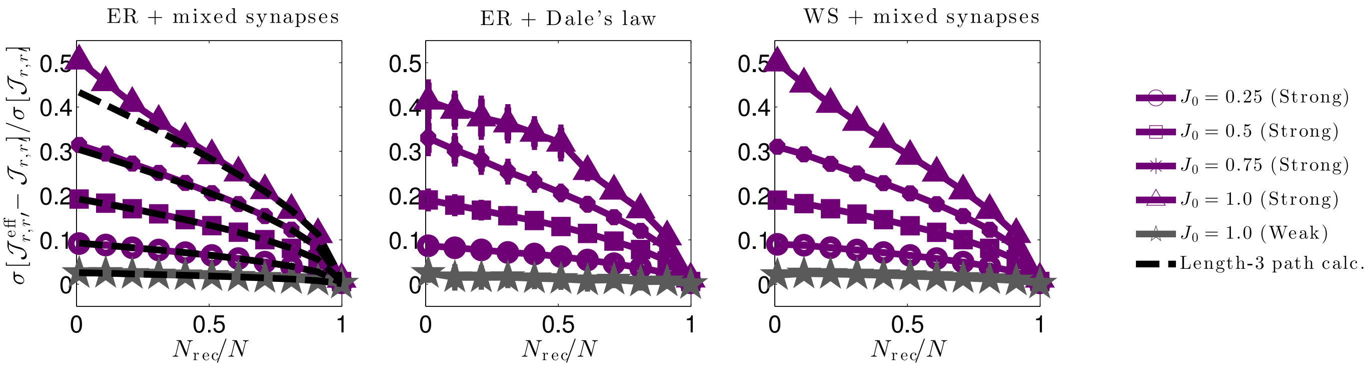

Fig 5. Relative changes in interaction strength due to hidden neurons for three network types. We quantify relative changes in interaction strength between effective ($\mathcal{J}^{\text{eff}}_{r,r'}$) and true ($\mathcal{J}_{r,r'}$) interactions by the (sample) root-square-mean deviation, $\sigma[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}]$, normalized by the true synaptic weight (sample) standard deviation $\sigma[\mathcal{J}_{r,r'}]$. We do so for three (sparse) network types: **Left**. An Erdős-Réyni (ER) network with "mixed synapses" (i.e., Dale's law not imposed) with normally distributed synaptic weights. **Middle**. An ER network with Dale's law imposed, (i.e., each neuron's outgoing synaptic weights all have the same sign). **Right**. A Watts-Strogatz (WS) small world network with 30% rewired connections and mixed synapses. All network types yield qualitatively similar results. In each plot solid lines are numerical estimates of the sample standard deviation of the difference between effective coupling weights $\mathcal{J}^{\text{eff}}_{r,r'}$ 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. Purple lines correspond to synaptic weights with standard deviation $J_0/\sqrt{pN}$ (strong coupling), while grey lines correspond to synaptic weights with standard deviation J_0/pN (weak coupling). For weak $1/N$ coupling (grey), the ratio of standard deviations is $\mathcal{O}(1/\sqrt{N})$. For strong $1/\sqrt{N}$ coupling (purple) the ratio is $\mathcal{O}(1)$ and grows in strength as the fraction of recorded neurons N_{rec}/N decreases or the typical synaptic strength J_0 increases. The dashed black lines in the left plot show theoretical estimates accounting only for hidden paths of length-3 connecting recorded neurons (Eq (4)). 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.

https://doi.org/10.1371/journal.pcbi.1006490.g005

Table 1. Network connectivity parameter values for Figs 5–8 and S1–S3. See individual captions for other figures.

Number of neurons N	**1000**
Number of hidden neurons N_{hid}	{1, 90, 190, 290, 390, 490, 590, 690, 790, 890, 990}
Number of recorded neurons N_{rec}	N-N_{hid}
Baselines μ_i	-1.0, $\forall i$
Sparsity p	0.2
Coupling weights $\mathcal{J}_{ij}$ $(i \neq j)$	$\mathcal{N}\left(0, J_0^2/(pN)^{2a}\right)$
Self-coupling weights $\mathcal{J}_{ii}$	0
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

https://doi.org/10.1371/journal.pcbi.1006490.t001

There are two striking effects. First, deviations are nearly negligible ($\mathcal{O}(1/\sqrt{pN})$) for $1/N$ scaling of connections (gray traces in Fig 5). Thus, for large 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 $\mathcal{O}(1)$ impact (purple traces in Fig 5). This is particularly the case when $f \ll 1$—the typical experimental case in which 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 analytical insight into these numerical results, we calculate the standard deviation $\sigma[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}]$, normalized by $\sigma[\mathcal{J}_{r,r'}]$, for contributions from paths up to length-3, focusing on the case of the ER network with mixed synapses (the Dale's law and WS networks are more complicated, as the moments of the synaptic weights depend on the identity of the neurons). For strong $1/\sqrt{N}$ coupling we find

$$\begin{aligned}\frac{\sigma[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}]}{\sigma[\mathcal{J}_{r,r'}]} &\approx \lambda_0 J_0 e^{\mu_0}\sqrt{1-f} \\ &\quad \times \left(1 + \frac{3}{2}(\lambda_0 J_0 e^{\mu_0})^2(1-f)\right),\end{aligned} \tag{4}$$

corresponding to the black dashed curves in Fig 5 left. Eq (4) is a truncation of a series in powers of $\lambda_0 J_0 e^{\mu_0}\sqrt{1-f}$, where $f = N_{\text{rec}}/N$ 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 $N \to \infty$. In contrast, the corresponding expression for $\sigma[\mathcal{J}^{\text{eff}}_{r,r'} - \mathcal{J}_{r,r'}]/\sigma[\mathcal{J}_{r,r'}]$ in the case of weak $1/N$ coupling is a series 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 [67] for derivation and results for $N = 100$.) Deviations of Eq (4) from the numerical solutions in Fig 5 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.

The above analysis demonstrates that the strength of the effective interactions can deviate from that of the true direct interactions by as much as 50%. However, changes in strength do not give us the full picture—we must also investigate how the temporal dynamics of the

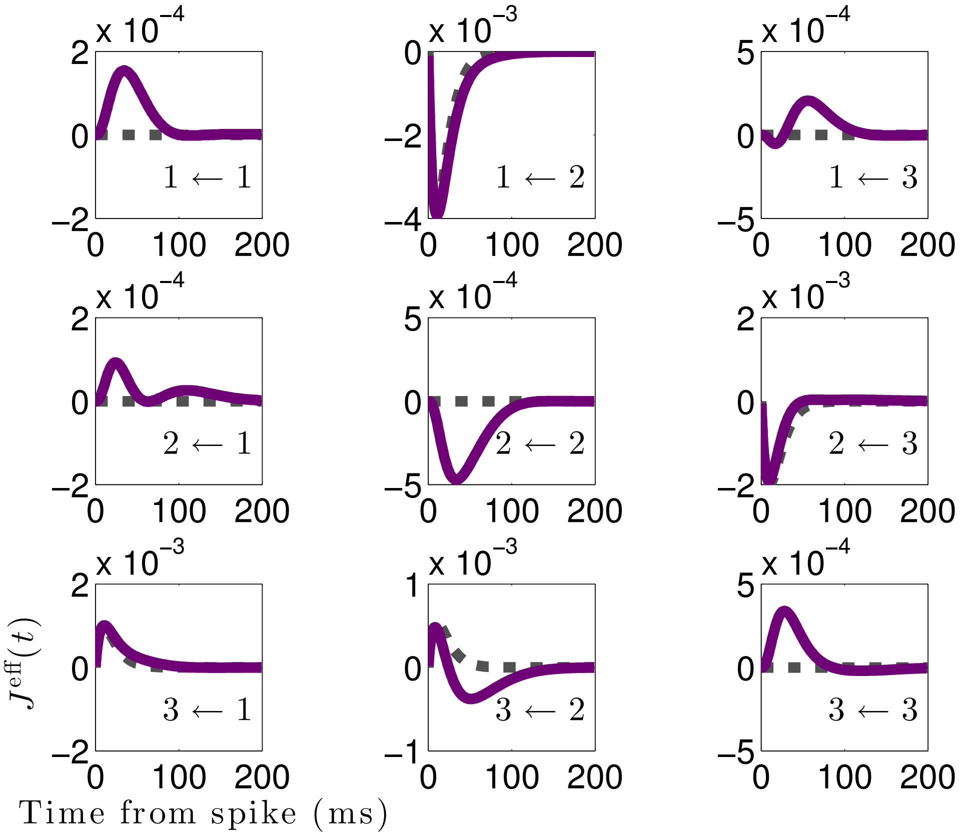

Fig 6. Effective interactions between recorded neurons differ qualitatively from true interactions. Effective interactions $J^{\mathrm{eff}}_{r,r'}(t)$ (solid purple) versus true coupling filters (dashed black) for $N_{\mathrm{rec}} = 3$ recorded neurons in a network of $N = 1000$ total neurons. Inset labels $i \leftarrow j$ indicate the interaction is from neuron j to i, for $i, j \in \{1, 2, 3\}$. 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 $1/\sqrt{pN}$. Although the network is sparse, the effective interactions are not: non-zero effective interactions develop where no direct connection exists. The effective interactions can differ *qualitatively* from the true interactions, as evidenced by the biphasic $3 \leftarrow 2$ effective interaction, whereas the true $3 \leftarrow 2$ is purely excitatory.

https://doi.org/10.1371/journal.pcbi.1006490.g006

effective interactions change. To illustrate how hidden units can skew temporal dynamics, in Fig 6 we plot the effective vs. true interactions between $N_{\mathrm{rec}} = 3$ neurons in an $N = 1000$ neuron network. Because the three network types considered in Fig 5 yield qualitatively similar results, we focus on the Erdős-Réyni network with mixed synapses.

Four of the true interactions between neurons shown in Fig 6 are non-zero ($J^{\mathrm{eff}}_{1,2}(t)$, $J^{\mathrm{eff}}_{3,2}(t)$, $J^{\mathrm{eff}}_{3,1}(t)$, and $J^{\mathrm{eff}}_{2,3}(t)$). Of these, three exhibit only slight differences between the true and effective interactions: $J^{\mathrm{eff}}_{1,2}(t)$ and $J^{\mathrm{eff}}_{3,1}(t)$ have slightly longer decay timescales than their true counterparts, while $J^{\mathrm{eff}}_{2,3}(t)$ has a slightly shorter timescale, indicating the contribution of the hidden network to these interactions was either small or cancelled out. However, the interaction $J^{\mathrm{eff}}_{3,2}(t)$ differs significantly from the true interaction, becoming initially excitatory but switching to inhibitory after a short time, as in our earlier example case of feedforward inhibition. This indicates that neuron 2 must drive a cascade of neurons that ultimately provide inhibitory input to neuron 3.

