Sparse code of conflict in a primate society

Bryan C. Daniels[a,b,1], David C. Krakauer[a,b], and Jessica C. Flack[a,b]

[a]Santa Fe Institute, Santa Fe, NM 87501; and [b]Center for Complexity and Collective Computation, Wisconsin Institute for Discovery, University of Wisconsin, Madison, WI 53716

Animals living in groups collectively produce social structure. In this context individuals make strategic decisions about when to cooperate and compete. This requires that individuals can perceive patterns in collective dynamics, but how this pattern extraction occurs is unclear. Our goal is to identify a model that extracts meaningful social patterns from a behavioral time series while remaining cognitively parsimonious by making the fewest demands on memory. Using fine-grained conflict data from macaques, we show that sparse coding, an important principle of neural compression, is an effective method for compressing collective behavior. The sparse code is shown to be efficient, predictive, and socially meaningful. In our monkey society, the sparse code of conflict is composed of related individuals, the policers, and the alpha female. Our results suggest that sparse coding is a natural technique for pattern extraction when cognitive constraints and small sample sizes limit the complexity of inferential models. Our approach highlights the need for cognitive experiments addressing how individuals perceive collective features of social organization.

cognition | statistical inference | collective computation | social complexity | social niche construction

Individuals in primate societies strategically join fights (1, 2), form coalitions and alliances (3–5), predict conflict participants, and mitigate escalating aggression (6–8). Consider the following examples.

Pigtailed macaques (*Macaca nemestrina*) preferentially solicit support from powerful individuals, suggesting they can predict which individuals are most effective at providing support or terminating conflict (9). Pigtailed macaques have been shown to reconcile with individuals related to their opponents even when these individuals are not involved or in proximity to the conflict (10). This suggests that they recognize alliances and can make predictions about subsets of individuals that are likely to support one another during fights. Chimpanzee (*Pan troglodytes*) mothers monitor the play bouts of their offspring. When there is a difference in age and size between play partners, older partners use signals to appease the mother of the younger partner. This suggests older partners predict that the mother will respond negatively if the play bout is too rough (11).

These studies indicate that individuals are able to predict the outcomes of behavioral interactions in a changing social environment and tune their behavior appropriately. Making predictions about the social environment requires that individuals can extract regularities in collective dynamics. Knowledge of these regularities allows individuals to tune decision-making strategies to avoid costly fights, solicit additional support to improve their odds of winning fights when they do occur, and reduce social tension that increases the probability of conflict. The ability of individuals to encode collective dynamics requires information about individual identities, subgroups such as alliances and matrilines, and memory of past affiliative interactions and fights. The storage and recall of this information is necessarily limited because individuals in social systems have finite computational resources including limited memory and processing capacities. Hence we need to consider the memory requirements of alternative decision-making strategies and how these constraints affect the ability to predict the future (out of sample data) (12–15).

One technique for efficiently representing data using only a small number of components is sparse coding. Sparse coding seeks to explain observed input as combinations of statistically independent higher-order structures that encapsulate fine-grained correlations in the system. Olshausen et al. (16, 17) found that sparse coding of natural images produces wavelet-like oriented filters resembling the receptive fields of simple cells in the visual cortex. This work suggests that sparse coding is not only an efficient method for describing statistics of visual scenes but a good candidate algorithm for how neurons in the visual system encode this information.

We consider three alternative methods—a modified sparse coding model, frequency model, and spin-glass model—for efficiently representing regularities in a highly resolved conflict event time series collected from a group of 84 pigtailed macaques (*Macaca nemestrina*) housed at the Yerkes National Primate Research Center in Lawrenceville, Georgia (*SI Text*).

Our goal is to determine which model is the best candidate for describing how the monkeys in our study group might compress the time series. We judge the performance of each method by its ability to predict out-of-sample data (both fine-grained and collective features including individual frequency of participation in fights and the distribution of fight sizes), its cognitive demands, and its ability to identify socially meaningful individuals and subgroups.

