Journal of Complex Networks (2017) **00**, 1–22
doi: 10.1093/comnet/cnx032

Loss of information in feedforward social networks

SIMON STOLARCZYK AND MANISHA BHARDWAJ
Department of Mathematics, University of Houston, Houston, TX 77204, USA

KEVIN E. BASSLER
Department of Mathematics, University of Houston, Houston, TX 77204, USA, Department of Physics, University of Houston, Houston, TX 77204, USA and Texas Center for Superconductivity, University of Houston, 77204, USA

WEI JI MA
Center for Neural Science and Department of Psychology, New York University, New York, NY 10003, USA

AND

KREŠIMIR JOSIĆ†
Department of Mathematics, University of Houston, Houston, TX 77204, USA, Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA and Department of BioSciences, Rice University, Houston, TX 77251, USA
†Corresponding author. Email: josic@math.uh.edu

Edited by: Mason Porter

[Received on 28 March 2017; editorial decision on 15 July 2017; accepted on 5 September 2017]

We consider social networks in which information propagates directionally across layers of rational agents. Each agent makes a locally optimal estimate of the state of the world, and communicates this estimate to agents downstream. When agents receive some information from a common source their estimates are correlated. We show that the resulting redundancy can lead to the loss of information about the state of the world across layers of the network, even when all agents have full knowledge of the network's structure. A simple algebraic condition identifies networks in which information loss occurs, and we show that all such networks must contain a particular network motif. We also study random networks asymptotically as the number of agents increases, and find a sharp transition in the probability of information loss at the point at which the number of agents in one layer exceeds the number in the previous layer.

1. Introduction

While there are billions of people on the planet, we exchange information with only a small fraction of them. How does information propagate through such social networks, shape our opinions, and influence our decisions? How do our interactions impact our choice of career or candidate in an election? More generally, how do we as agents in a network aggregate noisy signals to infer the state of the world?

These questions have a long history. The general problem is not easy to describe using a tractable mathematical model, as it is difficult to provide a reasonable probabilistic description of the state of the world. We also lack a full understanding of how perception [1, 2], and the information we exchange [3]

shapes our decisions. Progress has therefore relied on tractable idealized models that mimic some of the main features of information exchange in social networks.

Early models relied on computationally tractable interactions, such as the majority rule assumed in Condorcet's Jury Theorem [4], or local averaging assumed in the DeGroot model [5]. More recent models rely on the assumption of rational (Bayesian) agents who use private signals, measurements or observations of each other's actions to maximize utility. Such models of information sharing are often used in the economics literature, sometimes in combination with ideas from game theory. For instance, in a series of papers Mossel, Tamuz and collaborators considered the propagation of information on an undirected network of rational agents, and showed that all agents on an irreducible graph integrate information optimally in a finite number of steps [6]. A similar setup was used by Acemoglu *et al.* [7] to examine herd behaviour in a network. Mueller-Frank [8] considered model social networks where private information of each agent is represented by a finite partition of the state space, and showed that in networks of non-Bayesian agents information is typically not aggregated optimally, but optimality is achieved in the presence of a single Bayesian agent [9]. These, and related works (reviewed in [10]), refer to such abstract models as "social networks", and we follow this convention for simplicity. However, we note that this is at odds with the more traditional definition of this term [11].

Simplified models about how information is exchanged are also used in the political science literature to explain tendencies observed in social groups, and to fit to data. For example, Ortoleva and Snowberg used dependent Gaussian random variables to model the experimentally observed neglect of redundancies in information received by human observers [12]. They used this model to show how neglect of correlations can explain overconfidence in a sample of 3000 adults from the 2010 Cooperative Congressional Election Study (CCES) [13]. On the other hand, Levy and Razin show that similar correlation neglect can also lead to positive outcomes, as observers rely on actual information in forming opinions, rather than political orientation [14].

Such social network models of information propagation are generally either sequential or iterative. In sequential models, agents are ordered and act in turn based on a private signal and the observed action of their predecessors [15, 16]. In iterative models, agents make a single or a sequence of measurements, and iteratively exchange information with their neighbours [6, 17]. Sequential models have been used to illustrate information cascades [18], while iterative models have been used to illustrate agreement and learning [19].

Here we consider a sequential model in which information propagates directionally through layers of rational agents. The agents are part of a structured network, rather than a simple chain. As in the sequential model, we assume that information transfer is directional, and the recipient does not communicate information to its source. This assumption could describe the propagation of information via print or any other fixed medium.

We assume that at each step, a layer of agents receive information from those in a previous layer. This is different from previous sequential models where agents received information in turn from all their predecessors as in [15, 20–22]. Importantly, the same information can reach an agent via multiple paths. Therefore, information received from agents in the previous layer can be redundant. Unlike in models of information neglect [13], we assume that agents take into account these redundancies in making decisions. We show that, depending on the network structure, even rational agents with full knowledge of the network structure cannot always resolve these redundancies. As a result, an estimate of the state of the world can degrade over layers. We also show that network architectures that lead to information loss can amplify an agent's bias in subsequent layers.

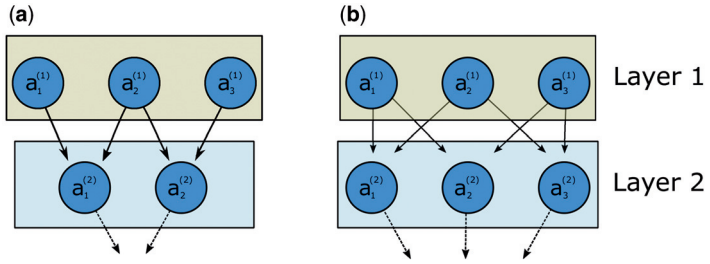

FIG. 1. Illustration of the general setup. Agents in the first layer (top layer in the figure) make measurements, x_1, x_2 and x_3, of a parameter s. In each layer agents make an estimate of this parameter, and communicate it to agents in the subsequent layer. Arrows indicate the direction in which information is propagated. We show that information about s degrades across layers in the network in panel (a), but not in the network in (b).

As an example, consider the network in Fig. 1(a). We assume that the first-layer agents make measurements x_1, x_2 and x_3 of the state of the world, s, and that these measurements are normally distributed with equal variance. This assumption means that minimum-variance unbiased estimators for these parameters are always linear combinations of individual measurements [23]. Each agent makes an estimate, $\hat{s}_1^{(1)}, \hat{s}_2^{(1)}$ and $\hat{s}_3^{(1)}$, of s. The superscript and subscript refer to the layer and agent number, respectively. An agent with global access to all first-layer estimates would be able to make the optimal (minimum-variance) estimate $\hat{s}_{\text{ideal}} = \frac{1}{3}\left(\hat{s}_1^{(1)} + \hat{s}_2^{(1)} + \hat{s}_3^{(1)}\right)$ of s.

All agents in the first layer then communicate their estimates to one or both of the second-layer agents. These in turn use the received information to make their own estimates, $\hat{s}_1^{(2)} = \frac{1}{2}(\hat{s}_1^{(1)} + \hat{s}_2^{(1)})$ and $\hat{s}_2^{(2)} = \frac{1}{2}(\hat{s}_2^{(1)} + \hat{s}_3^{(1)})$. An agent receiving the two estimates from the second layer then takes their linear combination to estimate s. However, in this network no linear combination of the locally optimal estimates, $\hat{s}_1^{(2)}$ and $\hat{s}_2^{(2)}$, equals the best estimate, $\hat{s}_{\text{ideal}}$, obtainable from all measurements in the first layer. Indeed,

$$\hat{s} = \beta_1 \hat{s}_1^{(2)} + \beta_2 \hat{s}_2^{(2)} = \beta_1 \left(\hat{s}_1^{(1)} + \hat{s}_2^{(1)}\right) + \beta_2 \left(\hat{s}_2^{(1)} + \hat{s}_3^{(1)}\right) \neq \hat{s}_{\text{ideal}} = \frac{1}{3}\left(\hat{s}_1^{(1)} + \hat{s}_2^{(1)} + \hat{s}_3^{(1)}\right),$$

with the inequality holding for any choice of β_1, β_2. Moreover, assume the estimates of first-layer agents are biased, and $\hat{s}_i^{(1)} = x_i + b_i$. If the other agents are unaware of this bias, then, as we will show, the final estimate is $\hat{s} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) \cdot (\hat{s}_1^{(1)} + b_1, \hat{s}_2^{(1)} + b_2, \hat{s}_3^{(1)} + b_3) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) \cdot \hat{s}^{(1)} + (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) \cdot (b_1, b_2, b_3)$. Thus the bias of the second agent in the first layer, $a_2^{(1)}$, has disproportionate weight in the final estimate.