Contrasting the true and effective interactions shown in Fig 6 highlights many of the ways in which hidden neurons skew the temporal properties of measured interactions. An immediately obvious difference is that although the true synaptic connections in the network are sparse, the effective interactions are not. This is a generic feature of the effective interaction matrix, as in order for an effective interaction from a neuron r' to r to be identically zero there cannot be any paths through the network by which r' can send a signal to r.[1] In a random network the probability that there are no paths connecting two nodes tends to zero as the network size N grows large. Note that this includes paths by which each neuron can send a signal back to itself, hence the neurons developed effective self-interactions, even though the true self-interactions are zero in these particular simulations.

Discussion

We have derived a quantitative relationship between "ground-truth" synaptic interactions and the effective interactions (interpreted here as post-synaptic membrane responses) that unobserved neurons generate between subsets of observed neurons. This relationship, Eq (2) and Fig 2, provides a foundation for studying how different network architectures and neural properties shape the effective interactions between subsets of observed neurons. Our approach can be also be used to study higher order effective interactions between 3 or more neurons, and can be systematically extended to account for corrections to our mean-field approximations and investigate effective noise generated by hidden neurons (using field theoretic techniques from [53], see SI), as well as time-dependent external drives or steady-states.

Here, as first explorations, we focused on the effective interactions corresponding to linear membrane responses. We first demonstrated that our approach applied to small feedforward inhibitory circuits yields effective interactions that capture the role of inhibition in shortening the time window for spiking, and are qualitatively similar to experimentally observed measurements [43]. Moreover, we used this example to demonstrate explicitly that different hidden networks can give rise to the same effective interactions between neurons. We then showed that the influence of hidden neurons can remain significant even in large networks in which the typical synaptic strengths scale with network size. In particular, when the synaptic weights scale as $1/\sqrt{N}$, the relative influence of hidden neurons depends only on the fraction of neurons recorded. Together with theoretical and experimental evidence for this scaling in cortical slices [44–48], this suggests that neural interactions inferred from cortical activity data may differ markedly from the true interactions and connectivity.

Dealing with degeneracy

The issue of degeneracy in complex biological systems and networks has been discussed at length in the literature, in the context of both inherent degeneracies—multiple different network architectures can produce the same qualitative behaviors [34, 37–39], as well as degeneracies in our model descriptions—many models may reproduce experimental observations, demanding sometimes arbitrary criteria for selecting one model over another. All have implications for how successfully one can infer unobserved network properties. One kind of model degeneracy, "sloppiness" [35, 65], describes models in which the behavior of the model is sensitive to changes in only a relatively small number of directions in parameter space. Many models of biological systems have been shown to be sloppy [35]; this could account for experimentally observed networks that are quite different in composition but produce remarkably similar behaviors. Sloppiness suggests that rather than trying to infer all properties of a hidden network, there may be certain parameter combinations that are much more important to the overall network operation, and could potentially be inferred from subsampled observations.

Another perspective on model degeneracy comes from the concepts of "universality" that occur in random matrix theory [68, 69] and Renormalization Group methods of statistical physics [64]. Many bulk properties of matrices (e.g., the distribution of eigenvalues) whose entrees are combinations random variables, such as our $\mathcal{J}^{\text{eff}}_{r,r'}$, are universal in that they depend on only a few key details of the distribution that individual elements are drawn from [70]. Similarly, one of the central results of the Renormalization Group shows that models with drastically disparate features may yield the same coarse-grained model structure when many degrees of freedom are averaged out, as in our case of approximately averaging out hidden neurons. Different distributions (in the case of random matrix theory) or different models (in the case of the Renormalization group) that yield the same bulk properties or coarse-grained models are said to be in the same "universality class." Measuring particular quantities under a range of experimental conditions (e.g., different stimuli) may be able to reveal which universality class an experimental system belongs to and eliminate models belonging to other universality classes as candidate generating models of the data, but these measurements cannot distinguish between models within a universality class.

Purely feedforward networks and recurrent networks are simple examples of broad universality classes in this context. In any randomly sampled feedforward network, only the feedforward interactions are modified or generated; no lateral or feedback connections develop because there is no path through hidden neurons that a recorded neuron can send signals to recorded neurons in the same or previous layers. Thus, the feedforward structure—a topological property of the network—is preserved. However, adding even a single feedback connection can destroy this topological structure if it joins two neurons connected by a feedforward path —i.e., such a link creates a cycle within the network, and it is no longer feedforward. If this network is heavily subsampled ($f \ll 1$) the resulting effective interactions $J^{\text{eff}}_{r,r'}(t)$ can even be fully recurrent. The majority of the effective interactions may be very weak, but nonetheless from a topological perspective the network has been fundamentally altered. Accordingly, any interactions $J^{\text{eff}}_{r,r'}(t)$ that represent a purely feedforward network could not have come from a network with recurrent interactions. In practice, we expect few, if any, cortical networks to be purely feedforward, so most networks will be recurrent if we consider only the network connections and not the connection strengths. Thus, a more interesting question is how the statistics and dynamics of synaptic weights further partition topologically-defined universality classes; for example, whether the distribution of synaptic weights can split the sets of $J^{\text{eff}}_{r,r'}(t)$ that arise from recurrent networks and predominantly feedforward networks with sparse feedback and lateral interactions into different universality classes. A thorough investigation of such phenomena will be the focus of future work.

Inference of hidden network features

Despite the many possible confounds network degeneracy produces, much of the work on inference of hidden network properties has focused on inferring the individual interactions between neurons, with varying degrees of success. Both Dunn and Roudi [40] as well as Tyrcha and Hertz [41] studied inference of hidden activity in kinetic Ising models, sometimes used as simple minimal models of neuronal network activity. They found that the synaptic weights between pairs of observed neurons and observed-hidden pairs could be recovered to within reasonably small mean-squared-error when the number of hidden neurons was less than the number of observed neurons. However, both methods also found it difficult to infer connections between pairs of hidden neurons, resorting to setting such connections to zero in order to stabilize their algorithms. Tyrcha and Hertz also note their method recovers only an equivalence class of connections due to degeneracy in the possible assignment of signs of the synaptic weights and hidden neuron labels. This suggests inferring hidden network structure will be nearly impossible in the realistic limit $N_{\text{rec}} \ll N_{\text{hid}}$.

A series of papers by Bravi and Sollich perform theoretical analyses of hidden dynamics inference in chemical reaction networks, modeled by a system of Langevin equations [57–60]. Although the applications the authors have in mind are signaling pathways such as epidermal growth factor reaction networks, one could imagine re-interpreting or adapting these equations to describe rate models. The authors develop a variety of approaches, including Plefka expansions [57–59] and variational Gaussian approximations [60], to study how observations constrain the inferred hidden dynamics, assuming particlar properties of the network structure. Ref. [60] in particular takes an approach most similar to ours, deriving an effective system of Langevin equations for the subsampled dynamics of the chemical reaction network. The effective system of equations contains a memory kernel that plays a role analogous to the correction to the interactions between neurons in our work (second term in our Eq (2). However, the structure of the memory kernel in [60] has a rather different form, being exponentially dependent on the integral of the hidden-hidden interactions, in contrast to our $\Gamma_{h,h'}(t-t')$, which depends on the inverse of $\delta_{h,h'}\delta(t-t') - \gamma_h J_{h,h'}(t-t')$ (see Methods). Though Bravi and Sollich do not expand their memory kernel in a series as we do, it would admit a similar series and interpretation in terms of paths through the hidden network, as in our Fig 2A. However, due to the exponential dependence on the hidden-hidden interactions, long paths of length ℓ through hidden networks are suppressed by factors $\ell!$, suggesting the hidden network may have less influence in such networks compared to the network dynamics we study here.

Closest to our choice of model, Pillow and Latham [28] and Soudry *et al*[25] both use modifications of nonlinear Hawkes models to fit neural data with unobserved neurons. Pillow and Latham outline a statistical approach for inferring not just interactions with and between hidden neurons, but also the spike trains of hidden neurons, testing the method on a network of two neurons (one hidden). To properly infer the spike train of the hidden neurons, the model must allow for *acausal* synaptic interactions. This is acceptable if the goal is inferring hidden spike trains: for example, if the hidden neuron were to make a strong excitatory synapse onto the observed neuron, then a spike from the observed neuron increases the probability that the hidden neuron fired a spike in the recent past. An acausal synaptic interaction captures this effect, but is of course an unphysical feature in a mechanistic model, precluding physiological interpretation of such an interaction.

Soudry *et al* are concerned with the fact that common input from hidden neurons will skew estimates of network connectivity. To get around this issue, they present a different take on the hidden unit problem: rather than attempt to infer connectivity in a fixed subsample of a network, they propose a shotgun sampling method, in which a sequence of overlapping random subsets of the network are sampled over a long experiment. Under this procedure, a large fraction of the network can be sampled, just not contiguously in time, and reconstruction of the entire network could in principle be accomplished. Soudry *et al* show this strategy works in their simulated networks (even when the generative model is a hard-threshold leaky-integrate-and-fire rather than the nonlinear Hawkes model, which can be interpreted as a soft-threshold leaky-integrate-and-fire model; see SI). However, sampling the entire network may only be feasible *in vitro*; sampling of neurons *in vivo*, such as in wide-field calcium imaging studies, will still necessarily miss neurons not in the field of view or too deep in the tissue; in such cases our work provides the means to properly interpret the inferred effective interactions obtained with such a method.

Although a thorough treatment of statistical inference of hidden network properties is beyond the scope of our present work, we may make some general remarks on future work in these directions. The nonlinear Hawkes model we use here is commonly used to fit neural population activity data, and one could infer the effective baselines $\mu_r^{\rm eff}$, interactions $J_{r,r'}^{\rm eff}(t)$, and noise $\xi_r(t)$ using existing techniques. In particular, Vidne et al. [22] explicitly fit the noise, which is likely important for proper inference, as otherwise effects of the noise could be artificially inherited by the effective interactions. Once such estimates are obtained, one could then in principle infer certain hidden network properties by combining a statistical model for these properties with the relationships between effective and true interactions derived in this work, such as Eq (2). (Detailed physiological measurements of ground-truth synaptic interactions in small volumes of neural tissue can be used to refine estimates). As we have stressed throughout this paper, inferring the exact connections between hidden neurons may be impossible due to a large number of degenerate solutions consistent with observations. However, one may be able to infer bulk properties of the network, such as the parameters governing the distribution of hidden-network connections, or even more exotic properties such as the eigenvalue distribution of the hidden network connection weight matrix. We leave these ideas as interesting directions for future work.

