Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates

Hong-Li Zeng,[1,*] Mikko Alava,[1] Erik Aurell,[2,3,4] John Hertz,[5,6] and Yasser Roudi[5,7]

[1]*Department of Applied Physics, Aalto University, FIN-00076 Aalto, Finland*
[2]*Department of Computational Biology, KTH-Royal Institute of Technology, SE-100 44 Stockholm, Sweden*
[3]*ACCESS Linnaeus Centre, KTH-Royal Institute of Technology, SE-100 44 Stockholm, Sweden*
[4]*Department of Information and Computer Science, Aalto University, FIN-00076 Aalto, Finland*
[5]*Nordita, KTH-Royal Institute of Technology and Stockholm University, 10691 Stockholm, Sweden*
[6]*The Niels Bohr Institute, 2100 Copenhagen, Denmark*
[7]*Kavli Institute for Systems Neuroscience, NTNU, 7030 Trondheim, Norway*

(Received 11 September 2012; revised manuscript received 26 November 2012; published 20 May 2013)

We describe how the couplings in an asynchronous kinetic Ising model can be inferred. We consider two cases: one in which we know both the spin history and the update times and one in which we know only the spin history. For the first case, we show that one can average over all possible choices of update times to obtain a learning rule that depends only on spin correlations and can also be derived from the equations of motion for the correlations. For the second case, the same rule can be derived within a further decoupling approximation. We study all methods numerically for fully asymmetric Sherrington-Kirkpatrick models, varying the data length, system size, temperature, and external field. Good convergence is observed in accordance with the theoretical expectations.

DOI: 10.1103/PhysRevLett.110.210601 PACS numbers: 05.10.−a, 02.50.Tt, 75.10.Nr

Introduction.—Inferring interactions between the elements of a network can be posed as an inverse problem in statistical physics in terms of either equilibrium models [1–3] or nonequilibrium ones. The latter has recently gained a lot of attention because of the wider generality and relevance to systems where one has data on the system over time [4,5].

In this connection, the asynchronous kinetic Ising model offers a powerful platform for theoretical insight and practical applications. Under detailed balance (symmetric couplings), it converges to the celebrated maximum entropy equilibrium Ising distribution [6]; that is, the asynchronous model *includes* as a subclass the Gibbs equilibrium Ising model. In many recent works, this equilibrium model is used for inferring functional connectivity and building statistical descriptions, e.g., for neuronal spike trains [2]. However, spike trains and many other real life data come in the form of time series. Since it is only under strict detailed balance that the asynchronous Ising model converges to the equilibrium Ising distribution, it is important to find the relation between the couplings found from the asynchronous model and those from the equilibrium Gibbs distribution. This becomes particularly important for analyzing data using fine time bins at which temporal correlations become important.

The asynchronous Ising model is also important from another perspective. Most of the work on the subject so far has focused on models with only one *type* of stochastic variables. The asynchronous Ising model, however, can be viewed as a doubly stochastic model where, in addition to spin configurations, the update times of the spins are themselves stochastic variables. This differs from the synchronously updated model where all spins are updated at all times, making the spin configurations the only stochastic variables [4]. Doubly stochastic processes are in fact abundant in real life. An example is a securities market [7] where traders place limit orders: conditional offers to buy securities if their market price falls below a threshold or to sell if the market price rises above it. If offers are made, other traders may respond or not; if they do, transactions take place. Whether or not limit offers are placed defines a first set of stochastic variables depending on which transactions may or may not occur, defining a second set.

The presence of two stochastic degrees of freedom raises a number of questions. How can we infer interactions if the data contain only the history of one of them, e.g., the transaction times? How does this compare with the case where everything is known? When do the two scenarios converge? Here, starting from two likelihood functions for the data, one in which update times are known, the other not, we derive two different learning rules. We show that these learning rules have different precisions for inferring the couplings, and that they have a nontrivial relation to each other: averaging over possible update times, they both lead to a third one, but with different learning rates. Surprisingly, this third learning rule can also be derived from the forward equations of motion for the correlations of the asynchronous Ising model [6] and without appealing to a likelihood function. This relates two previously unrelated approaches of learning the couplings. Applying the averaged rule to data from retinal ganglion cells, we find that the connections of the effective asynchronous model are nearly identical to those of the equilibrium Ising model.