In this example the information about the state of the world, s, available from second-layer agents is less than that available from first-layer agents. In the preceding example the measurement x_2 is used by both agents in the second layer. The estimates of the two second-layer agents are therefore correlated, and the final agent cannot disentangle them to recover the ideal estimate. We will show that the type of subgraph shown in Fig. 1(a), which we call a *W-motif*, provides the main obstruction to obtaining the best estimate in subsequent layers.

2. The model

We consider feedforward networks having n layers and identify each node of a network with an agent. The structure of the network is thus given by a directed graph with agents occupying the vertices. Agents in each layer only communicate with those in the next layer. For convenience, we will assume that layer n consists of a single agent that receives information from all agents in layer $n-1$. This final agent in

the last layer therefore makes the best estimate based on all the estimates in the next-to-last layer. We will use this last agent's estimate to quantify information loss in the network. Two example networks are given in Fig. 1, with the single agent in the final, third layer not shown.

We assume that all agents are Bayesian, and know the structure of the network. Every agent estimates an unknown parameter, $s \in \mathbb{R}$, but only the agents in the first layer make a measurement of this parameter. Each agent makes the best possible estimate given the information it receives and communicates this estimate to a subset of agents in the next layer. We also assume that measurements, x_i, made by agents in the first layer are independent and normally distributed with mean s, and variance σ_i^2, that is $x_i \sim \mathcal{N}(s, \sigma_i^2)$. Furthermore, every agent in the network knows the variance of each measurement in the first layer, σ_i^2. Also, for simplicity, we will assume that all agents share an improper, flat prior over s. This assumption does not affect the main results.

An agent with access to all of the measurements, $\{x_i\}_i$, has access to all the information available about s in the network. This agent can make an *ideal* estimate, $\hat{s}_{\text{ideal}} = \text{argmax}_s\, p(s|x_1, ..., x_n)$. We assume that the actual agents in the network are making locally optimal, maximum-likelihood estimates of s, and ask when the estimate of the final agent equals the ideal estimate, $\hat{s}_{\text{ideal}}$.

Individual estimate calculations

Each agent in the first layer only has access to its own measurement, and makes an estimate equal to this measurement. We therefore write $\hat{s}_i^{(1)} = x_i$. We denote the jth agent in layer k by $a_j^{(k)}$. Each of these agents makes an estimate, $\hat{s}_j^{(k)}$ of s, using the estimates communicated by its neighbours in the previous layer. Under our assumptions, the posterior computed by any agent is normal and the vector of estimates in a layer follows a multivariate Gaussian distribution. As agents in the second layer and beyond can share upstream neighbours, the covariance between their estimates is typically non-zero. We show that under the assumption that the variance of the initial measurements and the structure of the network are known to all agents, each agent knows the full joint posterior distribution over s for all agents it receives information from.

Weight matrices

We define the connectivity matrix $C^{(k)}$ for $1 \leq k \leq n-1$ as,

$$C_{ij}^{(k)} = \begin{cases} 1, & \text{if } a_j^{(k)} \text{ communicates with } a_i^{(k+1)} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

An agent receives a subset of estimates from the previous layer determined by this connectivity matrix. The agent then uses this information to make its own, maximum-likelihood estimate of s. By our assumptions, this estimate will be a linear combination of the communicated estimates [23]. Denoting by $\hat{\mathbf{s}}^{(k)}$ the vector of estimates in the kth layer, we can therefore write $\hat{\mathbf{s}}_i^{(k+1)} = \mathbf{w}_i^{(k+1)} \cdot \hat{\mathbf{s}}^{(k)}$, and

$$\hat{\mathbf{s}}^{(k+1)} = W^{(k+1)} \hat{\mathbf{s}}^{(k)}.$$

Here $W^{(k+1)}$ is a matrix of weights applied to the estimates in the kth layer.

Weighting by precision

We can write $\hat{\mathbf{s}}^{(1)} = W^{(1)}\mathbf{x}$ where $W^{(1)}$ is the identity matrix and $\mathbf{x}$ is the vector of measurements made in the first layer. We assume that all measurements have finite, non-zero variance. Using standard estimation theory results [23], we can compute the optimal estimates for agents in the second layer. Defining $w_i := \frac{1}{\sigma_i^2}$, we can calculate $W^{(2)}$ entrywise: $w_{ij}^{(2)}$ is 0 if agent $a_i^{(2)}$ does *not* communicate with $a_j^{(1)}$. Otherwise $w_{ij}^{(2)} = \frac{w_j^{(1)}}{\sum_{k\to i} w_k^{(1)}}$, where the sum is taken over all agents in the first layer that communicate with agent $a_i^{(2)}$. Therefore,

$$\hat{\mathbf{s}}^{(2)} = W^{(2)}\,\hat{\mathbf{s}}^{(1)} = W^{(2)}W^{(1)}\mathbf{x}\,. \tag{2.2}$$

Covariance matrices

The estimates in the second layer and beyond can be correlated. Let L_k be the number of agents in the kth layer and for $2 \le k \le n-1$ define $\Omega^{(k)} = (\xi_{ij}^{(k)})$ as the $L_k \times L_k$ covariance matrix of estimates in the kth layer,

$$\xi_{ij}^{(k)} = \mathrm{Cov}(\hat{s}_i^{(k)}, \hat{s}_j^{(k)}).$$

When all of the weights are known, we have

$$\hat{\mathbf{s}}^{(k)} = W^{(k)}\hat{\mathbf{s}}^{(k-1)} = W^{(k)}W^{(k-1)}\hat{\mathbf{s}}^{(k-2)} = \cdots = \left(\prod_{l=0}^{k-2} W^{(k-l)}\right)\hat{\mathbf{s}}^{(1)}. \tag{2.3}$$

The ith row of $\left(\prod_{l=0}^{k-2} W^{(k-l)}\right)$ is the vector of weights that the agent $a_i^{(k)}$ applies to the first-layer estimates, since its entries are the coefficients in $s_i^{(k)}$.

The complete covariance matrix, $\Omega^{(k)}$, can therefore be written as

$$\begin{aligned}\Omega^{(k)} &= \mathrm{Cov}(\hat{\mathbf{s}}^{(k)}) = \mathrm{Cov}(W^{(k)}\hat{\mathbf{s}}^{(k-1)}) = W^{(k)}\,\mathrm{Cov}(\hat{\mathbf{s}}^{(k-1)})\,\left(W^{(k)}\right)^{\mathrm{T}} \\ &= \left(\prod_{l=0}^{k-2} W^{(k-l)}\right)\mathrm{Cov}(\hat{\mathbf{s}}^{(1)})\left(\prod_{l=0}^{k-2} W^{(k-l)}\right)^{\mathrm{T}} \\ &= \left(\prod_{l=0}^{k-2} W^{(k-l)}\right)\mathrm{Diag}\left(\frac{1}{w_1},\cdots,\frac{1}{w_{L_1}}\right)\left(\prod_{l=0}^{k-2} W^{(k-l)}\right)^{\mathrm{T}}.\end{aligned} \tag{2.4}$$

Now the ith agent in layer $k \ge 3$, $a_i^{(k)}$, can use $\Omega^{(k-1)}$ to calculate $\mathbf{w}_i^{(k)}$. If the agent is not connected to all agents in the $(k-1)$th layer, it uses the submatrix of $\Omega^{(k-1)}$ with rows and columns corresponding to the agents in the previous layer that communicate their estimates to it. We denote this submatrix $R_i^{(k-1)}$. As in [24], we assume that we remove edges from the graph so that all submatrices $R_i^{(k-1)}$ are invertible, but all estimates are the same as in the original network.

An agent thus receives estimates that follow a multivariate normal distribution, $\mathcal{N}(\hat{\mathbf{s}}_{j\to i}^{(k-1)}, R_i^{(k-1)})$, see [23]. The weights assigned by agent $a_i^{(k)}$ to the estimates of agents in the previous layer are therefore (see also [24]),

$$\tilde{\mathbf{w}}_i^{(k)} = \frac{\mathbf{1}^{\mathrm{T}} \left(R_i^{(k-1)}\right)^{-1}}{\mathbf{1}^{\mathrm{T}} \left(R_i^{(k-1)}\right)^{-1} \mathbf{1}}. \tag{2.5}$$

We define $\mathbf{w}_i^{(k)}$ by using the corresponding entries from $\tilde{\mathbf{w}}_i^{(k)}$ and setting the remainder to zero. In the following, we describe the maximum-likelihood estimate that can be made from all the estimates in a layer. For simplicity, we denote this final estimate by $\hat{s}$. The following results are standard [23].