Hidden neurons and dimensionality reduction

Given the challenges that hidden network inference poses, one might wonder if there are network properties that can be reliably measured even with subsampled neural activity. Collective, low-dimensional dynamics have emerged as a possible candidate: recent work has investigated the effect that subsampled measurements have on estimating collective low-dimensional dynamics of trial-averaged network activity (using, e.g., principal components analysis). During a task, the effective dimensionality of a network's dynamics is constrained [71, 72], opening the possibility that the subsampled population may be sufficient to accurately represent these task-constrained low-dimensional dynamics. Indeed, under certain assumptions—in particular that the collective dynamical modes are approximately random superpositions of neural activity and that sampled neurons are statistically representative of the hidden population—Gao et al. [72] calculate a conservative upper bound on the number of sampled neurons necessary to reconstruct the collective dynamics, finding it is often less than the effective dimensionality of the network.

The assumption that the collective dynamics are random superpositions of neural activity is crucial, because it means that each neuron's trial-averaged dynamics are in turn a superposition of the collective modes. Hence, every neuron's activity contains some information about the collective modes, and if only a few of these modes are important, then they can be extracted from any sufficiently large subset of neurons.

While modes of collective activity alone may be sufficient for answering certain questions, such as decoding task parameters or elucidating circuit function, explaining the structure of these modes—and in particular how the dynamical patterns that emerge under different task conditions or sensory environments are related—will ultimately require an understanding of the distribution of possible underlying network properties, which remain difficult to estimate from subsampled populations. We may be able to establish such structure-function relationships using our theory of effective interactions presented in this work: if we can relate the collective dynamics extracted from subsampled neurons to the properties of the effective interactions $J^{\mathrm{eff}}_{rr'}(t)$, then we can link them to the true interactions through our Eq (2). With an understanding of how network properties shape such collective dynamics, we can begin to understand what network manipulations achieve desired patterns of activity, and therefore circuit function.

Implications beyond experimental limitations

The fact that many different hidden networks may yield the same set of effective interactions or low-dimensional dynamics suggests that the effective interactions themselves may yield direct insight into a circuit's functions. For instance, many circuits consist of principal neurons that transmit the results of circuit computation to downstream circuitry, but often do not make direct connections with one another, instead interacting through (predominantly inhibitory) intermediaries called interneurons. From the point of view of a downstream circuit, the principal neurons are "recorded" and the interneurons are "hidden." A potential reason for this general arrangement is that direct synaptic interactions alone are insufficient to produce the membrane responses required to perform the circuit's computations, and the network of interneurons reshapes the membrane responses of projection neurons into effective interactions that can perform the desired computations—it may thus be that the effective interactions should be of primary interest, not necessarily the (possibly degenerate choices of) physiological synaptic interactions. For example, in the feedforward inhibitory circuits of Figs 3 and 4, the roles of the hidden inhibitory neurons may simply be to act as interneurons that reshape the interaction between the excitatory projection neurons 1 and 2, and the choice of which particular circuit motif is implemented in a real network is determined by other physiological constraints, not only computational requirements.

One of the greatest achievements in systems neuroscience would be the ability to perform targeted modifications to a large neural circuit and *selectively* alter its suite of computations. This would have powerful applications for both studying a circuit's native computations, but also repurposing circuits or repairing damaged circuitry (due to, e.g., disease). If the computational roles of circuits are indeed most sensitive to the effective interactions between principal neurons, this suggests we can exploit potential degeneracies in the interneuron architecture and intrinsic properties to find *some* circuit that achieves a desired computation, even if it is not a physiologically natural circuit. Our main result relating effective and true interactions, Eq (2), provides a foundation for future work investigating how to identify sets of circuits that perform a desired set of computations. We have shown in this work that it can be done for small circuits (Figs 3 and 4), and that the effective interactions in large random networks can be significantly skewed away from the true interactions when synaptic weights scale as $1/\sqrt{N}$, as observed in experiments [48]. This holds promise for identifying principled approaches to tuning or controlling neural interactions, such as by using neuromodulators to adjust interneuron properties or inserting artificial or synthetic circuit implants into neural tissue to act as "hidden" neurons. If successful, this could contribute to the long term goal of using such interventions to aid in reshaping the effective synaptic interactions between diseased neurons, and thereby restore healthy circuit behaviors.

Methods

Model definition and details

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

$$\lambda_i(t) = \lambda_0 \phi\left(\mu_i + \mu_i^{\text{ext}}(t) + \sum_j \int_{-\infty}^{\infty} dt' J_{ij}(t-t') \dot{n}_j(t')\right), \tag{5}$$

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, $\mu_i^{\text{ext}}(t)$ is an external input current, and $J_{ij}(t - t')$ 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. We will take $\mu_i^{\text{ext}}(t) = 0$ for simplicity in this work, focusing on the activity of the network due to the tonic drives μ_i (which could be still be interpreted as external tonic inputs, so the activity of the network need not be interpreted as spontaneous activity).

While we need not attach a mechanistic interpretation to these filters, a convenient interpretation is that the nonlinear Hawkes model approximates the stochastic dynamics of a leaky integrate-and-fire network model driven by noisy inputs [55, 56]. In fact, the nonlinear Hawkes model is equivalent to a current-based integrate-and-fire model in which the deterministic spiking rule (a spike fires when a neuron's membrane potential reaches a threshold value V_{th}) is replaced by a stochastic spiking rule (the higher a neuron's membrane potential, the higher the probability a neuron will fire a spike). (It can also be mapped directly to a conductance-based in special cases [73]). For completeness, we present the mapping from a leaky integrate-and-fire model with stochastic spiking to Eq (5) in the Supporting Information (SI).

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 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

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

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

To simplify notation, we write $J_{i,j} * \dot{n}_j = \int_{-\infty}^{\infty} dt'\ J_{i,j}(t - t')\dot{n}_j(t')$. If we seek to describe the firing of the recorded neurons only in terms of their own spiking history, input from hidden neurons effectively acts like noise with some mean amount of input. We thus begin by splitting the hidden input to the recorded neurons up into two terms, the mean plus fluctuations around the mean:

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

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

By construction, the mean of the fluctuations is identically zero, while the cross-correlations can be expressed as

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

where $C_{h_1,h_2}(t_1, t_2)$ is the cross-covariance between hidden neurons h_1 and h_2 (conditioned on the spiking of recorded neurons). If the autocorrelation of the fluctuations ($r = r'$) is small compared to the mean input to the recorded neurons ($\sum_h J_{r,h} * \mathbb{E}[\dot{n}_h(t)|\{\dot{n}_r\}]$), 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 SI) to estimate the size of this noise. Even when this noise is not strictly negligible, it can simply be treated as a separate input to the recorded neurons, as shown in the main text, and hence will not alter the form of the effective couplings between neurons. Averaging out the effective noise, however, would 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 $\mathbb{E}[\dot{n}_h|\{\dot{n}_r\}]$. 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

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

This is not a tractable calculation. Taylor series expanding the nonlinearity $\phi(x)$ 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

$$\mathbb{E}[\dot{n}_h|\{\dot{n}_r\}] = \lambda_0\phi\left(\mu_h + \sum_r J_{h,r} * \dot{n}_r + \sum_{h'} J_{h,h'} * \mathbb{E}[\dot{n}_{h'}|\{\dot{n}_r\}]\right). \tag{6}$$

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

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

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 nonlinear Hawkes 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 nonlinear Hawkes model so that firing rates are modulated by filtering the *deviations* of spikes from the mean firing rate, rather than filtering the spikes themselves. Here, we focus on the zero-activity reference state. We present the formulation for the mean field reference state in the SI.

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

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

The mean inputs $\mathbb{E}[\dot{n}_h|0]$ are the mean field approximations to the firing rates of the hidden neurons in the absence of the recorded neurons. Defining $\nu_h \equiv \mathbb{E}[\dot{n}_h|0]$, 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 assumed 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 $\mathcal{J}_{h,h'} = \int_0^\infty dt\ J_{h,h'}(t)$. This assumption will generally be valid for at least some parameter regime of the network, but there can be cases where it breaks down, such as if the nonlinearity $\phi(x)$ is bounded, in which case a transition to chaotic firing rates $\nu_h(t)$ may exist (c.f. [74]). 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 the solution to the integral equation

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

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.

For time-independent drives μ_r and steady states ν_h (and hence γ_h), we may solve for $\Gamma_{h,h'}(t - t')$ by first converting to the frequency domain and then performing a matrix inverse:

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

where $\hat{V}_{h,h'}(\omega) = \gamma_h J_{h,h'}(\omega)$.

If the zero and first order Taylor series coefficients in our expansion of $\mathbb{E}[\dot{n}_h(t)|\{\dot{n}_r\}]$ are the dominant terms—i.e., if we may neglect higher order terms in this expansion—then 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

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

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 techniques of [53]; we do so to estimate the size of the effective noise correlations in the SI.

In the main text, we decompose our expression for $\hat{J}^{\text{eff}}_{r,r'}(\omega)$ into contributions from all paths that a signal can travel from neuron r' to r. To arrive at this interpretation, we note that we can expand $\hat{\Gamma}_{h,h'}(\omega)$ in a series over paths through the hidden network. To start, we note that if $||\hat{\mathbf{V}}(\omega)|| < 1$ for some matrix norm $||\cdot||$, then the matrix $[\mathbb{I} - \mathbf{V}(\omega)]^{-1}$ admits a convergent series expansion [75]

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

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

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

This expression can be interpreted 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. However, such loops can be easily factored, contributing a factor $\sum_{m=0}^{\infty}\left(\gamma_h\hat{J}_{h,h}(\omega)\right)^m = 1/\left(1-\gamma_h\hat{J}_{h,h}(\omega)\right)$. We 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, as discussed in the main text and depicted in Fig 2. (The contribution of the self-feedback loops can be derived rigorously; see the SI for the full derivation).

Although we have worked here in the frequency domain, our formalism does adapt straightforwardly to handle time-dependent inputs; however, among the consequences of this explicit time-dependence are that the mean field rates $\nu_h(t)$ are not only time-dependent, but solutions of a system of nonlinear integral equations, and hence more challenging to solve. Furthermore, quantities like the linear response of the hidden network, $\Gamma_{h,h'}(t, t')$, will depend on both absolute times t and t', rather than just their difference, $t - t'$, and hence we must also (numerically) solve for $\Gamma_{h,h'}(t, t')$ directly in the time domain. We leave these challenges for future work.

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 inhibitory 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.

Specific networks—Common features. The specific network architectures we study in the main text share several features in common: all are sparse networks with sparsity p (i.e., only a fraction p of connections are non-zero) and non-zero synaptic weight strengths drawn independently from a random distribution with zero population mean and population standard deviation $J_0/(pN)^a$; the overall standard deviation of weights, accounting for the expected $1 - p$ fraction of zero weights is $\sqrt{p}J_0/(pN)^a$. The parameter a determines whether the synaptic strengths are “strong” ($a = 1/2$) or “weak” ($a = 1$). 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. 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 synaptic currents.