Since the learning rules we derive, as opposed to those for the equilibrium Ising model, do not require calculating a partition function and Monte Carlo sampling, the asynchronous model offers a much faster way of inferring functional connectivity.

Kinetic Ising model with asynchronous updates.—Consider N binary spins, $s_i = \pm 1$, $i = 1, \ldots, N$, coupled to each other through a matrix J_{ij} and each subject to an external field θ_i. The coupling matrix need not be symmetric and, consequently, the system may not possess a Gibbs equilibrium state [8]. One can describe this stochastic dynamical system in either of two ways.

(1) Consider a time discretization with steps of size δt. At each step, update spin i with probability $\gamma_i \delta t$, where γ_i are constants with dimension of inverse time. We assume γ_i to be known *a priori*, not a parameter of the model to be determined. For simplicity, we also assume $\gamma_i = \gamma$ for all i, but all our derivations follow in the general case as well. By "update" we mean assigning a new value $s_i(t + \delta t)$ with probability $[1 + s_i(t + \delta t) \tanh H_i(t)]/2 = \exp[s_i(t + \delta t) H_i(t)]/2 \cosh H_i(t)$, where $H_i(t) = \theta_i + \sum_j J_{ji} s_j(t)$ is the total field acting on spin i at time t. Of course, the new value $s_i(t + \delta t)$ may be equal to the old one; updating does not necessarily mean flipping. Multiple spins can be updated in one time step, but for $\delta t \ll 1$ (the limit we consider) in most steps at most one spin is updated. The synchronously updated model is recovered when $\gamma \delta t = 1$. Thus, one can interpolate between the synchronous and asynchronous models by varying γ. In this formulation, the model is doubly stochastic: the dynamics of one set of stochastic variables (the spins) are conditional on the dynamics of the other (the updates). Here we set the temperature that conventionally appears in this model equal to 1, because it can be absorbed into the definitions of the fields and couplings. Equivalently, our fields and couplings are in units of temperature.

(2) Start from the Glauber master equation [6]. Then at every step every spin is flipped with a probability $\gamma \delta t[1 - s_i(t) \tanh H_i(t)]/2$. As in scheme (1), multiple spins can flip in a single time step, but this happens with probability of order $(\delta t)^2$. Thus, if $\delta t \ll 1$, in most time intervals at most one spin is flipped.

The difference between the schemes is that in scheme (1) we have two sets of random variables, the update times (which we denote by $\{\tau_i\}$) and the spin histories $\{s_i(t)\}$, while scheme (2) contains only the $\{s_i(t)\}$. One can easily show that marginalizing out the $\{\tau_i\}$ in scheme (1) leads exactly to scheme (2), even if $\gamma \delta t$ is not small. Thus, all averages over histories involving spins only (i.e., not involving the update times) will be the same in the two schemes. Nevertheless, knowing "the history of the system" (i.e., a realization of its stochastic evolution) means something different in the two schemes. In the first we know all the update times, while in the second we know only those at which the updated spins flipped. We will see below that knowing these extra data influences the performance in reconstructing the couplings. Which scheme is relevant for inferring the couplings from data depends on the specific nature of the system being modeled and the data available. The "update times" may be meaningful and, if so, available in some cases and not in others.

Two likelihoods to maximize.—Consider scheme (1) above. Suppose we are given a history of the system, i.e., the data $s \equiv \{s_i(t)\}$ and $\tau \equiv \{\tau_i\}$, of length $L = T/\delta t$ steps, and we are asked to reconstruct the couplings and fields. We do this by maximizing the likelihood $P(s, \tau) = P(s|\tau)p(\tau)$ over these parameters. For each spin i, the τ_i are a (discretized) Poisson process; i.e., every t has probability $\gamma \delta t$ of being a member of the set τ. Thus the probability of the update history $p(\tau)$ is independent of the model parameters, and we can take as the objective function $\log P(s|\tau)$, i.e.,

$$\mathcal{L}_1 = \sum_i \sum_{\tau_i} [s_i(\tau_i + \delta t) H_i(\tau_i) - \log 2 \cosh H_i(\tau_i)].$$

This is just like the synchronous-update case except that the sum over times is only over the update times. It leads to a learning rule:

$$\delta J_{ij} \propto \frac{\partial \mathcal{L}_1}{\partial J_{ij}} = \sum_{\tau_i} [s_i(\tau_i + \delta t) - \tanh(H_i(\tau_i))] s_j(\tau_i). \quad (1)$$

Defining $J_{i0} = \theta_i$, $s_0(t) = 1$, this equation also includes the learning rule for θ_i. We call this algorithm "spin- and update-history-based." or "SUH".