We also consider what we call collective efficient coding, asking how much consensus (*SI Text*) there is across the local minima produced by sparse coding, our best performing method. Our interest is how much agreement there could be across individuals in their representations of the data if they were using sparse coding to compress the time series.

The dataset we analyze consists of a series of fights occurring during approximately 150 h of observation over a four-month period; in these analyses we use data from 47 of the group's 84 individuals (*SI Text*, Figs. S1 and S2). A fight includes any interaction in which one individual threatens or aggresses a second individual. A fight can involve multiple individuals. Third parties can become involved in pair-wise conflict through an aggressive, affiliative, or passive intervention, through redirection, or when a family member of a conflict participant attacks a fourth party. Individuals may intervene to break up conflicts (7) and intervene in support of one of the conflict participants. Such interventions can reflect temporary coalitions or more long-term alliances (reviewed in ref. 3). We ask if the time series can be efficiently represented using only this "who-what" (14, 15, 18–20) information. We treat each fight as an independent sample, removing temporal biases by randomly drawing 500 of the 1,078 fights from the time series to use as "in-sample" data and use the remaining fights as "out-of-sample" data.

We first test two maximum entropy models, which estimate the probability distribution over possible fights as the distribution with maximum uncertainty that fits chosen statistical features of in-sample data (21). The simplest of the models is the frequency model. It assumes that each individual has a fixed independent probability to join any given fight, producing the maximum entropy distribution constrained to fit in-sample individual frequencies. The negative log-likelihood of a fight x in the frequency model is $L_{\text{freq}}(x) = \sum_i h_i x_i$, where x_i is a binary indicator of whether individual i was involved in the given fight (see *Materials and Methods*).

Secondly, the spin-glass model [Ising model (22)] describes the distribution with maximum entropy that exactly reproduces both individual frequencies and all pair-wise correlations in the in-sample data (23). The spin-glass model has the form $L_{\text{SG}}(x) = \sum_{i,j} x_i J_{ij} x_j$. Given that some individuals appear infrequently, estimating the frequency of pair co-occurrences will suffer from small-number fluctuations. To avoid overfitting, we regularize the spin-glass model by including only a subset of individuals in the model (either those who appear in pairs with the largest covariances or the most significant covariances—*SI Text*).

Thirdly, we propose sparse coding as an alternative model that may be able to pick out important correlations of any order while making more efficient use of memory. Sparse coding attempts to represent a given set of fights using combinations of basis vectors with the constraint of sparsity: Only a few basis vectors should be used to reconstruct any given fight. Sparse basis vectors are found by minimizing a cost function that makes a tradeoff between accurate representation of the data (measured by the reconstruction error R^2) and sparseness of the fit coefficients (measured by a sparseness term λE_s that penalizes large coefficients). In matrix form, we define the best set of basis vectors B and coefficients A for a given set of fights X as the one that minimizes

$$E = R^2 + \lambda E_s = \sum_{i,j} |X_{ij} - \tanh(BA)_{ij}|^2 + \lambda \sum_{k,j} S(A_{kj}|B|), \quad [1]$$

where i indexes individuals, j indexes fights, and k indexes basis vectors. The tradeoff between accurate representation and sparseness, and hence the amount of structure identified by the model, is controlled by a single parameter, λ (see *Materials and Methods* for more details). Fig. 1 depicts the representation of a fight in terms of a sparse basis B. As groups of individuals are correlated in their appearance in fights, sparse coding represents fights in terms of those groups in order to use fewer large coefficients to represent each fight.

Results

Testing Model Predictions. We first evaluate these three models by making a functionally important fine-grained prediction: Given a subset of participants involved in a fight, which other individuals are also likely to be involved? This test is accomplished by removing each participant in turn from each out-of-sample fight and using the model to calculate the most likely individual to appear in the fight given the other participants.