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 $1/\sqrt{N}$.

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.

Specific networks—Differences in architecture and synaptic constraints. Beyond the common features outlined above, we perform our analysis of the distribution of effective synaptic interaction strengths for three network architectures commonly studied in network models. These architectures are not intended to be realistic representations of neuronal network structures, but to capture basic features of network architecture and therefore give insight into the basic features of the effective interaction networks.

Erdős-Réyni + mixed synapses—The first network we consider (and the one we perform most of our later analyses on as well) is an Erdős-Réyni random network architecture with “mixed synapses.” That is, each connection between neurons is chosen randomly with probably p. By “mixed synapses” we mean that each neuron’s outgoing synaptic weights are chosen completely independently. i.e., in this network there are no excitatory or inhibitory neurons; each neuron make make both excitatory and inhibitory connections. The corresponding analysis is shown in Fig 5A.

Erdős-Réyni + Dale's law imposed—Real neurons appear to split into separate excitatory and inhibitory classes, a dichotomy know as "Dale's law" (or alternatively, "Dale's principle" to highlight that it is not really a law of nature). Neurons in a network that obeys this law will have coupling filters $J_{i,j}(t)$ that are strictly positive for excitatory neurons and strictly negative for inhibitory neurons. This constraint complicates analytic calculations slightly, as the moments of the synaptic weights now depend on the identity of the neuron, and more care must be taken in calculating expected values or population averages. We instead impose this numerically to generate the results shown in Fig 5B. The trends are the same as in the network with mixed synapses, with the resulting ratios being slightly reduced.

As a technical point, 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 5B: less random samplings for which the hidden network was stable were available to perform the computation. One way this artifact can be prevented is 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 8). Another is to choose a different reference state of network activity around which we perform our expansion of $\mathbb{E}[\dot{n}_h|\{\dot{n}_r\}]$, such as the mean field state discussed in the SI.

Watts-Strogatz network + mixed synapses—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. Therefore, we also consider a network architecture with more structure, a Watts-Strogatz (small world) network. 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 generated the adjacency matrices of the Watts-Strogatz networks using code available in [76]. Here we consider only a Watts-Strogatz network with mixed synapses; a network with spatial structure and Dale's law would become sensitive to both the spatial distribution of excitatory and inhibitory neurons in the network as well as the way in which the neurons are sampled, an investigation we leave for future work. The results for the Watts-Strogatz network with mixed synapses are shown in Fig 5C, and are qualitatively similar to the Erdős-Réyni network with mixed synapses.

Because all three network types we considered yield qualitatively similar results, for the remainder of our analyses, we focus on the Erdős-Réyni + mixed synapses network for simplicity in both simulations and analytical calculations.

Parameter values used to generate our networks are given in Table 1.

Choice of nonlinearity $\phi(x)$

The nonlinear function $\phi(x)$ sets the instantaneous firing rate for the neurons in our model. Our main analytical results (e.g., Eq (2) hold for arbitrary choice of $\phi(x)$. Where specific choices are required in order to perform simulations, we used $\phi(x) = \max(x, 0)$ for the results presented in Figs 3 and 4 and $\phi(x) = \exp(x)$ otherwise. The rectified linear choice is convenient for small networks, as high-order derivatives are zero, which eliminates corresponding high-order "loop corrections" to mean field theory [53]. The exponential function is the "canonical" choice of nonlinearity for the nonlinear Hawkes process [16–18, 20]. The exponential has particularly nice theoretical properties, but is also convenient for fitting the nonlinear Hawkes model to data, as the log-likelihood function of the model simplifies considerably and is convex (though some similar families of nonlinearities also yield convex log-likelihood functions).

An important property that both choices of nonlinearity possess is that they are unbounded. This property is necessary to *guarantee* that a neuron spikes given enough input. A bounded nonlinearity imposes a maximum firing rate, and neurons cannot be forced to spike reliably by providing a large bolus of input. 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.

This can be a problem in simulations of networks obeying Dale's law. For unbounded nonlinearities, 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 selection of recorded of neurons. However, the Dale's law network is stable if the nonlinearity is bounded. We demonstrate this below in Figs 7 and 8, comparing simulations of the effective

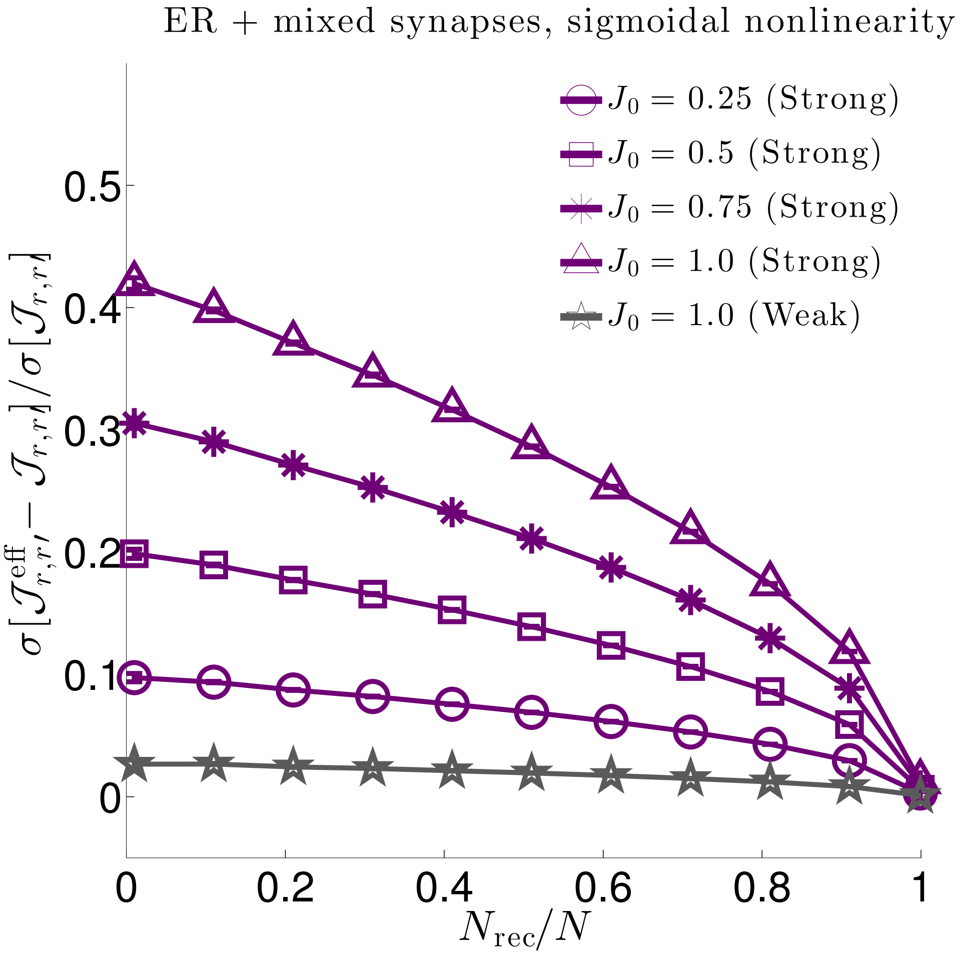

Fig 7. Same as Fig 5A in main text, but for a sigmoidal nonlinearity $\phi(x) = 2/(1 + e^{-x})$.

https://doi.org/10.1371/journal.pcbi.1006490.g007

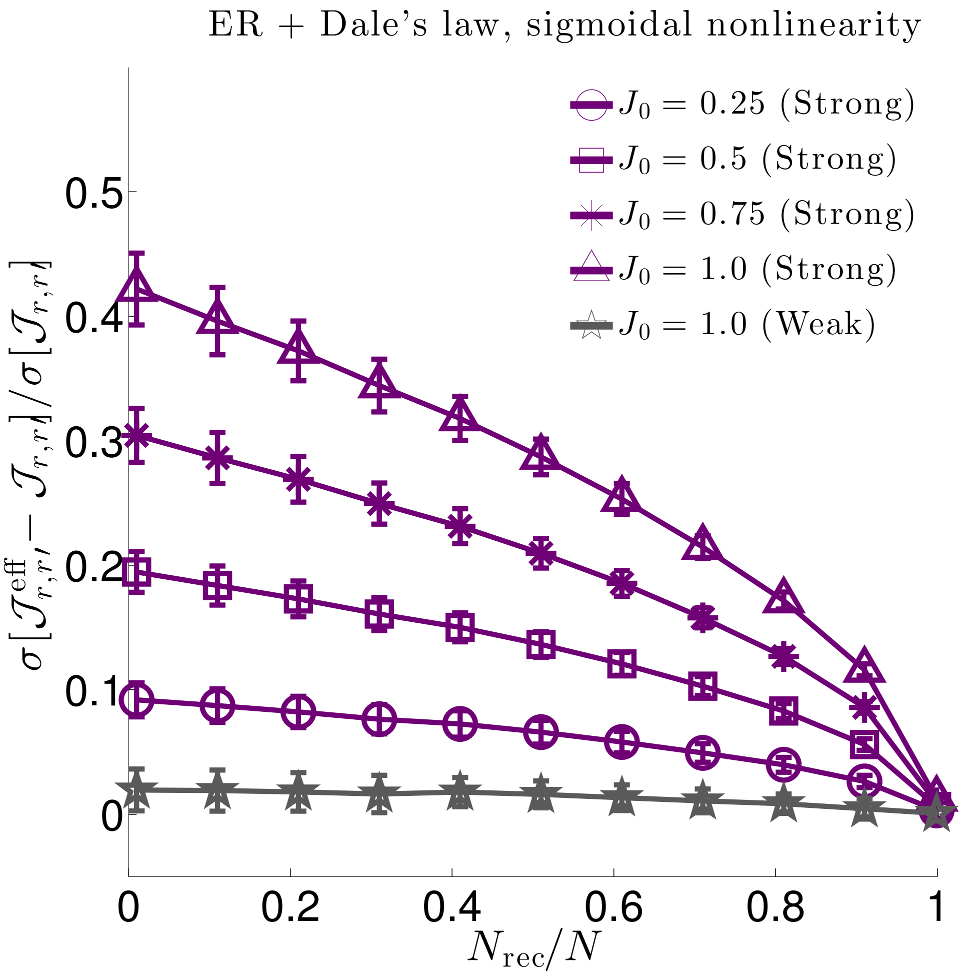

Fig 8. Same as Fig 5B in the main text, but for a sigmoidal nonlinearity $\phi(x) = 2/(1 + e^{-x})$. Because the sigmoid is bounded the mean field solution cannot diverge, yielding better results.

https://doi.org/10.1371/journal.pcbi.1006490.g008

interaction statistics in Erdős-Réyni networks with and without Dale's law for a sigmoidal nonlinearity $\phi(x) = 2/(1 + e^{-x})$.