In scheme (2), we know only the spin history, not the update times. Since this scheme is equivalent to the first one with the τ_i marginalized out, we treat it by maximizing $P(s) = \sum_\tau P(s|\tau)p(\tau)$ [9], leading to

$$\mathcal{L}_2 = \sum_{i,t} \log\left[(1 - \gamma \delta t)\delta_{s_i(t+\delta t), s_i(t)} + \gamma \delta t \frac{e^{s_i(t+\delta t) H_i(t)}}{2 \cosh H_i(t)}\right],$$

as the objective function. Separating terms with and without spin flips, the resulting learning rules will be

$$\delta J_{ij} \propto \frac{\partial \mathcal{L}_2}{\partial J_{ij}} = \sum_{\text{flips}} [s_i(t + \delta t) - \tanh(H_i(t))] s_j(t) + \frac{\gamma \delta t}{2} \sum_{\text{no flips}} q_i(t) s_i(t + \delta t) s_j(t), \quad (2)$$

where $q_i(t) \equiv 1 - \tanh^2(H_i(t))$, and it includes the rule for the θ_i with the convention $J_{i0} = \theta_i$, $s_0(t) = 1$. We call this the "spin-history-only" (SHO) algorithm.

Reconstruction errors for both algorithms can be calculated by analyzing the Fisher information matrices. For SHO the Fisher matrix elements read

$$-\frac{\partial^2 \mathcal{L}_2}{\partial J_{ij}\partial J_{kl}} = \delta_{ik}\sum_{\text{flips}} q_i(t)s_j(t)s_l(t) + 2\delta_{ik}\gamma\delta t \times \sum_{\text{no flips}} q_i(t)s_i(t+\delta t)\tanh(H_i(t))s_j(t)s_l(t). \tag{3}$$

In the weak coupling limit, this matrix has nonzero elements only for $j = l$, and the mean value of these nonzero elements yields the inverse of the mean square reconstruction error (MSE). Without external fields, the second term in Eq. (3) vanishes; thus, the MSE in this case is $2/(T\gamma)$, noting that the probability that a time step is a flip is $\gamma\delta t/2$. For SUH the calculation is analogous, and for $\theta_i = 0$ and weak couplings, the MSE will be $(T\gamma)^{-1}$, i.e., a factor of 2 smaller than for SHO.

History-averaged learning.—SUH and SHO utilize explicitly their respective full model histories, both $\{s_i(t)\}$ and τ_i for SUH and $\{s_i(t)\}$ for SHO. Below we derive a third rule by averaging the one for SUH, Eq. (1), over all update histories. Defining $C_{ij}(t) \equiv \langle s_i(t_0 + t)s_j(t_0)\rangle$, we have

$$\dot{C}_{ij}(t) = \lim_{\delta t\to 0}\frac{\langle s_i(t+\delta t)s_j(t_0)\rangle - \langle s_i(t)s_j(t_0)\rangle}{\delta t},$$

where $\langle\cdots\rangle$ means an average over all realizations of the stochastic dynamics. Separating time steps into those at which an update occurred and those at which no update occurred yields

$$\dot{C}_{ij}(t) = \lim_{\delta t\to 0}\left\{\gamma\delta t\frac{\langle s_i(\tau_i+\delta t)s_j(t_0)\rangle_{\tau_i} - \langle s_i(\tau_i)s_j(t_0)\rangle_{\tau_i}}{\delta t}\right\}.$$

There is no contribution from steps with no flip because then $s_i(t + \delta t) = s_i(t)$ and the numerator would be zero. Thus we have expressed the average over all realizations of the first term in Eq. (1) in terms of spin correlation functions and their time derivatives:

$$\langle s_i(\tau_i+\delta t)s_j(\tau_i)\rangle_{\tau_i} = \frac{1}{\gamma}\dot{C}_{ij}(0) + C_{ij}(0). \tag{4}$$

In averaging the second term in Eq. (1), the average over $\{\tau_i\}$ can be replaced by an average over all times, since the quantity $\tanh H_i(t)s_j(t)$ is insensitive to whether an update is being made. Thus, averaging Eq. (1) over all possible histories yields

$$\delta J_{ij} \propto \gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle\tanh(H_i(t))s_j(t)\rangle. \tag{5}$$

We will refer to the update rule given by Eq. (5) as the averaged-SUH rule, or AVE. This rule has the same structure as the one for the synchronous-update model [4], with $\langle s_i(t+1)s_j(t)\rangle$ replaced by $C(0) + \gamma^{-1}\dot{C}(0)$.