In Fig. 2*A*, we plot the proportion of correct predictions for out-of-sample data for the three models, varying the sparseness parameter λ in the case of sparse coding. The frequency model performs worst and sparse coding becomes competitive with spin-glass models for a range of λs. The fact that the spin-glass and sparse coding models perform significantly better than the frequency model demonstrates that there is indeed useful information in participation correlations (we can quantify this information for subsets of our data; *SI Text* and Fig. S3). Note that the maximum possible predictive success rate depends on how much structure there is in the data, which is not computable without knowledge of the full underlying fight probability distribution; we can only test competing methods to obtain lower bounds.
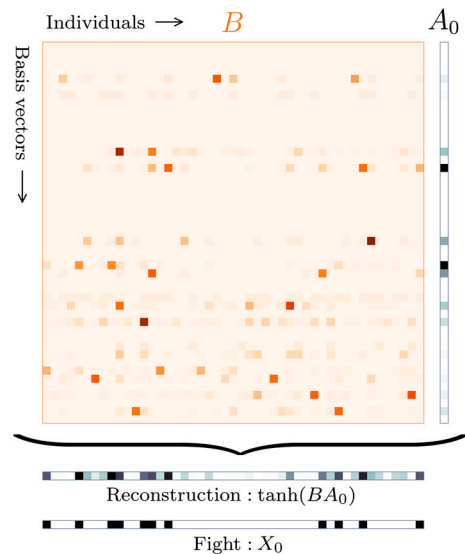

Fig. 1. Representing a fight using a sparse basis. To represent a fight X_0, coefficients A_0 define a sparse linear combination of basis vectors BA_0 that is passed through the nonlinear activation function tanh. Here λ has been tuned to produce optimal out-of-sample predictions; note that this produces sparse basis vectors that consist of small sets of individuals. Sparsity of the data implies that if one knows B and only the few active elements of A_0, one can come close to reconstructing the entire fight. Furthermore, by remembering B, the likelihood of future fights can be estimated by determining the necessary magnitudes of coefficients A_0: fights that can be represented using fewer large coefficients are more likely. In this figure and in Fig. 5, darker colors represent elements with larger magnitude (see *Materials and Methods*).

In the sparse coding model, moving from small to large sparseness λ corresponds to changing from an exact overfit representation to fitting in-sample data decreasingly well using subgroups of increasing size (see Fig. 2 *B* and *C*). If λ is too small, we find a trivial basis consisting of vectors with single individuals, such that any fight can be represented, but no interactions are encoded. If λ is too large, the model is forced to use too few basis vectors. For the remainder of the analysis, we choose $\lambda = 10^{-1/2} = 0.316$ to maximize the predictive power, implicitly fitting the complexity of the model to the data.

We next create generative versions of each model (see *Materials and Methods*) and test the ability of each to match coarse-grained statistics of the data to which they were not fit. First, we test the distribution of fight sizes, shown in Fig. 3. Sparse coding fits the fight size distribution better than the spin-glass model, with a Kullback-Leibler divergence from the empirical distribution of 0.04 bits compared to the spin-glass model's 0.21 bits. In predicting other collective statistics (e.g. the average fight size given the appearance of a particular individual), the spin-glass and sparse models perform about equally well (*SI Text*, Tables S1 and S2 and Fig. S4). These results echo recent work asserting that pair-wise interactions often contain most of the available multi-information in natural datasets (22, 24), which can be explicitly verified for subsets of our data (*SI Text*).