Another consequence of unbounded nonlinearities is that the mean firing rates are either finite or they diverge. Bounded nonlinearities, on the other hand, may allow for the possibility of a transition to chaotic dynamics in the mean-field firing rate dynamics (cf. the results of the [74]).

Specific choices of network properties used to generate figures

Feedforward-inhibitiory circuit model details. *3 neuron circuit* (Fig 3).

Using our graphical rules (Fig 2), we calculated the effective interaction from neuron 1 to 2 for the circuit shown in Fig 3A, giving Eq 3. In principle, our mean field approximation would not be expected to hold for such a small circuit; in particular, loop corrections [53] to our calculation of the rate ν_3 and associated gain γ_3 might be significant. However, as loop corrections depend on derivatives of the nonlinearity $\phi(x)$, we can minimize these errors by choosing $\phi(x) = \max(x, 0)$, for which $\phi'(x) = \Theta(x)$, the Heaviside step function. Accordingly, we can solve for $\nu_3 = \lambda_0\mu_3/(1 - \lambda_0\mathcal{J}_{33})$ and $\gamma_3 = \lambda_0$ for this particular network.

To generate the plots shown in Fig 3C, we take the inter-neuron couplings to have the form $J_{i,j}(\tau) = \mathcal{J}_{i,j}\alpha_{i,j}^2\tau e^{-\alpha_{i,j}\tau}$ and the self-history couplings to have the form $J_{i,i}(\tau) = \mathcal{J}_{i,i}\beta_{i,i}e^{-\beta_{i,i}\tau}$.

Using Mathematica to perform the inverse Fourier transform, we obtain an explicit expression for the effective interaction,

$$\begin{aligned}
J_{2,1}^{\text{eff}}(\tau) \quad &= \mathcal{J}_{21}\alpha_{21}^2\tau e^{-\alpha_{21}\tau} \\
&\quad + \mathcal{J}_{23}\mathcal{J}_{31}\alpha_{23}^2\alpha_{31}^2 \times \\
&\Bigg[\frac{\beta_{33}\mathcal{J}_{33}}{(\alpha_{23}-\beta_{33}(1-\lambda_0\mathcal{J}_{33}))^2(\alpha_{31}-\beta_{33}(1-\lambda_0\mathcal{J}_{33}))^2}e^{-\beta_{33}(1-\lambda_0\mathcal{J}_{33})\tau} \\
&\quad + \frac{\left(-2\alpha_{31}^2+\beta_{33}\alpha_{31}(4-\lambda_0\mathcal{J}_{33})-2\beta_{33}^2(1-\lambda_0\mathcal{J}_{33})-\beta_{33}\mathcal{J}_{33}\alpha_{23}\right)}{(\alpha_{23}-\alpha_{31})^2(\alpha_{31}-\beta_{33}(1-\lambda_0\mathcal{J}_{33}))^2}e^{-\alpha_{31}\tau} \\
&\quad + \frac{\left(-2\alpha_{23}^2+\beta_{33}\alpha_{23}(4-\lambda_0\mathcal{J}_{33})-2\beta_{33}^2(1-\lambda_0\mathcal{J}_{33})-\beta_{33}\mathcal{J}_{33}\alpha_{31}\right)}{(\alpha_{31}-\alpha_{23})^2(\alpha_{23}-\beta_{33}(1-\lambda_0\mathcal{J}_{33}))^2}e^{-\alpha_{23}\tau} \\
&\quad + \frac{\alpha_{23}-\beta_{33}}{(\alpha_{23}-\alpha_{31})^2(\alpha_{23}-\beta_{33}(1-\lambda_0\mathcal{J}_{33}))^2}\tau e^{-\alpha_{23}\tau} \\
&\quad + \frac{\alpha_{31}-\beta_{33}}{(\alpha_{31}-\alpha_{23})^2(\alpha_{31}-\beta_{33}(1-\lambda_0\mathcal{J}_{33}))^2}\tau e^{-\alpha_{31}\tau}\Bigg].
\end{aligned}$$

In order for the inverse Fourier transform to converge and result in a causal function, we require that $1-\lambda_0\mathcal{J}_{33} > 0$.

Parameter values used to generate the plots in Fig 3C are given in Table 2.

4 neuron circuit (Fig 4).

Like for the 3-neuron circuit, we can use our graphical rules (Fig 2) to calculate the effective interaction for our 4-neuron circuit (Fig 4A) in the frequency domain:

$$\begin{aligned}
\hat{J}_{21}^{\text{eff}}(\omega) - \hat{J}_{21}(\omega) &= \hat{J}_{23}(\omega)\left[\sum_{m=0}^{\infty}(\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega))^m\right]\hat{J}_{31}(\omega) \\
&\quad + \hat{J}_{23}(\omega)\left[\sum_{m=0}^{\infty}(\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega))^m\right]\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{41}(\omega) \\
&= \hat{J}_{23}(\omega)\frac{1}{1-\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega)}\hat{J}_{31}(\omega) \\
&\quad + \hat{J}_{23}(\omega)\frac{1}{1-\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega)}\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{41}(\omega) \\
&= \hat{J}_{23}(\omega)\hat{J}_{31}(\omega) + \hat{J}_{23}(\omega)\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{41}(\omega) \\
&\quad + \frac{\hat{J}_{23}(\omega)\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega)\hat{J}_{31}(\omega)}{1-\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega)} + \frac{\hat{J}_{23}(\omega)\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega)\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{41}(\omega)}{1-\gamma_3\hat{J}_{34}(\omega)\gamma_4\hat{J}_{43}(\omega)};
\end{aligned}$$

in going to the last equality we have separated the terms out into contributions from each of the paths, in order, shown in Fig 4B.

To generate the plots in Fig 4C, we choose $\phi(x) = \max(x, 0)$, which gives $\gamma_i = \lambda_0$, as in Fig 3C, and interaction filters $J_{2,1}(\tau) = \mathcal{J}_{21}\alpha_{21}^2\tau e^{-\alpha\tau}$ for the direct interaction and $J_{i,j}(\tau) = \mathcal{J}_{ij}\alpha^2\tau e^{-\alpha\tau}$ for all other interactions shown—i.e., all other interactions have the same decay time α^{-1} for simplicity.

Table 2. Parameter values for Fig 3C. Setting $\lambda_0 = 1.0$ simply sets the units of frequency and time to be measured relative to λ_0 (e.g., the value $\alpha_{31} = 1.8$ really means $\alpha_{31} = 1.8\lambda_0$ and $\mathcal{J}_{31} = 2.0$ really means $\mathcal{J}_{31} = 2.0/\lambda_0$).

Parameter	value
λ_0	1.0
$\mathcal{J}_{21}$	1.0
$\mathcal{J}_{23}$	−2.0
$\mathcal{J}_{31}$	2.0
$\mathcal{J}_{33}$	−0.9
$\alpha_{21} = \alpha_{23} = \beta_{33}$	1.0
α_{31}	1.8

https://doi.org/10.1371/journal.pcbi.1006490.t002

Inverting the Fourier transform using Mathematica yields

$$\begin{aligned} J^{\text{eff}}_{2,1}(\tau) &= \mathcal{J}_{21}\alpha^2\tau e^{-\alpha\tau} - \frac{\mathcal{J}_{23}\mathcal{J}_{31}\alpha e^{-\alpha\tau}}{2|\mathcal{J}_{34}|^{3/4}|\mathcal{J}_{43}|^{3/4}}\left(\sin(\alpha(|\mathcal{J}_{34}||\mathcal{J}_{43}|)^{1/4}\tau) - \sinh(\alpha(|\mathcal{J}_{34}||\mathcal{J}_{43}|)^{1/4}\tau)\right) \\ &+ \frac{\mathcal{J}_{23}\mathcal{J}_{41}\alpha e^{-\alpha\tau}}{2|\mathcal{J}_{34}|^{1/4}|\mathcal{J}_{43}|^{5/4}}\left(2\alpha(|\mathcal{J}_{34}||\mathcal{J}_{43}|)^{1/4}\tau - \sin(\alpha(|\mathcal{J}_{34}||\mathcal{J}_{43}|)^{1/4}\tau) - \sinh(\alpha(|\mathcal{J}_{34}||\mathcal{J}_{43}|)^{1/4}\tau)\right) \end{aligned}$$

In order for this result to converge, we require $|\mathcal{J}_{34}||\mathcal{J}_{43}| < 1$. Splitting this result up into the contributions to each plot in Fig 4C, using the specific parameter choices $\lambda_0 = 1$ and $\mathcal{J}_{34} = \mathcal{J}_{43} \equiv \mathcal{J}$, gives

$$\begin{aligned} 2 \leftarrow 3 \leftarrow 1 &: \frac{1}{6}\alpha^4\mathcal{J}_{23}\mathcal{J}_{31}\tau^3 e^{-\alpha\tau}, \\ 2 \leftarrow 3 \leftarrow 4 \leftarrow 1 &: -\frac{1}{120}\alpha^6|\mathcal{J}|\mathcal{J}_{23}\mathcal{J}_{41}\tau^5 e^{-\alpha\tau}, \\ 2 \leftarrow 3 \leftrightarrow 4 \leftarrow 3 \leftarrow 1 &: \frac{\alpha\mathcal{J}_{23}\mathcal{J}_{31}(\cosh(\alpha\tau) - \sinh(\alpha\tau))(-2\alpha^3|\mathcal{J}|^{3/2}\tau^3 - 6\sin(\alpha\sqrt{|\mathcal{J}|}\tau) + 6\sinh(\alpha\sqrt{|\mathcal{J}|}\tau)}{12|\mathcal{J}|^{3/2}}, \\ 2 \leftarrow 3 \leftrightarrow 4 \leftarrow 1 &: \frac{\alpha e^{-\alpha\tau}\mathcal{J}_{23}\mathcal{J}_{41}(\alpha\sqrt{|\mathcal{J}|}\tau(120 + \alpha^4\mathcal{J}^2\tau^4) - 60\sin(\alpha\sqrt{\mathcal{J}}\tau) - 60\sinh(\alpha\sqrt{|\mathcal{J}|}\tau))}{120|\mathcal{J}|^{3/2}}. \end{aligned}$$

Parameter values used to generate the plots in Fig 4C are given in Table 3.