PROPOSITION 1 The posterior distribution over s of the final agent is normal with

$$\hat{s} = \frac{\mathbf{1}^{\mathrm{T}} (\Omega^{(n-1)})^{-1}}{\mathbf{1}^{\mathrm{T}} (\Omega^{(n-1)})^{-1} \mathbf{1}} \hat{\mathbf{s}}^{(n-1)} \quad \text{and} \quad Var\,[\hat{s}] = \frac{1}{\mathbf{1}^{\mathrm{T}} (\Omega^{(n-1)})^{-1} \mathbf{1}}, \tag{2.6}$$

where $\Omega^{(n-1)}$ is defined by Equations (2.4) and (2.5). Here $\hat{s}$ is the maximum-likelihood, as well as minimum-variance, unbiased estimate of s.

It follows from Equation (2.3) that the estimate of any agent in the network is a convex linear combination of the estimates in the first layer.

Examples

Returning to the example in Fig. 1(a) we have

$$C^{(1)} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},\ W^{(2)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix},\ \Omega^{(2)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix},\ (\Omega^{(2)})^{-1} = \frac{16}{3} \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

The final agent applies the weights $W^{(3)} = \left(\frac{1}{2}, \frac{1}{2}\right)$ to the estimates from the second layer. We thus have the final estimate $\hat{s} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) \cdot \hat{\mathbf{s}}^{(1)}$ with Var $[\hat{s}] = \frac{3}{8}$. The variance of the ideal estimate is $\frac{1}{3}$.

On the other hand, the final agent in the example in Fig. 1(b) makes an ideal estimate: Here $W^{(2)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, $\Omega^{(2)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$, and after inverting $\Omega^{(2)}$ we see that applying a weight of $\frac{1}{3}$ to every agent in the second layer gives the ideal estimate, $\hat{s} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \cdot \hat{\mathbf{s}}^{(1)}$.

REMARK If the agents have a proper normal prior with mean χ and variance σ_p^2, then agents in the first layer make the estimate,

$$\hat{s}_i^{(1)} = \frac{\sigma_i^{-2}}{\sigma_i^{-2} + \sigma_p^{-2}} x_i + \frac{\sigma_p^{-2}}{\sigma_i^{-2} + \sigma_p^{-2}} \chi,$$

with a similar form in the following layers. This does not change the subsequent results as long as all agents have the same prior. Also, if each agent in the network makes a measurement, the general ideas remain unchanged.

3. Results

We ask what graphical conditions need to be satisfied so that the agent in the final layer makes an ideal estimate. That is, when does knowing all estimates of the agents in the $(n-1)$st layer give an estimate that is as good as possible given the measurements of all first-layer agents. We refer to a network in which the final estimate is ideal as an *ideal* network.

PROPOSITION 2 A network with n layers and $\sigma_i^2 \neq 0$ for $i = 1, \ldots, L_1$, is ideal if and only if the vector of inverse variances, $(w_1, ..., w_{L_1})$, is in the row space of the weight matrix product $(\prod_{l=0}^{n-3} W^{(n-1-l)})$.

Proof. In this setting the ideal estimate is

$$\hat{s}_{\text{ideal}} = \frac{1}{\sum_i w_i} \sum_{i=1}^{L_1} w_i \hat{s}_i^{(1)}. \tag{3.1}$$

The network is ideal if and only if there are coefficients $\beta_j \in \mathbb{R}$ such that

$$\hat{s}_{\text{ideal}} = \sum_{j=1}^{L_{n-1}} \beta_j \hat{s}_j^{(n-1)}.$$

Matching coefficients with Equation (3.1), we need

$$\frac{1}{\sum_j w_j} \sum_{i=1}^{L_1} w_i \hat{s}_i^{(1)} = \left(\beta_1, ..., \beta_{L_{n-1}}\right) \cdot \hat{\mathbf{s}}^{(n-1)},$$

or equivalently,

$$\begin{aligned} \frac{1}{\sum_j w_j} \left(w_1, ..., w_{L_1}\right) \cdot \hat{\mathbf{s}}^{(1)} &= \left(\beta_1, ..., \beta_{L_{n-1}}\right) \cdot W^{(n-1)} \hat{\mathbf{s}}^{(n-2)} \\ &= \left(\beta_1, ..., \beta_{L_{n-1}}\right) \cdot \left(\prod_{l=0}^{n-3} W^{(n-1-l)}\right) \hat{\mathbf{s}}^{(1)}. \end{aligned}$$

Equality holds exactly when $(w_1, ..., w_{L_1})$ is in the row space of $\left(\prod_{l=0}^{n-3} W^{(n-1-l)}\right)$. □

In particular, a three-layer network with $\sigma_i^2 = \sigma$ for all $i \in \{1, \ldots, L_1\}$ is ideal if and only if the vector $\vec{1} = (1, 1, ..., 1)$ is in the row space of the connectivity matrix $C^{(1)}$ defined by Equation (2.1). We will use and extend this observation below.

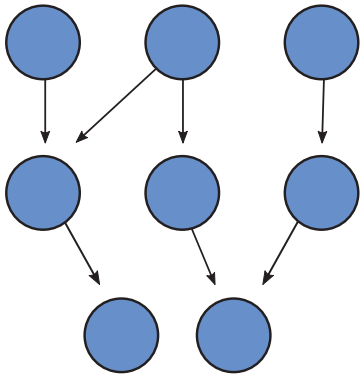

FIG. 2. A W-motif spanning three layers.

3.1 *Graphical conditions for ideal networks*

We say that a network contains a *W-motif* if two agents downstream receive common input from a first-layer agent, as well as private input from two distinct first-layer agents. Examples are shown in Figs 1(a) and 2. A rigorous definition follows.

We will show that *all networks that are not ideal contain a W-motif*. However, the converse is not true: The network in Fig. 1(b) contains many W-motifs, but is ideal. Therefore ideal networks can contain a W-motif, as the redundancy introduced by a W-motif can sometimes be resolved. Hence, additional graphical conditions determine if the network is ideal.

As shown in Fig. 2, in a W-motif there is a directed path from a single agent in the first layer to two agents in the third layer. There are also paths from distinct first-layer agents to the two third-layer agents. This general structure is captured by the following definitions.

DEFINITION 1 The path matrix P^{kl}, $l < k$, from layer l to layer k is defined by,

$$P_{ij}^{kl} = \begin{cases} 1, & \text{if there is a directed path from agent } a_j^{(l)} \text{ to agent } a_i^{(k)} \\ 0, & \text{otherwise.} \end{cases}$$

DEFINITION 2 A network contains a W-motif if a path matrix from the first layer, P^{k1}, has a 2×3 submatrix equal to $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ (modulo column permutation). Graphically, two agents in layer k are connected to one common, and two distinct agents in layer 1.

THEOREM 1 A non-ideal network in which every agent communicates its estimate to the subsequent layer must contain a W-motif. Equivalently, if there are no W-motifs, then the network is ideal.

The proof of this theorem can be found in Appendix A. Intuitively, any agent receives estimates that are a linear combination of first-layer measurements. If there are no W-motifs, any two estimates are either obtained from disjoint sets of measurements, or the measurements in the estimate of one agent contain the measurements in the estimate of another. When measurements are disjoint, there are no correlations between the estimates and thus no degradation of information. When one set of measurements contains

the other, then the estimates in the subset are redundant and can be discarded. Therefore, this redundant information does not cause a degradation of the final estimate.

3.2 *Sufficient conditions for ideal three-layer networks*

We next consider only three-layer networks. This allows us to give a graphical interpretation of the algebraic condition describing ideal networks in Proposition 2. To do so, we will use the following corollary of the proposition.

COROLLARY 1 Let $C^{(1)}$ be defined as in Equation (2.1). Then a three-layer network is ideal if and only if the vector $m\vec{1}$ is in the row space of $C^{(1)}$ over $\mathbb{Z}$ for some non-zero $m \in \mathbb{N}$.

The proof is straightforward and provided in Appendix B for completeness. Note that the corollary is not restricted to the case where first-layer agents have equal variance measurements; whether the network is ideal or not depends entirely on the connection matrix $C^{(1)}$. The ith row of the matrix $C^{(1)}$ corresponds to the inputs of agent $a_i^{(2)}$, and the sum of the jth column is the out-degree of agent $a_j^{(1)}$. Therefore, Corollary 1 is equivalent to the following: If each second-layer agent applies equal integer weights to all of its received estimates, then a three-layer network is ideal if and only if, for some choice of weights, the weighted out-degrees of all agents in the first layer are equal. Hence, we have the following special case:

COROLLARY 2 A three-layer network is ideal if all first-layer agents have equal out-degree in each connected component of the network restricted to the first two layers.

In the connected network in Fig. 1(a), the second agent in the first layer has greater out-degree than the others, while the agents in the first layer of the connected network in Fig. 1(b) have equal out-degree.