AVE requires knowing the equal-time correlations, their derivatives at $t = 0$, and $\langle\tanh(H_i(t))s_j(t)\rangle$. This latter quantity depends on the model parameters (through $H_i(t)$), so, in practice, estimating it at each learning step requires knowing the entire spin history, the same data as SHO learning needs.

Can we derive an algorithm like Eq. (5) from SHO learning by averaging over spin flip times in the same way we did by averaging SUH learning over update times? Denote the local fields at time t generated by the true model (the one that generated the data) by $\tilde{H}_i(t)$, and, as before, the local field calculated using the inferred parameters as $H_i(t)$. At each time step t, then, the probability of flipping spin i is $\gamma\delta t[1 - s(t)\tanh\tilde{H}_i(t)]/2$. We thus have to allot the first term in Eq. (2) a weight $\gamma\delta t[1 - s(t)\tanh\tilde{H}_i(t)]/2$ and the second a weight $1 - \gamma\delta t[1 - s(t)\tanh\tilde{H}_i(t)]/2 \approx 1$, getting

$$\delta J_{ij} \propto \left\langle\frac{\partial\mathcal{L}_1}{\partial J_{ij}}\right\rangle_0 = \frac{\gamma}{2T}\int dt[\tanh\tilde{H}_i(t) - \tanh H_i(t)] \times [1 + s_i(t)\tanh H_i(t)]s_j(t). \tag{6}$$

The learning thus converges when the discrepancy $\tanh(H(t)) - \tanh(\tilde{H}(t))$ is zero. Noting also that the arguments above leading to Eq. (4) yield $\langle\tanh\tilde{H}(t)s_j(t)\rangle_t = \gamma^{-1}\dot{C}(0) + C(0)$, we write Eq. (6) as

$$\delta J_{ij} \propto \gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle\tanh H_i(t)s_j(t)\rangle_t + \langle[\tanh\tilde{H}_i(t) - \tanh H_i(t)]s_i(t)\tanh H_i(t)s_j(t)\rangle_t. \tag{7}$$

The first line is identical to Eq. (5). We can obtain a learning rule heuristically by an *ad hoc* factorization of the average in the second line as $\langle[\tanh\tilde{H}_i(t) - \tanh H_i(t)]s_i(t)\tanh H_i(t)\times s_j(t)\rangle_t \approx \langle\tanh\tilde{H}_i(t) - \tanh H_i(t)s_j(t)\rangle_t\langle s_i(t)\tanh H_i(t)\rangle_t$, yielding

$$\delta J_{ij} \propto [\gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle\tanh H_i(t)s_j(t)\rangle_t] \times \langle[1 + s_i(t)\tanh H_i(t)]\rangle_t. \tag{8}$$

This just amounts to varying the learning rate in Eq. (5) proportional to the time-averaged probability of not flipping according to the model. Thus we arrive by a different route at the AVE rule, Eq. (5).

We compared the performance of the algorithms SUH, SHO, and AVE to each other and to the naive mean-field (nMF) and Thouless-Anderson-Palmer (TAP) approximations to AVE investigated in [5] for fully asymmetric Sherrington-Kirkpatrick models [10]. The couplings are zero-mean independent and identically distributed normal variables with variance g^2/N (J_{ij} is independent of J_{ji}). We study these for different values of g and θ, the system size N, and the data length L. As a performance measure, we use the MSE on the J_{ij}.