Large networks. To generate the results in Fig 6 in the main text, we choose the coupling filters to be $J_{i,j}(t) = \mathcal{J}_{i,j}\alpha^2 te^{-\alpha t}$, for $i \neq j$, which has Fourier transform

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

Table 3. Parameter values for Fig 4C. Setting $\lambda_0 = 1.0$ simply sets the units of frequency and time to be measured relative to λ_0 (e.g., the value $\alpha = 1.294$ really means $\alpha = 1.294\lambda_0$ and $\mathcal{J}_{23} = -3.0$ really means $\mathcal{J}_{23} = -3.0/\lambda_0$).

Parameter	value
λ_0	1.0
$\mathcal{J}_{21} = \mathcal{J}_{31} = \mathcal{J}_{41}$	1.0
$\mathcal{J}_{23}$	−3.0
$\mathcal{J}_{34} = \mathcal{J}_{43} = \mathcal{J}$	−0.9
α_{21}	1.0
α	1.294

https://doi.org/10.1371/journal.pcbi.1006490.t003 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)

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

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

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

and

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

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

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

where $\hat{V}^{\text{HH}}_{h,h'}(\omega) = \gamma_h \mathcal{J}_{h,h'}(\omega)$ and $\text{diag}(\vec{\gamma})$ is an $N_{\text{hid}} \times N_{\text{hid}}$ 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,

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

We then return to the time domain by inverse Fourier transforming the result, achieved by treating $\hat{\mathbf{J}}^{\text{eff}}(\omega)$ as an $N_{\text{rec}} \times N_{\text{rec}} \times N_{\text{freq}}$ array (where N_{freq} is the number of frequencies at which we evaluate the effective coupling) and multiplying along the frequency dimension by an $N_{\text{freq}} \times N_{\text{time}}$ 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 S1 Table.

To generate Fig 5, we focus on the zero-frequency component of $\hat{J}^{\text{eff}}(\omega)$, which is also equal to the time integral of $\mathbf{J}^{\text{eff}}(t)$. As in the main text, we label the elements of this component $\mathcal{J}^{\text{eff}}_{r,r'} = \hat{J}^{\text{eff}}_{r,r'}(\omega = 0)$, which is equal to

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

We do not need to simulate the full network to study the statistics of $\mathcal{J}^{\text{eff}}_{r,r'}$. We only need to generate samples of the matrix $\mathcal{J}$ and evaluate $\hat{\boldsymbol{\Gamma}}(0)$. This is where the choice of an Erdős-Réyni network that is not restricted to obey Dale's law 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

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

and sample variance

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

will tend to the expected values $\mathbb{E}[\mathcal{J}^{\text{eff}}_{r,r'}]$ and $\text{var}[\mathcal{J}^{\text{eff}}_{r,r'}]$ 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 5 (after normalization by the population variance of the true off-diagonal weights), to the expected variance calculated analytically below.

The error bars in Fig 5 are generated by first drawing a single sample of true weights $\mathcal{J}$, and then taking 100 random subsets of N_{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 $\tilde{\mathcal{J}}_{\text{var}}$ 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 $\mathcal{J}^{\text{eff}}_{r,r'}$. We note that for an Erdős-Réyni network with mixed synapses, 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 (i.e., the network is self-averaging). However, this will not be true for networks with more structure, such as the Watts-Strogatz or Dale's law networks we also considered, for which the second stage of averaging over the global network architecture is necessary to average over network configurations.

Series approximation for the mean field firing rates for the case of exponential nonlinearity $\phi(x) = e^x$

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 we do not need to calculate a separate 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,

$$\begin{aligned}a_h^{(1)} &= \sum_{h'_1} \mathcal{J}_{h,h'_1}, \\ a_h^{(2)} &= \sum_{h'_1,h'_2} \left\{ \mathcal{J}_{h,h'_1} \mathcal{J}_{h'_1,h'_2} + \frac{1}{2} \mathcal{J}_{h,h'_1} \mathcal{J}_{h,h'_2} \right\}, \\ a_h^{(3)} &= \sum_{h'_1,h'_2,h'_3} \left\{ \mathcal{J}_{h,h'_1} \mathcal{J}_{h'_1,h'_2} \mathcal{J}_{h'_2,h'_3} + \frac{1}{2} J_{h,h'_1} \mathcal{J}_{h'_1,h'_2} \mathcal{J}_{h'_1,h'_3} + \mathcal{J}_{h,h'_1} \mathcal{J}_{h,h'_2} \mathcal{J}_{h'_2,h'_3} \right. \\ &\qquad \left. + \frac{1}{3!} \mathcal{J}_{h,h'_1} \mathcal{J}_{h,h'_2} \mathcal{J}_{h,h'_3} \right\}.\end{aligned}$$

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

$$\nu_h = \lambda_0 \epsilon_h \left(1 + \sum_{h'} \mathcal{J}_{h,h'} \lambda_0 \epsilon_{h'} + \sum_{h'_1,h'_2} \left\{ \mathcal{J}_{h,h'_1} \mathcal{J}_{h'_1,h'_2} + \frac{1}{2} \mathcal{J}_{h,h'_1} \mathcal{J}_{h,h'_2} \right\} \lambda_0 \epsilon_{h'_1} \lambda_0 \epsilon_{h'_2} + \ldots \right).$$

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 $\mathcal{J}_{r,r'} \equiv \int_0^{\infty} dt \; J_{r,r'}(t)$, 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 $\mathrm{var}(\mathcal{J}) = p\frac{J_0^2}{(pN)^{2a}}$ 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 $\mathcal{J}^{\text{eff}}_{r,r'}$:

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

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

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

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 $1 - f \equiv N_{\text{hid}}/N$.

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

$$\begin{aligned} \frac{\mathrm{var}[\mathcal{J}^{\mathrm{eff}}_{r,r'}]}{\mathrm{var}[\mathcal{J}]} &= 1 + (\lambda_0 \epsilon)^2 N_{\mathrm{hid}} \mathrm{var}[\mathcal{J}] \\ &= 1 + (\lambda_0 J_0 \epsilon)^2 (1-f) \frac{1}{(pN)^{2a-1}}. \end{aligned} \tag{7}$$

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, though we could also use the path-length series expression for $\hat{\Gamma}_{h,h'}(\omega)$, keeping terms up to fourth order in ϵ. We begin with the 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

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

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

$$
\begin{aligned}
v_h^2 &\approx (\lambda_0\epsilon)^2\left(1+\sum_{h_1}\mathcal{J}_{h,h_1}\lambda_0\epsilon+\sum_{h_1,h_2}a^{(2)}_{h,h_1,h_2}(\lambda_0\epsilon)^2\right)\\
&\qquad\times\left(1+\sum_{h'_1}\mathcal{J}_{h,h'_1}\lambda_0\epsilon+\sum_{h'_1,h'_2}a^{(2)}_{h,h'_1,h'_2}(\lambda_0\epsilon)^2\right)\\
&\approx(\lambda_0\epsilon)^2\left(1+2\left(\sum_{h_1}\mathcal{J}_{h,h_1}\lambda_0\epsilon+\sum_{h_1,h_2}a^{(2)}_{h,h_1,h_2}(\lambda_0\epsilon)^2\right)+\left(\sum_{h_1}\mathcal{J}_{h,h_1}\lambda_0\epsilon\right)^2\right)\\
&=(\lambda_0\epsilon)^2\left(1+2\sum_{h_1}\mathcal{J}_{h,h_1}\lambda_0\epsilon+\sum_{h_1,h_2}\left\{2a^{(2)}_{h,h_1,h_2}+\mathcal{J}_{h,h_1}\mathcal{J}_{h,h_2}\right\}(\lambda_0\epsilon)^2\right),
\end{aligned}
$$

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 $\overline{v_h^2}$. The third-order term will vanish upon averaging, and

$$
\overline{2a^{(2)}_{h,h_1,h_2}+\mathcal{J}_{h,h_1}\mathcal{J}_{h,h_2}}=\overline{2\mathcal{J}_{h,h_1}\mathcal{J}_{h_1,h_2}+2\mathcal{J}_{h,h_1}\mathcal{J}_{h,h_2}}=2\mathrm{var}[\mathcal{J}]\delta_{h_1,h_2}(1-\delta_{h,h_1}),
$$

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

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

The first fourth order term in $\hat{\Gamma}^2_{h,h'}$, $2\sum_{h''}v_h^3 v_{h''}\mathcal{J}_{h,h''}\mathcal{J}_{h'',h}\delta_{h,h'}$, 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 $\mathrm{var}[\mathcal{J}](1-\delta_{h,h'})$, where the factor of $(1-\delta_{h,h'})$ again accounts for the fact that this term does not contribute when $h=h'$ due to no self-couplings. We thus arrive at

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

Putting everything together,

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

For weak coupling, this tends to 1 in the $N\gg1$ limit, as $N_{\mathrm{hid}}\mathrm{var}[\mathcal{J}]=(1-f)J_0^2/N\to0$, for fixed fraction of observed neurons $f=N_{\mathrm{rec}}/N$. For strong coupling, $N_{\mathrm{hid}}\mathrm{var}[\mathcal{J}]=(1-f)J_0^2$, which is constant as $N \to \infty$, and hence

$$\frac{\mathrm{var}[\mathcal{J}^{\mathrm{eff}}_{r,r'}]}{\mathrm{var}[\mathcal{J}]} = 1 + (\lambda_0 J_0 \epsilon)^2 (1-f) + 3(\lambda_0 J_0 \epsilon)^4 (1-f)^2 + o\big((\lambda_0 J_0 \epsilon)^4 (1-f)^2\big), \qquad (8)$$