Some row reduction operations can be interpreted graphically. Let g be the *input-map* which maps an agent, $a_i^{(2)}$, to the subset of agents in the first layer that it receives estimates from. Formally, let $\mathcal{P}(A)$ denote the power set of a set A, then $g\colon \{a_1^{(2)}, \ldots, a_{L_2}^{(2)}\} \to \mathcal{P}\{a_1^{(1)}, \ldots, a_{L_1}^{(1)}\}$ is defined by $a_j^{(1)} \in g(a_i^{(2)})$ if agent $a_j^{(1)}$ communicates with agent $a_i^{(2)}$, *that is* if $C_{ij}^{(1)} = 1$.

If $g(a_i^{(2)}) \subseteq g(a_j^{(2)})$ for some $i \neq j$, then some of the information received by $a_j^{(2)}$ is redundant, as it is already contained in the estimate of agent $a_i^{(2)}$. We can then reduce the network by eliminating the directed edges from $g(a_i^{(2)})$ to $a_j^{(2)}$, so that in the reduced network $g(a_i^{(2)}) \cap g(a_j^{(2)}) = \emptyset$. This reduction is equivalent to subtracting row i from row j of $C^{(1)}$ resulting in a connection matrix with the same row space. By Proposition 2, the reduced network is ideal if and only if the original network is ideal. This motivates the following definition.

DEFINITION 3 A three-layer network is said to be reduced if $g(a_i^{(2)})$ is not a subset of $g(a_j^{(2)})$ for all $1 \leq i \neq j \leq L_2$.

Reducing a network eliminates edges, and results in a simpler network structure. In a three-layer network, this will not affect the final estimate: Since reduction leaves the row space of $C^{(1)}$ unchanged, the final estimate in the reduced and unreduced network is the result of applying the same weights to the first-layer estimates. This reduction procedure often simplifies identification of ideal networks to a counting of out-degrees (see Corollary 2).

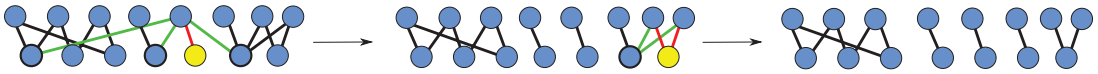

FIG. 3. Example of a two-step network reduction. It is difficult to tell whether the network on the left is ideal. However, after the reduction, all first-layer agents in each of the five connected components have equal out-degree. The network is therefore ideal.

Example

In Fig. 3, we illustrate a two-step reduction of a network. In both steps, an agent (colored differently) has an input set which is overlapped by the input sets of some other second-layer agents (with bolded borders). We use this to cancel the common inputs to the bolded agents and simplify the network. In the first step, note that the lighter agent receives input (in a lighter shade) from a single first-layer agent. We use this to remove all of the other connections (in the lightest shade) emanating from this first-layer agent. In the second step, we again see that the lighter agent receives input (in the medium shade) that is overlapped by input to the agent next to it. We can thus remove the redundant inputs (in the lightest shade) to the bolded agent. The reduced network has 5 connected components all containing vertices with equal out-degree. Hence, this network is ideal by Corollary 2.

3.3 *Variance and bias of the final estimate*

We next consider how the variance and bias of the estimate in layer n depend on the network structure. By definition, the variance of the ideal estimate is $\mathrm{Var}(\hat{s}) = \left(\sum_{i=1}^{L_1} w_i\right)^{-1}$. If the variances of the individual estimates are bounded above as the size of the network increases, the final estimate in an ideal network is *consistent*: As the number of measurements increases the final estimate converges in probability to the true value of s [23]. We next show that the final estimate in non-ideal networks is not necessarily consistent. We also show that biases of certain first-layer agents can have a disproportionate impact on the bias of the final estimate.

Example (variance maximizing network structure)

Figure 4 shows an example of a network structure for which the variance of the final estimate converges to a positive number as the number of agents in the first layer increases. We assume that all first-layer agents make measurements with unit variance. We will show that as the number of agents in both layers increases, the variance of the final estimate approaches $1/4$. Let the estimate of the central agent be $s_1^{(1)}$. Then each agent in the second layer makes an estimate $\frac{1}{2}(s_1^{(1)} + s_i^{(1)})$ for some $i \neq 1$. By symmetry the single agent in the last layer averages all estimates from the second layer to obtain $\hat{s} = \frac{1}{2}(s_1^{(1)} + \frac{1}{L_1 - 1}\sum_{i=2}^{L_1} s_i^{(1)})$. Therefore, the estimate of the central agent (which communicates with all agents in the second layer) receives a much higher weight than all other estimates from the first layer. The variance of the final estimate thus equals

$$\mathrm{Var}(\hat{s}) = \frac{1}{4} + \frac{1}{4(L_1 - 1)}.$$

Hence, the final estimate is not consistent, as its variance remains positive as the number of first-layer agents, L_1, diverges. Given a restriction on the number of second-layer agents, we show that this network leads to the highest possible variance of the final estimate:

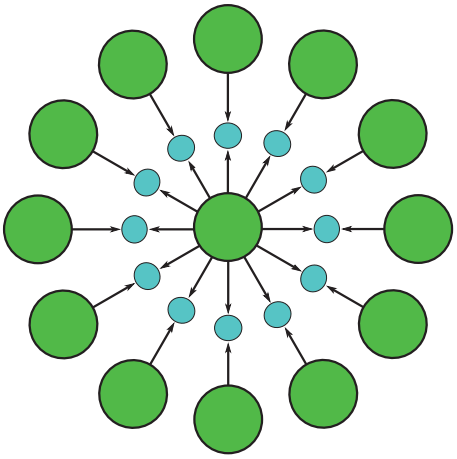

FIG. 4. Example of a network with an inconsistent final estimate. The larger and smaller nodes represent agents in the first and second layer, respectively. Each second-layer agent receives input from the common, central agent and a distinct first-layer agent, and thus $L_2 = L_1 - 1$.

PROPOSITION 3 The final estimate in the network in Fig. 4 has the largest variance among all three-layer networks with a fixed number $L_1 \geq 4$ of first-layer, and $L_2 \geq L_1 - 1$ second-layer agents, assuming that every first-layer agent makes at least one connection.

The idea of the proof is to limit the possible out-degrees of the agents in the first layer and show that the structure in Fig. 4 has the highest variance for this restriction. The proof is provided in Appendix C.

In general, we conjecture that for the final estimate to have large variance, some agents upstream must have a disproportionately large out-degree, with the remaining agents making few connections. On the other hand, as the in-degree of a second-layer agent increases, the variance of its estimate shrinks. Thus when a few agents communicate information to many, the resulting redundancy is difficult to resolve downstream. But when downstream agents receive many estimates, we expect the estimates to be good. We next show that the biases of the agents with the highest out-degrees can have an outsized influence on the estimates downstream.

Propagation of biases

We next ask how biases in the measurements of agents in the first layer propagate through the network. Ideally, such biases would be averaged out in subsequent layers. To simplify the analysis we assume constant, additive biases, $\hat{s}_i^{(1)} = x_i + b_i$, with the constant bias, b_i. Downstream agents are unaware of these biases, and therefore assume them to be zero. Since all estimates in the network are convex linear combinations of first-layer measurements, the final estimate will have the form

$$\hat{s} = \sum \alpha_i \, (x_i + b_i) = \sum \alpha_i x_i + \sum \alpha_i b_i, \tag{3.2}$$

and thus will have finite bias bounded by the maximum of the individual biases.

We have provided examples of network structures where the estimate of a first-layer agent was given higher weight than others, even when all first-layer measurements had equal variance. Equation (3.2)

shows that this agent's bias will also be disproportionately represented in the bias of the final estimate. Indeed, in the example in Fig. 1(a), the estimate of second agent in first layer has weight $\frac{1}{2}$, and its bias will have twice the weight of the other agents in the final estimate. Similarly, the bias of the central agent in Fig. 4 will account for half the bias of the final estimate as $n \to \infty$. Thus even if the biases, b_i, are distributed randomly with zero mean, the asymptotic bias of the final estimate does not always disappear as the number of measurements increases.

More generally, networks that contain W-motifs can result in biases of first-layer agents with disproportionate impact on the final estimate. As with the variance, we conjecture that the bias of agents that communicate their estimates to many agents downstream will be disproportionately represented in the final estimate. Equivalently, if the network contains agents that receive many estimates, we expect the bias of the final estimate to be reduced.

3.4 *Inference in random feedforward networks*

We have shown that networks with specific structures can lead to inconsistent and asymptotically biased final estimates. We now consider networks with randomly and independently chosen connections between layers. Such networks are likely to contain many W-motifs, but it is unclear whether these motifs are resolved and whether the final estimate is ideal. We will use results of random matrix theory to show that there is a sharp transition in the probability that a network is ideal when the number of agents from one layer exceeds that of the previous layer [25].