Figure 1 shows the performance of the algorithms. As anticipated above, the error for SUH is half of that for SHO learning; see Fig. 1(a). The same panel also shows that AVE and SHO appear to perform equally well for large enough L. In retrospect, this is not surprising, since both

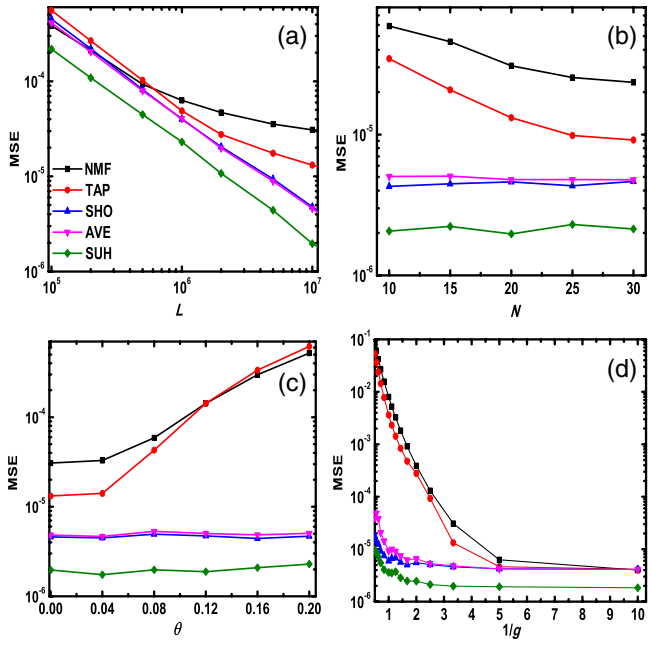

FIG. 1 (color online). Mean square error (MSE) versus (a) data length L, (b) system size N, (c) external field θ, and (d) temperature $1/g$. Black squares show nMF, red circles TAP, blue up triangles SHO, pink down triangles AVE, and green diamonds SUH, respectively. The parameters are $g = 0.3$, $N = 20$, $\theta = 0$, $L = 10^7$ except when varied in a panel.

algorithms effectively use the same data (the spin history). For small L, the averaging that yields AVE from SHO may be prone to fluctuations yielding the two learning rules behaving differently. Figure 1(b) shows that the MSE for the exact algorithms is insensitive to N, while the approximate algorithms improve as N becomes larger [note, however, the opposite trend in Fig. 1(a)]; in these calculations, the average numbers of updates and flips per spin were kept constant, taking $L = 5 \times 10^5 N$. Figure 1(c) shows that the performance of the three exact algorithms is also not sensitive at all to θ, while nMF and TAP work noticeably less well with a nonzero θ. Finally, the effects of (inverse) g are depicted in Fig. 1(d). For fixed L, all the algorithms do worse at strong couplings (large g). The nMF and TAP do so in a much clearer fashion at smaller g, growing approximately exponentially with g for g greater than ≈ 0.2. In the weak coupling limit, all algorithms perform roughly similarly, except that SUH enjoys its factor-2 advantage (conferred by knowledge of the update times), as already seen in Fig. 1(a).

We applied the learning rule Eq. (5) to spike trains from 20 retinal ganglion cells and compared the inferred couplings with those of the Gibbs equilibrium model (see Supplemental Material for details [11]). Figure 2(a) shows that the Gibbs equilibrium and kinetic Ising couplings are very similar. Furthermore, the asynchronous model allows the inference of self-couplings (diagonal elements of the coupling matrix) which are not present in the equilibrium

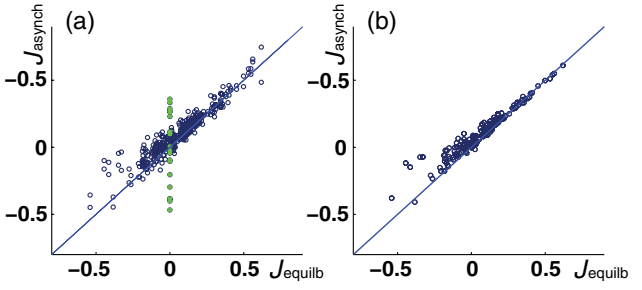

FIG. 2 (color online). Asynchronous versus equilibrium couplings for retinal data. (a) The full asynchronous model. Green stars show the self-couplings which by convention are equal to zero for the equilibrium model. (b) The results when at every iteration the self-couplings were put to zero and the matrix was symmetrized.

model. This result provides a rationale for the use of the maximum entropy equilibrium Ising model: if the asynchronous couplings were very different form the equilibrium ones, it would have meant that the real dynamical process did not satisfy the Gibbs equilibrium conditions and that the final distribution of states is not the Gibbs equilibrium Ising model. In fact, we also tested what happens to the couplings of the asynchronous model if during learning we symmetrized the couplings matrix at each iteration by adding its transpose to itself and dividing by two and also putting the self-couplings to zero. Figure 2(b) shows that the resulting couplings now get even closer to the equilibrium ones. Since inferring the equilibrium model is an exponentially difficult problem, requiring time-consuming Monte Carlo sampling, these results have an important pragmatic consequence for inferring retinal functional connectivity. This is because the asynchronous approach does not require Monte Carlo sampling: the averages on the right-hand side of Eq. (5) are all over the data. The asynchronous learning rules thus allow the inference of functional connections that for the retinal data largely agree with the maximum entropy equilibrium model, but the inference is much faster.