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 (8) as

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

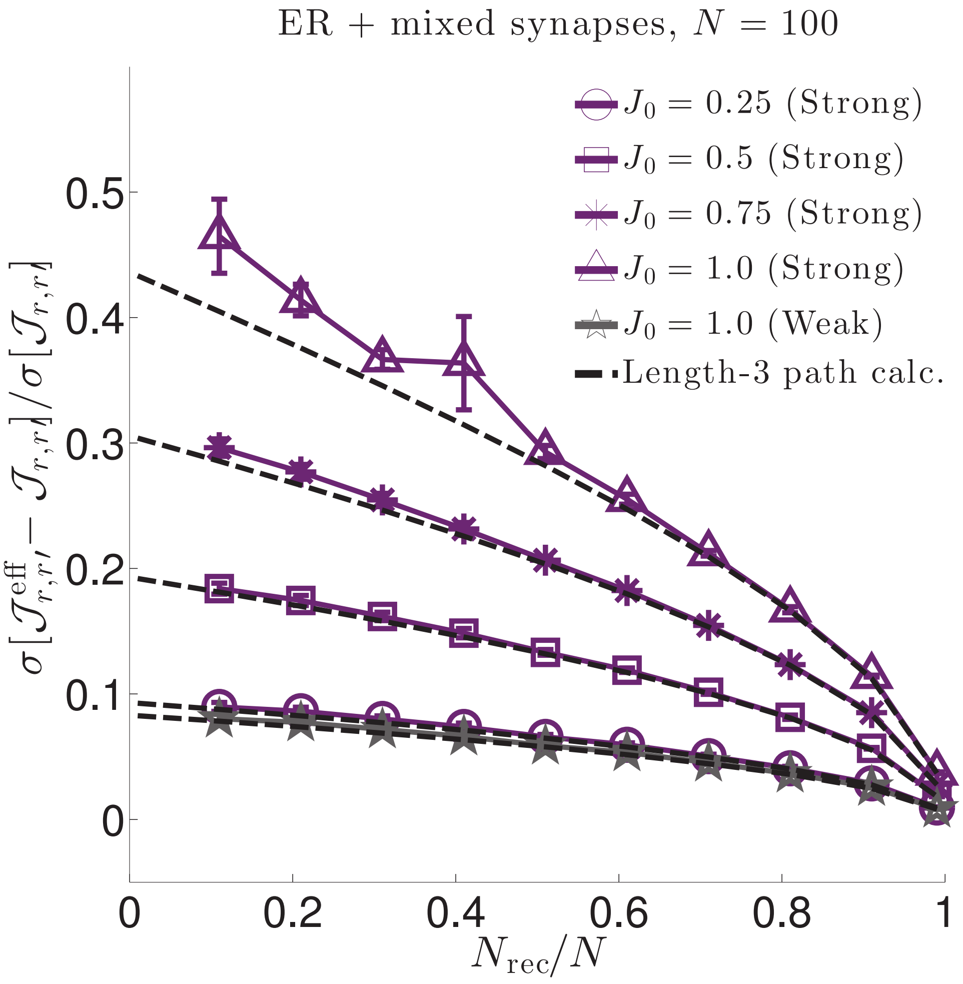

Fig 9. Same as Fig 5, but for $N = 100$ neurons and $N_{\mathrm{rec}} = \{1, 11, 21, 31, 41, 51, 61, 71, 81, 91, 99\}$ recorded neurons. Because we plot the relative deviations of the coupling strength against the fraction of observed neurons, the curves for the strongly coupled case are the same as for $N = 1000$, as expected, while the weakly coupled case yields stronger deviations.

https://doi.org/10.1371/journal.pcbi.1006490.g009

or, in terms of the ratio of standard deviations,

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

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 5A). 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 9 we remake Fig 5 for $N = 100$ neurons. We see that the strongly coupled results have not been significantly altered, whereas the weakly coupled results yield stronger deviations (as the deviations are $\mathcal{O}(1/\sqrt{N})$).

Supporting information

S1 Fig. 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 the mean field theory approximation up to $J_0 = 1.0$.
(EPS)

S2 Fig. 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 $N_{\text{rec}} = 100$.
(EPS)

S3 Fig. Same as S2 Fig 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.
(EPS)

S1 Table. Network activity simulation parameter values.
(PDF)

S1 Text. Supporting information.
(PDF)

[1] Perfect cancellation, though in principle possible, is almost surely impossible in the untuned random networks that we study here.

Acknowledgments

We thank Gabe Ocker for providing the nonlinear Hawkes 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. 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.

Author Contributions

Conceptualization: Braden A. W. Brinkman, Fred Rieke, Eric Shea-Brown, Michael A. Buice.

Formal analysis: Braden A. W. Brinkman, Michael A. Buice.

Funding acquisition: Fred Rieke, Eric Shea-Brown, Michael A. Buice.

Investigation: Braden A. W. Brinkman.

Methodology: Braden A. W. Brinkman, Michael A. Buice.

Software: Braden A. W. Brinkman.

Validation: Braden A. W. Brinkman.

Visualization: Braden A. W. Brinkman.

Writing – original draft: Braden A. W. Brinkman.

Writing – review & editing: Braden A. W. Brinkman, Fred Rieke, Eric Shea-Brown, Michael A. Buice.

References

1. Bassett DS, Bullmore E, Verchinski BA, Mattay VS, Weinberger DR, Meyer-Lindenberg A. Hierarchical Organization of Human Cortical Networks in Health and Schizophrenia. Journal of Neuroscience. 2008; 28(37):9239–9248. https://doi.org/10.1523/JNEUROSCI.1929-08.2008 PMID: 18784304
2. Kramer MA, Kolaczyk ED, Kirsch HE. Emergent network topology at seizure onset in humans. Epilepsy Research. 2008; 79(2–3):173–186. http://dx.doi.org/10.1016/j.eplepsyres.2008.02.002 PMID: 18359200
3. Supekar K, Menon V, Rubin D, Musen M, Greicius MD. Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease. PLOS Computational Biology. 2008; 4(6):1–11.
4. Lo CY, Wang PN, Chou KH, Wang J, He Y, Lin CP. Diffusion Tensor Tractography Reveals Abnormal Topological Organization in Structural Cortical Networks in Alzheimer's Disease. Journal of Neuroscience. 2010; 30(50):16876–16885. https://doi.org/10.1523/JNEUROSCI.4136-10.2010 PMID: 21159959
5. Chavez M, Valencia M, Navarro V, Latora V, Martinerie J. Functional Modularity of Background Activities in Normal and Epileptic Brain Networks. Phys Rev Lett. 2010; 104:118701. https://doi.org/10.1103/PhysRevLett.104.118701 PMID: 20366507
6. Douw L, de Groot M, van Dellen E, Heimans JJ, Ronner HE, Stam CJ, et al. 'Functional Connectivity' Is a Sensitive Predictor of Epilepsy Diagnosis after the First Seizure. PLOS ONE. 2010; 5(5):1–7.
7. van Diessen E, Hanemaaijer JI, Otte WM, Zelmann R, Jacobs J, Jansen FE, et al. Are high frequency oscillations associated with altered network topology in partial epilepsy? NeuroImage. 2013; 82:564–573. http://dx.doi.org/10.1016/j.neuroimage.2013.06.031 PMID: 23792218
8. Reijmer YD, Leemans A, Caeyenberghs K, Heringa SM, Koek HL, Biessels GJ, et al. Disruption of cerebral networks and cognitive impairment in Alzheimer disease. Neurology. 2013; 80(15):1370–1377. https://doi.org/10.1212/WNL.0b013e31828c2ee5 PMID: 23486876
9. Seo EH, Lee DY, Lee JM, Park JS, Sohn BK, Lee DS, et al. Whole-brain Functional Networks in Cognitively Normal, Mild Cognitive Impairment, and Alzheimer's Disease. PLOS ONE. 2013; 8(1):1–11. https://doi.org/10.1371/journal.pone.0053922
10. Stam CJ. Modern network science of neurological disorders. Nat Rev Neurosci. 2014; 15(10):683–695. https://doi.org/10.1038/nrn3801 PMID: 25186238

11. Warren DE, Power JD, Bruss J, Denburg NL, Waldron EJ, Sun H, et al. Network measures predict neuropsychological outcome after brain injury. Proceedings of the National Academy of Sciences. 2014; 111(39):14247–14252. https://doi.org/10.1073/pnas.1322173111

12. Olde Dubbelink KTE, Hillebrand A, Stoffers D, Deijen JB, Twisk JWR, Stam CJ, et al. Disrupted brain network topology in Parkinson's disease: a longitudinal magnetoencephalography study. Brain. 2014; 137(1):197–207. https://doi.org/10.1093/brain/awt316 PMID: 24271324

13. Bernhardt BC, Bonilha L, Gross DW. Network analysis for a network disorder: The emerging role of graph theory in the study of epilepsy. Epilepsy & Behavior. 2015; 50:162–170. http://dx.doi.org/10.1016/j.yebeh.2015.06.005.

14. Medaglia JD, Bassett DS. Network Analyses and Nervous System Disorders. ArXiv e-prints. 2017;.

15. Honey CJ, Thivierge JP, Sporns O. Can structure predict function in the human brain? NeuroImage. 2010; 52(3):766–776. http://dx.doi.org/10.1016/j.neuroimage.2010.01.071 PMID: 20116438

16. Simoncelli E, Paninski L, Pillow J, Schwartz O. Characterization of Neural Responses with Stochastic Stimuli. In: Gazzaniga M, editor. The Cognitive Neurosciences. 3rd ed. MIT Press; 2004. p. 327–338.

17. Paninski L. Maximum likelihood estimation of cascade point-process neural encoding models. Network: Computation in Neural Systems. 2004; 15(4):243–262. https://doi.org/10.1088/0954-898X_15_4_002

18. Pillow J, Paninski L, Uzzell VJ, Simoncelli E, Chichilnisky EJ. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2005; 25(47):11003–13. https://doi.org/10.1523/JNEUROSCI.3305-05.2005

19. Kulkarni JE, Paninski L. Common-input models for multiple neural spike-train data. Network: Computation in Neural Systems. 2007; 18(4):375–407. https://doi.org/10.1080/09548980701625173

20. Pillow J, Shlens J, Paninski L, Sher A, Litke A, Chichilnisky E, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature. 2008; 454(7207):995–9. https://doi.org/10.1038/nature07140 PMID: 18650810

21. Field G, Gauthier JL, Sher A, Greschner M, Machado Ta, Jepson LH, et al. Functional connectivity in the retina at the resolution of photoreceptors. Nature. 2010; 467(7316):673–7. https://doi.org/10.1038/nature09424 PMID: 20930838

22. Vidne M, Ahmadian Y, Shlens J, Pillow J, Kulkarni J, Litke A, et al. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. J Comput Neurosci. 2012; 33(1):97–121. https://doi.org/10.1007/s10827-011-0376-2 PMID: 22203465

23. Paninski L. Maximum likelihood estimation of cascade point- process neural encoding models. Network: Computation in Neural Systems. 2015; 6536(October).

24. Huang H. Effects of hidden nodes on network structure inference. Journal of Physics A: Mathematical and Theoretical. 2015; 48(35):355002. https://doi.org/10.1088/1751-8113/48/35/355002

25. Soudry D, Keshri S, Stinson P, Oh Mh, Iyengar G, Paninski L. Efficient "Shotgun" Inference of Neural Connectivity from Highly Sub-sampled Activity Data. PLOS Computational Biology. 2015; 11(10):1–30.

26. Dahlhaus R. Graphical interaction models for multivariate time series1. Metrika. 2000; 51(2):157–172. https://doi.org/10.1007/s001840000055

27. Eichler M, Dahlhaus R, Sandkühler J. Partial correlation analysis for the identification of synaptic connections. Biological Cybernetics. 2003; 89(4):289–302. https://doi.org/10.1007/s00422-003-0400-3 PMID: 14605893

28. Pillow JW, Latham PE. Neural characterization in partially observed populations of spiking neurons. In: c Platt J, Koller D, Singer Y, Roweis S, editors. Advances in Neural Information Processing Systems 20. Cambridge, MA: MIT Press; 2007. p. 1161–1168. Available from: http://books.nips.cc/papers/files/nips20/NIPS2007_0995.pdf.