We assume that connections between agents in different layers are random, independent and made with fixed probability, p. We will use the following result of [26], also discussed by [25]:

THEOREM 2 (Komlos) Let ξ_{ij}, $i,j = 1,\ldots,n$ be i.i.d. with non-degenerate distribution function $F(x)$. Then the probability that the matrix $X = (\xi_{ij})$ is singular converges to 0 with the size of the matrix,

$$\lim_{n\to\infty} P(\det X = 0) = 0.$$

COROLLARY 3 For a three-layer network with independent, random, equally probable ($p = 1/2$) connections from first to second-layer, as the number of agents L_1 and L_2 increases,

$$\frac{L_1}{L_2} \leq 1 \implies P(\hat{s} = \hat{s}_{\text{ideal}}) \to 1,$$

and

$$\frac{L_1}{L_2} > 1 \implies P(\hat{s} = \hat{s}_{\text{ideal}}) \to 0.$$

The proof is given in Appendix D. The same proof works when $L_1/L_2 \leq 1$ and the probability of a connection is arbitrary, $p \in (0, 1]$. We conjecture that the result also holds for $L_1/L_2 > 1$ and arbitrary p, but the present proof relies on the assumption that $p = 1/2$. Figure 5 shows the results of simulations which support this conjecture: The different panels correspond to different connection probabilities, and the curves to different numbers of agents in the first layer. As the number of agents in the second layer exceeds that in the first, the probability that the network is ideal approaches 1 as the number first-layer agents increases. With 100 agents in the first layer, the curve is approximately a step function for all connection probabilities we tested.

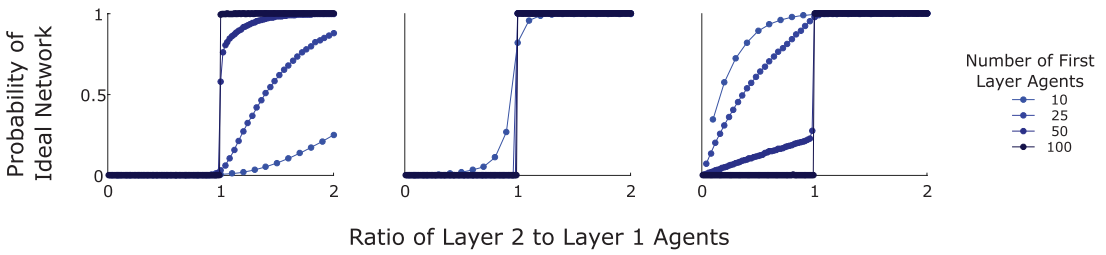

FIG. 5. The probability that a random, three-layer network is ideal for connection probabilities $p = 0.1$ (left), 0.5 (centre) and 0.9 (right). In each panel, the different curves correspond to different, but fixed numbers of agents in the first layer. The number of agents in the second layer is varied. There is a sharp transition in the probability that a network is ideal when the number of agents in the second layer exceeds the number in the first. Simulation details can be found in Appendix E.

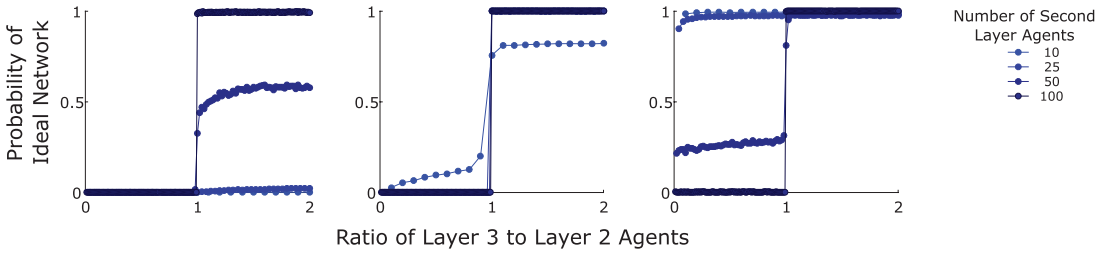

FIG. 6. The probability that a random, four-layer network is ideal for connection probabilities $p = 0.1$ (left), 0.5 (centre) and 0.9 (right). Each curve corresponds to equal, fixed numbers of agents in the first two layers, with a changing number of agents in the third layer. Simulation details can be found in Appendix E.

More than 3 layers

We conjecture that a similar result holds for networks with more than three layers:

CONJECTURE For a network with n layers with independent, random, equally probable connections between consecutive layers, as the total number of agents increases,

$$L_k \leq L_{k+1} \text{ for } 1 \leq k < n-1 \implies P(\hat{s} = \hat{s}_{\text{ideal}}) \to 1$$

and

$$L_1 > L_k \text{ for some } 1 < k < n \implies P(\hat{s} = \hat{s}_{\text{ideal}}) \to 0.$$

Figure 6 shows the results with four-layer networks with different connection probabilities across layers. The number of agents in the first and second layers are equal, and we varied the number of agents in the third layer. The results support our conjecture.

With multiple layers ($n \geq 4$) , if $L_1 > L_2$ then the network will not be ideal as in the limit the estimate of s will not be ideal already in the second layer by Corollary 3. If the number of agents does not decrease across layers, we conjecture that the probability that information is lost across layers is small when the number of agents is large. Indeed, it seems reasonable that the products of the random weight matrices

will be full rank with increasing probability allowing us to apply Proposition 2. However, the entries in these matrices are no longer independent, so classical results of random matrix theory no longer apply.

4. Conclusion

We examined how information about the world propagates through layers of rational agents. We assumed that at each step, a group of agents makes an inference about the state of the world from information provided by their predecessors. The setup is related, but different from information cascades where a chain of rational agents make decisions in turn [15, 20–22], or recurrent networks where agents exchange information iteratively [6]. The assumption that the observed variables in our analysis follow a Gaussian distribution simplified the analysis considerably. However, we believe that the main results hold under more general assumptions. Our preliminary work shows that when agents in the first layer make a Boolean measurement the presence of W-motif is necessary to prevent ideal information propagation. For more general measurements, for instance a sample from the exponential family of distribution, a non-linear estimator would be needed, and the analysis becomes more complicated.

Related results have been obtained by Acemoglu, *et al.* [7] who considered social networks in which individuals receive information from a random neighbourhood of agents. They show that agents can make the right choice, or infer the correct state of the world as network size increases when a finite group of agents does not account for most of the information that is propagated through the network. However, the setting of this study is somewhat different from ours: Agents are assumed to only observe each other's actions, but do not share their belief about the binary state of the world.

We translated the question about whether the estimate of the state of the world degrades across layers in the network to a simple algebraic condition. This allowed us to use results of random matrix theory in the case of random networks, find equivalent networks through an intuitive reduction process, and identify a class of networks in which estimates do not degrade across layers, and another class in which degradation is maximal.

Networks in which estimates degrade across layers must contain a W-motif. This motif introduces redundancies in the information that is communicated downstream and may not be removed. Such redundancies, also known as 'bad correlations,' are known to limit the information that can be decoded from neural responses [27, 28]. This suggests that agents with large out-degrees and small in-degrees can hinder the propagation of information, as they introduce redundant information in the network. On the other hand, agents with large in-degrees integrate information from many sources, which can help improve the final estimate. However, the detailed structure of a network is important: For example, an agent with large in-degree in the second layer can have a large out-degree without hindering the propagation of information as it has already integrated most available first-layer measurements.

To make the problem tractable, we have made a number of simplifying assumptions. We made the strong assumption that agents have full knowledge of the network structure. Some agents may have to make several calculations in order to make an estimate, so we also do not assume bounded rationality [29]. This is unlikely to hold in realistic situations. Even when making simple decisions, pairs of agents are not always rational [3]: When two agents each make a measurement with different variance, exchanging information can degrade the better estimate.

The assumption that only agents in the first layer make a measurement is not crucial. We can obtain similar results if all agents in the network make independent measurements, and the information is propagated directionally, as we assume here. However, in such cases, the confidence (inverse variance of the estimates) typically becomes unbounded across layers.

Funding

NSF-DMS-1517629 to S.S. and K.J., NSF/NIGMS-R01GM104974 to K.J., NSF-DMR-1507371 K.B. and NSF-IOS-1546858 to K.B.