Discussion.—A surprising observation is that Eq. (5), that we derived by maximizing the likelihood, can also be derived from a totally different route. For a kinetic Ising model, the equation of motion for the correlations given θ and J is $\gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) = \langle \tanh H_i(t) s_j(t) \rangle_t$ [6]. This equation holds for correct couplings, and thus a heuristic learning is given by just adjusting the couplings proportional to the difference of the two sides. This again yields Eq. (5), and the linearized version of it would, in fact, be the mean-field inference algorithm for the asynchronous model used in [5]. Our results show that this rule is not merely heuristic: it can be derived starting from the likelihood of the data, whether assuming that update times are known or not, and averaging over the update times.

Here we addressed the problem of inferring the couplings in a nonequilibrium system: the asynchronous,

asymmetrically coupled kinetic Ising model. We showed how to derive three different learning algorithms, utilizing three different levels of detail of the history of the system: the full spin and update history, the spin history only, and spin correlations at and near $t = 0$ only. The methods show performance that is promising in practical terms, agrees with theoretical expectations, and, in particular, is superior to approximate methods found earlier. We expect that the reasoning behind our results on deriving and relating different learning rules can be extended to a variety of inverse statistical mechanics problems beyond the particular case of the kinetic Ising model.

This work has been supported by the Finnish graduate school for Computational Science (FICS), the Academy of Finland as part of its Finland Distinguished Professor program project 129024/Aurell, and the Centers of Excellence COMP and COIN, as well as NORDITA and the Kavli Foundation. The authors acknowledge Manfred Opper for discussions and Michael Berry for providing the retinal data.

*hong.zeng@aalto.fi

[1] H. J. Kappen and F. B. Rodriguez, Neural Comput. **10**, 1137 (1998); T. Tanaka, Phys. Rev. E **58**, 2302 (1998); Y. Roudi, J. Tyrcha, and J. Hertz, *ibid.* **79**, 051915 (2009); Y. Roudi, E. Aurell, and J. A. Hertz, Front. Comput. Neurosci. **3**, 22 (2009); E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).

[2] E. Schneidman, M. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006); J. Shlens, G. Field, J. Gauthier, M. Grivich, D. Petrusca, A. Sher, A. Litke, and E. Chichilnisky, J. Neurosci. **26**, 8254 (2006); S. Cocco, S. Leibler, and R. Monasson, Proc. Natl. Acad. Sci. U.S.A. **106**, 14 058 (2009).

[3] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).

[4] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli, Nature (London) **454**, 995 (2008); Y. Roudi and J. Hertz, Phys. Rev. Lett. **106**, 048702 (2011); J. Stat. Mech. (2011) P03031; M. Mezard and J. Sakellariou, J. Stat. Mech. (2011) L07001; I. Mastromatteo and M. Marsili, J. Stat. Mech. (2011) P10012; J. Tyrcha, Y. Roudi, M. Marsili, and J. Hertz, J. Stat. Mech. (2013) P03005; P. Zhang, J. Stat. Phys. **148**, 502 (2012).

[5] H.-L. Zeng, E. Aurell, M. Alava, and H. Mahmoudi, Phys. Rev. E **83**, 041135 (2011).

[6] R. J. Glauber, J. Math. Phys. (N.Y.) **4**, 294 (1963).

[7] A. Ranaldo, J. Financ. Market. **7**, 53 (2004); S. Maslov, Physica (Amsterdam) **278A**, 571 (2000).

[8] D. Gillespie, J. Phys. Chem. **81**, 2340 (1977).

[9] C. Kipnis and C. Landim, *Scaling Limits of Interacting Particle Systems* (Springer-Verlag, Berlin, 1999), Vol. 320.

[10] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[11] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.110.210601 for details of fitting the model to the retinal data.