29. Stevenson IH, Rebesco JM, Miller LE, Körding KP. Inferring functional connections between neurons. Current Opinion in Neurobiology. 2008; 18(6):582–588. http://dx.doi.org/10.1016/j.conb.2008.11.005 PMID: 19081241

30. Stevenson IH, London BM, Oby ER, Sachs NA, Reimer J, Englitz B, et al. Functional Connectivity and Tuning Curves in Populations of Simultaneously Recorded Neurons. PLOS Computational Biology. 2012; 8(11):1–14. https://doi.org/10.1371/journal.pcbi.1002775

31. Liégeois R, Mishra B, Zorzi M, Sepulchre R. Sparse plus low-rank autoregressive identification in neuroimaging time series. In: 2015 54th IEEE Conference on Decision and Control (CDC); 2015. p. 3965–3970.

32. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2016; 78(5):947–1012. https://doi.org/10.1111/rssb.12167

33. Foti NJ, Nadkarni R, Lee AK, Fox EB. Sparse plus low-rank graphical models of time series for functional connectivity in MEG. In: 2nd KDD Workshop on Mining and Learning from Time Series; 2016. Available from: http://www-bcf.usc.edu/~liu32/milets16/paper/MiLeTS_2016_paper_22.pdf.

34. Prinz Aa, Bucher D, Marder E. Similar network activity from disparate circuit parameters. Nature neuroscience. 2004; 7(12):1345–52. https://doi.org/10.1038/nn1352 PMID: 15558066

35. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally Sloppy Parameter Sensitivities in Systems Biology Models. PLOS Computational Biology. 2007; 3(10):1–8. https://doi.org/10.1371/journal.pcbi.0030189

36. Apgar JF, Witmer DK, White FM, Tidor B. Sloppy models, parameter uncertainty, and the role of experimental design. Mol BioSyst. 2010; 6:1890–1900. https://doi.org/10.1039/b918098b PMID: 20556289

37. Gutierrez GJ, O'Leary T, Marder E. Multiple Mechanisms Switch an Electrically Coupled, Synaptically Inhibited Neuron between Competing Rhythmic Oscillators. Neuron. 2013; 77(5):845–858. https://doi.org/10.1016/j.neuron.2013.01.016 PMID: 23473315

38. Fisher D, Olasagasti I, Tank DW, Aksay ERF, Goldman MS. A Modeling Framework for Deriving the Structural and Functional Architecture of a Short-Term Memory Microcircuit. Neuron. 2013; 79(5):987–1000. https://doi.org/10.1016/j.neuron.2013.06.041 PMID: 24012010

39. Marder E, Gutierrez GJ, Nusbaum MP. Complicating connectomes: Electrical coupling creates parallel pathways and degenerate circuit mechanisms. Developmental Neurobiology. 2017; 77(5):597–609. https://doi.org/10.1002/dneu.22410 PMID: 27314561

40. Dunn B, Roudi Y. Learning and inference in a nonequilibrium Ising model with hidden nodes. Phys Rev E. 2013; 87:022127. https://doi.org/10.1103/PhysRevE.87.022127

41. Tyrcha J, Hertz J. Network inference with hidden units. Mathematical Biosciences and Engineering. 2014; 11(1):149–165. https://doi.org/10.3934/mbe.2014.11.149 PMID: 24245678

42. Dunn B, Battistin C. The appropriateness of ignorance in the inverse kinetic Ising model. ArXiv e-prints. 2016;.

43. Pouille F, Scanziani M. Enforcement of Temporal Fidelity in Pyramidal Cells by Somatic Feed-Forward Inhibition. Science. 2001; 293(5532):1159–1163. https://doi.org/10.1126/science.1060342 PMID: 11498596

44. van Vreeswijk C, Sompolinsky H. Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity. Science. 1996; 274(5293):1724–1726. https://doi.org/10.1126/science.274.5293.1724 PMID: 8939866

45. Litwin-Kumar A, Doiron B. Slow dynamics and high variability in balanced cortical networks with clustered connections. Nat Neurosci. 2012; 15(11):1498–1505. https://doi.org/10.1038/nn.3220 PMID: 23001062

46. Rosenbaum R, Doiron B. Balanced Networks of Spiking Neurons with Spatially Dependent Recurrent Connections. Phys Rev X. 2014; 4:021039.

47. Deneve S, Machens CK. Efficient codes and balanced networks. Nat Neurosci. 2016; 19(3):375–382. https://doi.org/10.1038/nn.4243 PMID: 26906504

48. Barral J, D Reyes A. Synaptic scaling rule preserves excitatory-inhibitory balance and salient neuronal network dynamics. Nat Neurosci. 2016; 19(12):1690–1696. https://doi.org/10.1038/nn.4415 PMID: 27749827

49. Nykamp DQ. Revealing Pairwise Coupling in Linear-Nonlinear Networks. SIAM Journal on Applied Mathematics. 2005; 65(6):2005–2032. https://doi.org/10.1137/S0036139903437072

50. Nykamp DQ. A mathematical framework for inferring connectivity in probabilistic neuronal networks. Mathematical Biosciences. 2007; 205(2):204–251. http://dx.doi.org/10.1016/j.mbs.2006.08.020 PMID: 17070863

51. Nykamp DQ. Exploiting History-Dependent Effects to Infer Network Connectivity. SIAM Journal on Applied Mathematics. 2007; 68(2):354–391. https://doi.org/10.1137/070683350

52. Nykamp DQ. Pinpointing connectivity despite hidden nodes within stimulus-driven networks. Phys Rev E. 2008; 78:021902. https://doi.org/10.1103/PhysRevE.78.021902

53. Ocker GK, Josić K, Shea-Brown E, Buice MA. Linking structure and activity in nonlinear spiking networks. PLOS Computational Biology. 2017; 13(6):1–46. https://doi.org/10.1371/journal.pcbi.1005583

54. Chornoboy ES, Schramm LP, Karr AF. Maximum likelihood identification of neural point process systems. Biological Cybernetics. 1988; 59(4):265–275. https://doi.org/10.1007/BF00332915 PMID: 3196770

55. Gerstner W, Kistler WM, Naud R, Paninski L. Neuronal Dynamics: From single neurons to networks and models of cognition. Cambridge, U.K.: Cambridge University Press; 2014.

56. Ostojic S, Brunel N. From Spiking Neuron Models to Linear-Nonlinear Models. PLOS Computational Biology. 2011; 7(1):1–16. https://doi.org/10.1371/journal.pcbi.1001056

57. Bravi B, Opper M, Sollich P. Inferring hidden states in Langevin dynamics on large networks: Average case performance. Phys Rev E. 2017; 95:012122. https://doi.org/10.1103/PhysRevE.95.012122 PMID: 28208380

58. Bravi B, Sollich P. Inference for dynamics of continuous variables: the extended Plefka expansion with hidden nodes. Journal of Statistical Mechanics: Theory and Experiment. 2017; 2017(6):063404. https://doi.org/10.1088/1742-5468/aa657d

59. Bravi B, Sollich P. Critical scaling in hidden state inference for linear Langevin dynamics. Journal of Statistical Mechanics: Theory and Experiment. 2017; 2017(6):063504. https://doi.org/10.1088/1742-5468/aa6bc4

60. Bravi B, Sollich P. Statistical physics approaches to subnetwork dynamics in biochemical systems. Physical Biology. 2017; 14(4):045010. https://doi.org/10.1088/1478-3975/aa7363 PMID: 28510539

61. Pernice V, Staude B, Cardanobile S, Rotter S. How Structure Determines Correlations in Neuronal Networks. PLOS Computational Biology. 2011; 7(5):1–14. https://doi.org/10.1371/journal.pcbi.1002059

62. Hu Y, Trousdale J, Josić K, Shea-Brown E. Motif statistics and spike correlations in neuronal networks. Journal of Statistical Mechanics: Theory and Experiment. 2013; 2013(03):P03012. https://doi.org/10.1088/1742-5468/2013/03/P03012

63. Hu Y, Trousdale J, Josić K, Shea-Brown E. Local paths to global coherence: Cutting networks down to size. Phys Rev E. 2014; 89:032802. https://doi.org/10.1103/PhysRevE.89.032802

64. Goldenfeld N. Lectures on Phase Transitions and the Renormalization Group. Westview Press; 1992.

65. Machta BB, Chachra R, Transtrum MK, Sethna JP. Parameter Space Compression Underlies Emergent Theories and Predictive Models. Science. 2013; 342(6158):604–607. https://doi.org/10.1126/science.1238723 PMID: 24179222

66. Cayco-Gajic NA, Zylberberg J, Shea-Brown E. Triplet correlations among similarly tuned cells impact population coding. Frontiers in Computational Neuroscience. 2015; 9:57. https://doi.org/10.3389/fncom.2015.00057 PMID: 26042024

67. See Supplementary Information.;.

68. Erdős L. Universality of Wigner random matrices: a survey of recent results. Russian Mathematical Surveys. 2011; 66(3):507. https://doi.org/10.1070/RM2011v066n03ABEH004749

69. Ahmadian Y, Fumarola F, Miller KD. Properties of networks with partially structured and partially random connectivity. Phys Rev E. 2015; 91:012820. https://doi.org/10.1103/PhysRevE.91.012820

70. TAO T, VU V. RANDOM MATRICES: THE CIRCULAR LAW. Communications in Contemporary Mathematics. 2008; 10(02):261–307. https://doi.org/10.1142/S0219199708002788

71. Mazzucato L, Fontanini A, La Camera G. Stimuli Reduce the Dimensionality of Cortical Activity. Frontiers in Systems Neuroscience. 2016; 10:11. https://doi.org/10.3389/fnsys.2016.00011 PMID: 26924968

72. Gao P, Trautmann E, Yu BM, Santhanam G, Ryu S, Shenoy K, et al. A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv. 2017;

73. Latimer K, Chichilnisky E, Rieke F, Pillow J. In: Inferring synaptic conductances from spike trains under a biophysically inspired point process model. vol. 2. january ed. Neural information processing systems foundation; 2014. p. 954–962.

74. Sompolinsky H, Crisanti A, Sommers HJ. Chaos in Random Neural Networks. Phys Rev Lett. 1988; 61:259–262. https://doi.org/10.1103/PhysRevLett.61.259 PMID: 10039285

75. Horn RA, Johnson CR, editors. Matrix Analysis. New York, NY, USA: Cambridge University Press; 1986.

76. Song HF, Wang XJ. Simple, distance-dependent formulation of the Watts-Strogatz model for directed and undirected small-world networks. Phys Rev E. 2014; 90:062801. https://doi.org/10.1103/PhysRevE.90.062801