References

1. Brunton, B. W., Botvinick, M. M. & Brody, C. D. (2013) Rats and humans can optimally accumulate evidence for decision-making. *Science*, **340**, 95–98.
2. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012) Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, **74**, 30–39.
3. Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G. & Frith, C. D. (2010) Optimally interacting minds. *Science*, **329**, 1081–1085.
4. de Condorcet, M. (1976) *Essay on the Application of Analysis to the Probability of Majority Decisions*. (K. M. Baker, ed.). Paris: Imprimerie Royale, 1785. Reprinted in Condorcet: Selected Writings.
5. DeGroot, M. H. (1974) Reaching a consensus. *J. Acoust Soc. Amer.*, **69**, 118–121.
6. Mossel, E., Sly, A. & Tamuz, O. (2014) Asymptotic learning on Bayesian social networks. *Probab. Theory Related Fields*, **158**, 127–157.
7. Acemoglu, D., Dahleh, M. A., Lobel, I. & Ozdaglar, A. (2011) Bayesian learning in social networks. *Rev. Econ. Stud.*, **78**, 1201–1236.
8. Mueller-Frank, M. (2013) A general framework for rational learning in social networks. *Theor. Econ.*, **8**, 1–40.
9. Mueller-Frank, M. (2014) Does one Bayesian make a difference? *J. Econ. Theory*, **154**, 423–452.
10. Golub, B. & Sadler, E. D. Learning in Social Networks. Available at SSRN: https://ssrn.com/abstract=2919146 (February 16, 2017).
11. Wasserman, S. & Faust, K. (1994) Social network analysis: Methods and applications. Cambridge: Cambridge University Press.
12. Enke, B. & Zimmermann, F. (2013) Correlation Neglect in Belief Formation (November 29, 2013). CESifo Working Paper Series No. 4483.
13. Ortoleva, P. & Snowberg, E. (2015) Overconfidence in political behavior. *Amer. Econ. Rev.*, **105**, 504–535.
14. Levy, G. & Razin, R. (2015) Correlation neglect, voting behavior, and information aggregation. *Amer. Econ. Rev.*, **105**, 1634–1645.
15. Banerjee, A. V. (1992) A simple model of herd behavior. *Q. J. Econ.*, 797–817.
16. Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.*, 992–1026.
17. Gale, D. & Kariv, S. (2003) Bayesian learning in social networks. *Games Econom. Behav.*, **45**, 329–346.
18. Bikhchandani, S., Hirshleifer, D. & Welch, I. (1998) Learning from the behavior of others: Conformity, fads, and informational cascades. *J. Econ. Perspect.*, **12**, 151–170.
19. Mossel, E. & Tamuz, O. (2014) Opinion exchange dynamics. *arXiv preprint arXiv:1401.4770*.
20. Easley, D. & Kleinberg, J. (2010) *Networks, Crowds, and Markets*, vol. 1. New York: Cambridge University Press.
21. Welch, I. (1992) Sequential sales, learning, and cascades. *J. Finance*, **47**, 695–732.
22. Bharat, K. & Mihaila, G. A. (2001) When experts agree: using non-affiliated experts to rank popular topics. *Proceedings of the 10th International Conference on World Wide Web*. New York, NY, USA: ACM, pp. 597–602.
23. Kay, S. M. (1993) *Fundamentals of Statistical Signal Processing*, vol. 1. Estimation Theory. Englewood Cliffs, N.J.: PTR Prentice-Hall.
24. Mossel, E., Olsman, N. & Tamuz, O. (2016) Efficient bayesian learning in social networks with gaussian estimators. In Communication, Control, and Computing (Allerton), 54th Annual Allerton Conference on IEEE, pp. 425–432.

25. Bollobás, B. (2001) *Random Graphs*. Number 73 in Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press.
26. Komlós, J. (1968) On the determinant of random matrices. *Stud. Sci. Math. Hung.*, **3**, 387–399.
27. Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P. & Pouget, A. (2014) Information-limiting correlations. *Nat. Neurosci.*, **17**, 1410–1417.
28. Bhardwaj, M., Carroll, S., Ma, W. J. & Josić, K. (2015) Visual decisions in the presence of measurement and stimulus correlations. *Neural Comput.*, **27**, 2318–2353.
29. Bala, V. & Goyal, S. (1998) Learning from neighbours. *Rev. Econ. Stud.*, **65**, 595–621.

Appendix A. Proof of Theorem 1

We start with the simpler case of a W-motif between the first two layers and then extend it to the general case. We begin with definitions that will be used in the proof.

Let g be the *input-map* which maps an agent to the subset of agents in the first layer that it receives information from (through some path). That is, $g(a_i^{(j)})$ is the set of agents in the first layer that provide input to $a_i^{(j)}$. It is intuitive—and we show it formally in Lemma A1—that a network contains a W-motif if each of the inputs to two agents, A and B are not contained in the other, and their intersection is not empty. That is, $g(A) \not\subseteq g(B)$ and $g(B) \not\subseteq g(A)$, but $g(A) \cap g(B) \neq \emptyset$. If these conditions are met, we also say that the inputs of A and B have a *non-trivial intersection*. If $g(A) \subseteq g(B)$, we say that the input of B *overlaps* the input of A: every agent which contributes to the estimate of A also contributes to the estimate of B.

Similarly, we let f be the *output-map* which maps an agent, $a_i^{(j)}$, to the set of all agents in the next, $j+1^{\text{st}}$, layer that receive input from $a_i^{(j)}$. We first prove a few lemmas essential to the proof of Theorem 1.

Lemma A1 Assume a network does not contain a W-motif and there are two agents, $a_{i_1}^{(k)}$ and $a_{i_2}^{(k)}$, with $g(a_{i_1}^{(k)}) \cap g(a_{i_2}^{(k)})$ non-empty. Then $g(a_{i_1}^{(k)})$ overlaps or is overlapped by $g(a_{i_2}^{(k)})$.

Proof. We prove the claim by contradiction. If one input does not overlap the other, then there are two distinct first-layer agents $a_{n_1}^{(1)}$ and $a_{n_2}^{(1)}$ such that $a_{n_1}^{(1)} \in g(a_{i_1}^{(k)}) \backslash g(a_{i_2}^{(k)})$ and $a_{n_2}^{(1)} \in g(a_{i_2}^{(k)}) \backslash g(a_{i_1}^{(k)})$. This means $P^{k1}_{i_1 n_1} = P^{k1}_{i_2 n_2} = 1$ and $P^{k1}_{i_1 n_2} = P^{k1}_{i_2 n_1} = 0$. Since the inputs of the agents have non-empty intersection, we also have $P^{k1}_{i_1 m} = P^{k1}_{i_2 m} = 1$ for some m. Thus there is a 2×3 submatrix of P^{k1} which, up to rearrangement of the columns, is equal to $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ and the network contains a W-motif, contrary to assumption. □

Every agent's estimate is a convex linear combination of estimates in the first layer, given by Equation (2.3). We will use the corresponding weight vectors in the following proofs. We show that in networks without W-motifs, agents will only be receiving collections of estimates with weight vectors which pairwise either have disjoint support (non-zero indices) or the support is contained in the support of the other agent. Thus, with no W-motifs, no two agents have inputs with non-trivial intersection. The next two lemmas will allow us to easily calculate the estimates of such agents.

LEMMA A2 Let r, s, t be positive integers, $w_i = \sigma_i^{-2}$, and consider three weight vectors applied by three agents in layer k, $a_1^{(k)}, a_2^{(k)}$ and $a_3^{(k)}$, to the estimates of the first layer:

$$v_1 = \left(\frac{w_1}{\sum_{i=1}^{r} w_i}, \ldots, \frac{w_r}{\sum_{i=1}^{r} w_i}, 0, \ldots, 0 \right)$$
$$v_2 = \left(\frac{w_1}{\sum_{i=1}^{r+s} w_i}, \cdots, \frac{w_{r+s}}{\sum_{i=1}^{r+s} w_i}, 0, \ldots, 0 \right)$$
$$v_3 = \left(0, \ldots, 0, \frac{w_{r+s+1}}{\sum_{i=r+s+1}^{r+s+t} w_i}, \ldots, \frac{w_{r+s+t}}{\sum_{i=r+s+t}^{r+s+t} w_i}, 0, \ldots, 0 \right).$$

An agent $a_i^{(k+1)}$ in $f(a_1^{(k)}) \cap f(a_2^{(k)})$, but not in $f(a_3^{(k)})$, will use weight vector v_2. An agent $a_i^{(k+1)}$ in $f(a_2^{(k)}) \cap f(a_3^{(k)})$, but not $f(a_1^{(k)})$, will use weight vector

$$v_4 = \left(\frac{w_1}{\sum_{i=1}^{r+s+t} w_i}, \ldots, \frac{w_{r+s+t}}{\sum_{i=1}^{r+s+t} w_i}, 0, ..., 0 \right).$$

Proof. First, consider an agent receiving the first two estimates with weights v_1 and v_2. Suppose that a fictitious agent receives a collection of estimates with weight vectors $\{z_1, ..., z_{r+s}\}$, where $z_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, *that is*, each estimate equals the measurement of agent $a_i^{(1)}$. This fictitious agent can obtain any linear combination of the first $r + s$ measurements. The linear combination with lowest variance has weights given by v_2. Therefore, an agent receiving measurements corresponding to the weight vectors v_1 and v_2 cannot do better than the estimate of agent $a_2^{(k)}$ with weights given by v_2.

A similar argument works when estimates are received from agents $a_2^{(k)}$ and $a_3^{(k)}$. Since these two agents make locally optimal estimates based on non-overlapping sets of measurements in the first layer, the best estimate is obtained by combining the two sets of measurements. This is precisely the estimate corresponding to the weights given by vector v_4. □

LEMMA A3 Suppose an agent, $a_i^{(k)}$, receives a collection of estimates such that for any pair, there is a relabelling of agents in the first layer that makes the pair look like v_1 and v_2 or like v_2 and v_3 in Lemma A2. Then, up to some relabelling of the agents in the first layer, that agent will make an estimate with corresponding weight vector

$$v = \left(\frac{w_1}{\sum_{i=1}^{r} w_i}, \ldots, \frac{w_r}{\sum_{i=1}^{r} w_i}, 0, \ldots, 0 \right).$$

Proof. Let the vectors z_i be defined as in the proof of Lemma A2. Relabel the first-layer agents so that only the first r entries of the weight vector applied by agent $a_i^{(k)}$ are non-zero. Then a fictitious agent receiving estimates with weight vectors $z_i, 1 \leq i \leq r$ can construct any estimate that agent $a_i^{(k)}$ can obtain. The optimal estimate of this fictitious agent has weight vector v. Hence if some linear combination of the weight vectors of estimates communicated to agent $a_i^{(k)}$ equals v, this linear combination defines the best estimate.

Then for each $j = 1, ..., r$, we can find a weight vector, v_j, which is non-zero in the jth entry with support that contains the support of every other weight vector which is non-zero in the jth entry. Such a vector exists by the assumption that any two vectors have disjoint support or the support of one contains the other. Therefore, we can find the weight vector with maximal support for each entry. If we take the

distinct elements of $\{v_j : 1 \leq j \leq r\}$, then these maximal weight vectors will have disjoint support that partitions the first r indices. Therefore,

$$v = \frac{1}{\sum_{i=1}^{r} w_i} \sum_{v_j \text{ distinct}} \left(\sum_{i=1, v_j^i \text{ non-zero}}^{r} w_i \right) v_j,$$

which shows the lemma. □

We now state and prove the three-layer case of Theorem 1 and then use it to finish the proof of Theorem 1.

PROPOSITION A1 If a three-layer network is not ideal and every first-layer agent communicates with at least one second-layer agent, then the network must contain a W-motif.

Proof. Assume the network does not contain a W-motif. Given a first-layer agent $a_i^{(1)}$, Lemma A1 says that for any two agents in $f(a_i^{(1)})$, one agent's input must overlap the other. Two second-layer agents thus receive estimates with input sets where one overlaps the other, or the sets do not intersect. Thus the set of weight vectors in the second layer satisfies the assumptions of Lemma A3. As all agents from the first layer communicate with the final agent, the network is ideal. □

To obtain the proof of Theorem 1, we use induction with Proposition A1 as a base case.

Proof of Theorem 1.. Assume the network has n layers, there are no W-motifs, and every agent (except those in the first layer) receives input from at least one other agent. Lemma A1 implies that in the second layer each pair of agents has either disjoint input or one overlaps the other. Thus in the third layer, by relabelling the agents, each agent makes an estimate with weight vector of the form: $\frac{1}{\sum_{i=1}^{r} w_i}(w_1, \ldots, w_r, 0, \ldots, 0)$.

Now assume that any estimate in layer k can be put in this form by relabelling the agents. Since there are no W-motifs, Lemma A1 implies that set of measurements used by agents $a_{i_1}^{(k)}$ and $a_{i_2}^{(k)}$ is disjoint or overlapping. This again allows us to apply Lemma A3 and any agent in layer $k+1$ makes an estimate whose weight vector again has the form $\frac{1}{\sum_{i=1}^{r} w_i}(w_1, \ldots, w_r, 0, \ldots, 0)$. Applying the same argument to the final agent, where every entry will be non-zero in some penultimate-layer agent's weight vector, we have that the network is ideal.

□

Appendix B. Proof of Corollary 1

We will show that a three-layer network is ideal if and only if $m\vec{1}$ is in the row space of $C^{(1)}$ over $\mathbb{Z}$ for some $m \in \mathbb{N}$. We do this by first showing that the network is ideal if and only if $\vec{1}$ is in the row space of $C^{(1)}$ over $\mathbb{R}$, and then we show that this is equivalent to $m\vec{1}$ being in the row space of $C^{(1)}$ over $\mathbb{Z}$.

By Proposition 2, a three-layer network is ideal if and only if $(w_1, \ldots, w_{L_1})$ is in the row space of $W^{(2)}$. We claim that this is equivalent to $\vec{1}$ being in the row space of $C^{(1)}$: Multiplying each row of $W^{(2)}$ by the common denominator of the non-zero entries gives

$$\mathcal{R}(W^{(2)}) = \mathcal{R}(C^{(1)}\text{Diag}(w_1, \ldots, w_{L_1})),$$

where $\mathcal{R}$ denotes the row space. By definition, $\vec{1}$ is a linear combination of the rows of $C^{(1)}$ if and only if

$$1 = \sum_i \beta_i C_{ij}^{(1)}, \quad \forall j.$$

This holds if and only if

$$w_j = \sum_i \beta_i w_j C_{ij}^{(1)}, \quad \forall j.$$

The last equality is equivalent to

$$(w_1, \ldots, w_{L_1}) = \sum_i \beta_i (C^{(1)} \mathrm{Diag}(w_1, \ldots, w_{L_1}))_i,$$

which means $(w_1, \ldots, w_{L_1})$ is in the row space of $W^{(2)}$. Hence, for three-layer networks, the network is ideal if and only if the vector $\vec{1}$ is in the row space of $C^{(1)}$ over $\mathbb{R}$.

Thus it remains to show that $\vec{1} \in \mathcal{R}(C^{(1)})$ over $\mathbb{R}$ is equivalent to $\vec{1} \in \mathcal{R}(C^{(1)})$ over $\mathbb{Z}$. If $m\vec{1} \in \mathcal{R}(C^{(1)})$ over $\mathbb{Z}$, then it is a linear combination of the rows of $C^{(1)}$ with integer coefficients. Multiplying the coefficients of this linear combination by $\frac{1}{m}$ shows that $\vec{1}$ is in the row space of $C^{(1)}$ and hence the network is ideal.

If $\vec{1}$ is in the row space of $C^{(1)}$ over $\mathbb{R}$, then by closure of $\mathbb{Q}^n$ this means there is some linear combination of the rows of $C^{(1)}$ over $\mathbb{Q}$ which is equal to $\vec{1}$:

$$\sum_{i=1}^{L_2} \alpha_i C_i^{(1)} = \vec{1}, \qquad \alpha_i \in \mathbb{Q}.$$

Multiplying both sides by the absolute value of the product of the denominators of the non-zero α_i shows that

$$\sum_{i=1}^{L_2} \beta_i C_i^{(1)} = m\vec{1}, \qquad \beta_i \in \mathbb{Z}$$

for some $m \in \mathbb{N}$ and thus $m\vec{1}$ is in the row space of $C^{(1)}$ over $\mathbb{Z}$.

Appendix C. Proof of Proposition 3

We will show that the network architecture that maximizes the variance of the final estimate for a given number of first and second-layer agents is the one shown in Fig. 4. To simplify notation we write $L_1 = n$ and $L_2 = m$.

LEMMA C1 If $\mathbf{d} = (d_1, ..., d_n)$ is the vector of out-degrees in the first layer, so $d_i = |f(a_i^{(1)})|$, then to maximize the variance of the final estimate, $\mathbf{d}$ must equal $(m, 1, \ldots, 1)$, up to relabelling.

Proof of Claim. Given a network structure consider the naïve estimate:

$$\frac{1}{Z} \sum_i |g(a_i^{(2)})| \hat{s}_i^{(2)} = \frac{1}{\sum_{ij} C_{ij}^{(1)}} \sum_i C_i^{(1)} \cdot \hat{\mathbf{s}}^{(1)}, \tag{C.1}$$

where Z is a normalizing factor that makes the entries of the corresponding vector of weights sum to 1. This estimate can always be made and is the same as using a linear combination of estimates of agents $a_j^{(1)}$ with weights $\frac{d_i}{\sum_{j=1}^n d_j}$. Thus the variance of the optimal estimate of the agent in the final layer is bounded above by the variance of the naïve estimate in Equation (C.1). By assumption $1 \leq d_j \leq m$ for all j. For the network in Fig. 4, this naïve estimate equals the final estimate. Thus it is sufficient to show that the naïve estimate has maximal variance when $\mathbf{d} = (m, 1, \ldots, 1)$, up to relabelling.

The variance, V, of the naïve estimate is:

$$V(d_1, \ldots, d_n) = \sum_j \left(\frac{d_j}{\sum_{k=1}^n d_k} \right)^2.$$

If we treat the degrees as continuous variables then V is continuous on $\mathbf{d} \in [1, m]^n$ and we can calculate the gradient of V to find the critical points.

$$\frac{\partial V}{\partial d_i} = 2 \left(\frac{d_i}{\sum_k d_k} \right) \frac{\sum_k d_k - d_i}{\left(\sum_k d_k \right)^2} + \sum_{j \neq i} 2 \left(\frac{d_j}{\sum_k d_k} \right) \frac{-d_j}{\left(\sum_k d_k \right)^2}.$$

Setting $\frac{\partial V}{\partial d_i} = 0$ and multiplying both sides by $\frac{1}{2} \left(\sum_{k=1}^n d_k \right)^3$ gives

$$0 = d_i (\sum_{k \neq i} d_k) - \sum_{j \neq i} d_j^2 = \sum_{j \neq i} d_j (d_i - d_j).$$

This shows that $d = k\vec{1}$ for $k = 1, \ldots, m$ are the only critical points, since if there exist $d_i \leq d_j$, for all $j \neq i$ and $d_i < d_k$ for some $k \neq i$ then the right hand side would be negative. These critical points are the first-layer out-degrees of ideal networks by Corollary 2, hence they are minima. This implies that V takes on its maximum values on the boundary.

The boundary of $[1, m]^n$ consists of points where at least one coordinate is 1 or m. Since V is invariant under permutation of the variables, we set d_1 equal to one of these values and investigate the behaviour of V on this restricted set.

First set $d_1 = m$. Setting $\frac{\partial V}{\partial d_i}$ to 0 on this boundary gives:

$$0 = m(d_i - m) + \sum_{j \neq i, 1} d_j (d_i - d_j).$$

One critical point is thus $m\vec{1}$. If $d_i \leq d_j$ for $j \neq i$ and $d_i < m$ then again the right hand side would be negative. Hence $d_i = m$ for all i, and there are no critical points on the interior of $\{m\} \times [1, d]^{n-1}$.

Next if $d_1 = 1$, setting $\frac{\partial V}{\partial d_i}$ to 0 on this boundary and multiplying by -1 gives:

$$0 = 1 - d_i + \sum_{j \neq i, 1} d_j (d_j - d_i).$$

Here a critical point is $\vec{1}$. If $d_i \leq d_j$ for $j \neq i$ and $1 < d_i < m$ then again the right hand side would be negative. Hence $d_i = 1$ for all i, and there are no critical points on the interior of $\{1\} \times [1, d]^{n-1}$. If we

iterate this procedure, we see that the maximum value of V must occur on the corners of the hypercube $[1, d]^n$.

Choose one of these corners, $\mathbf{c}$, and, without loss of generality, assume that the first l coordinates are m and the last $n - l$ coordinates are 1, $1 \leq l < n$. Then

$$\begin{aligned} V(\mathbf{c}) &= \sum_{j=1}^{l} \left(\frac{m}{\sum_{k=1}^{n} d_k} \right)^2 + \sum_{j=l+1}^{n} \left(\frac{1}{\sum_{k=1}^{n} d_k} \right)^2 \\ &= \left(\frac{1}{lm + (n-l)} \right)^2 \left(lm^2 + (n-l) \right) \\ &= \frac{lm^2 + n - l}{l^2m^2 + 2lm(n-l) + (n-l)^2} \\ &= \frac{l(m^2-1) + n}{l^2(m-1)^2 + l2n(m-1) + n^2}. \end{aligned}$$

Under the assumption that $m \geq n - 1$, a lengthy algebra calculation that we omit shows that this is maximized for $l = 1$. Hence the maximum value of V is achieved at $(m, 1, \ldots, 1)$, or any of its coordinate permutations.

□

Finally, to have $\mathbf{d} = (m, 1, \ldots, 1)$, one first-layer agent, $a_1^{(1)}$, communicates with all second-layer agents and every other agent has exactly one output. Since there are at least $n - 1$ agents in the second layer, this means that each first-layer agent must communicate with a distinct second-layer agent and each second-layer agent must receive input from $a_1^{(1)}$. Otherwise, some agent in the second layer would receive only the input from $a_i^{(1)}$ and thus the final estimate could use that estimate to decorrelate all of the second-layer estimates.

So, the naïve estimate for an alternative network has smaller variance than the ideal estimate for the ring network in Fig. 4. Hence the final estimate in any alternative network will have smaller variance. Since the only network with $\mathbf{d} = (m, 1, \ldots, 1)$ is the network in Fig. 4, we have shown that this structure maximizes the variance of the final estimate among all networks with $L_2 \geq L_1 - 1$.

Appendix D. Proof of Corollary 3

Whether or not $\hat{s}_{\text{ideal}} = \hat{s}$ is determined by $C^{(1)}$. For simplicity, we drop the superscript and refer to this connectivity matrix as C. By our assumption, this is a random matrix with $P(C_{ij} = 0) = P(C_{ij} = 1) = 1/2$.

First assume that there are at least as many second-layer agents as there are first-layer agents: $L_2 \geq L_1$ or $\frac{L_1}{L_2} \leq 1$. Then C is a random $L_2 \times L_1$ matrix with i.i.d. non-degenerate entries that has more rows than columns. By Theorem 2, this means that the $L_1 \times L_1$ submatrix formed by the first L_1 rows and columns is non-singular with probability approaching 1 as $L_1, L_2 \to \infty$. Thus the probability that the row space of C contains the vector $\vec{1}$ converges to 1 with the size of the network.

Next assume that there are fewer second-layer agents than first-layer agents, that is $L_2 < L_1$ or $\frac{L_1}{L_2} > 1$. We will show that the probability that the row space of C contains $\vec{1}$ goes to zero as $L_1, L_2 \to \infty$. Since increasing the number of rows will not decrease the probability that C contains a vector in its row space

we assume that $L_2 = L_1 - 1$ and let $L_1 = n$:

$$\lim_{L_1, L_2 \to \infty} P(\hat{s} = \hat{s}_{\text{ideal}}) \leq \lim_{n \to \infty} P(\vec{1} \in R(C(n-1, n))),$$

where $C(n-1, n)$ refers to the random matrix as before, and identifies that it has $n-1$ rows and n columns. We first use:

$$P(\vec{1} \in R(C(n-1, n))) \leq P(\begin{pmatrix} \vec{1} \\ C \end{pmatrix} \text{ is singular})$$

since if $\vec{1}$ is the row space of C, then attaching that row of ones to it would create a singular matrix.

LEMMA D1 $P\left(\det\left(\begin{pmatrix} \vec{1} \\ C \end{pmatrix}\right) = 0\right) \to 0$ as $n \to \infty$.

We can rewrite $C = \begin{pmatrix} B & \mathbf{v} \end{pmatrix}$, where $\mathbf{v}$ is the nth column of C and B is the remaining submatrix. We claim

$$\det\left(\begin{pmatrix} \vec{1} \\ C \end{pmatrix}\right) = -1^k \det\left(\begin{pmatrix} \vec{1} & 1 \\ \tilde{B} & \vec{0} \end{pmatrix}\right) = -1^{k+n+1} * \det(\tilde{B}), \tag{D.1}$$

where $\tilde{B}$ is a random $(n-1) \times (n-1)$ matrix distributed like C. Assuming this claim, then by [26] :

$$P\left(\det\left(\begin{pmatrix} \vec{1} \\ C = 0 \end{pmatrix}\right)\right) = P\left(\det(\tilde{B}) = 0\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Thus $P(\vec{1} \in R(M(n-1, n))) \to 0$ as $n \to \infty$.

To prove the first equality in Equation (D.1), we use row operations on $\begin{pmatrix} \vec{1} & 1 \\ B & \mathbf{v} \end{pmatrix}$: If $v_i = 1$ then subtract the first row from the ith row, $(B_i \ v_i)$, to get a vector whose entries are all 0 and -1. Then $(B_i \ v_i) \to -(\tilde{B}_i \ 0)$ where $(\tilde{B}_i \ 0)$ is a vector of entries which are again either 0 or 1 with equal probability. We do this for every row which has a 1 in its last entry and multiply the determinant a factor -1 and denote the number of these reductions as k. Since $P(C_{ij} = 0) = \frac{1}{2}$ we also have $P(\tilde{B}_{ij} = 0) = \frac{1}{2}$.

Appendix E. Details of simulations

All simulations were done in MATLAB. For the three-layer networks, we randomly generated binary connection matrices and tested whether or not the vector $\vec{1}$ was in the row space. Each point in the plots corresponds to the number of agents in the first two layers for a given connection probability and was generated using at least 10,000 samples. The code used for these simulations can be found at the repository https://github.com/Spstolar/FFNetInfoLoss.