Estimating Cognitive Burden. Though the models are predictive and hence useful for finding regularities in data, each has an associated computational requirement. Quantifying this computational requirement is a first step toward determining if the monkeys have the necessary cognitive abilities to use these different methods to find regularities in conflict. We focus here on the amount of memory required, which we measure as the number of bits needed to store each model's parameterization of its estimate of the fight probability distribution after it has been fit to 500 in-sample fights. We test each model's performance as we
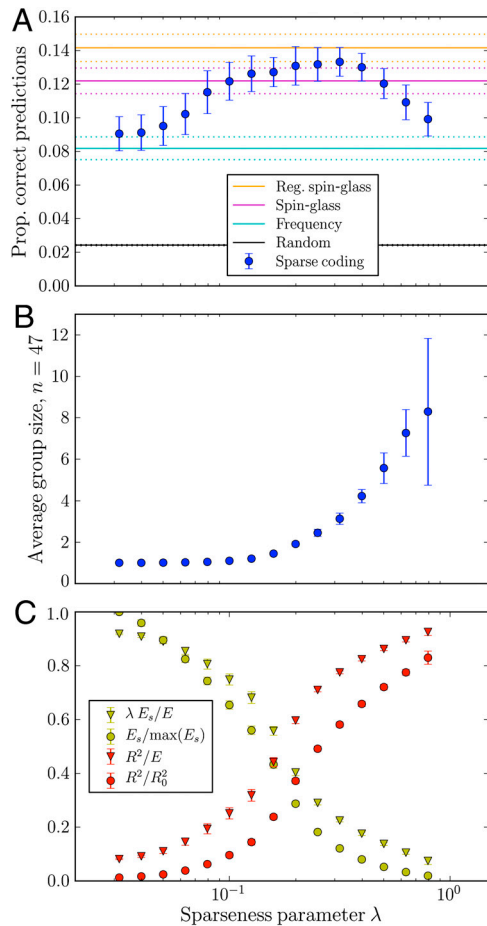
Fig. 2. (*A*) Predictive power versus sparseness. The best predictions of out-of-sample participant identities happens at an intermediate value of λ. Sparse coding performs significantly better than the frequency model and as well as pair-wise methods. The regularized spin-glass model is limited to include only 26 individuals (*SI Text*). (*B*) Average group size (the average number of large elements in each basis vector containing at least one large element, considering the $n = 47$ largest elements) versus sparseness. For increasing sparseness, individuals are collected into fewer groups with larger group size. (*C*) The two terms in E as a function of λ. As λ increases, we first have overfit representations (small R^2) that are not sparse (large E_s), and then vice versa. Over the range of λ that resolves the peak in predictive power, the sizes of R^2 and E_s relative to their sum (triangles) and to their maximum values (circles) vary appreciably. Error bars and dotted lines indicate +/- one standard deviation over 100 instances: 10 random selections of in-sample/out-of-sample data times 10 random initial conditions for sparse minimization.

decrease the amount of memory required to use it by degrading each model in one of three ways (see Fig. 4 and its caption).

We find that the performance of sparse coding is remarkably robust to "forgetting" small basis elements. Retaining only the 30 largest elements does not significantly degrade performance (Fig. S5). Similar efficient performance is also possible in the spin-glass model but requires regularizing the model, refitting using only data from a subset of individuals. In Fig. 4*B*, we find that the optimal performance of sparse coding and regularized spin-glass methods are remarkably similar under equal memory constraints. This suggests that both models are making use of the same underlying structure in the data, which we verify below. Although it is impossible to prove any model is optimal, this result also suggests that, given knowledge of 500 in-sample fights, the best performance in reconstructing incomplete out-of-sample fights requires an individual to remember on the order of 1,000 bits of information.
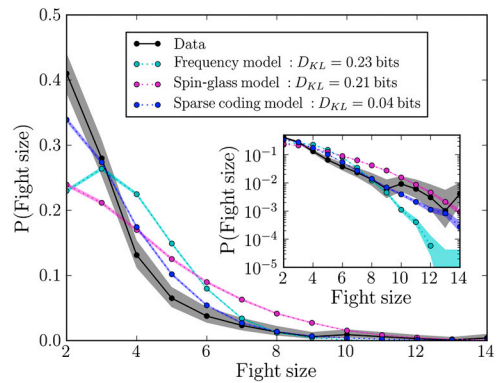
Fig. 3. Fight size distribution for the three generative models. The sparse coding model performs best with the smallest Kullback-Leibler divergence D_{KL} from the true distribution. Here we use 10^5 fight samples from each model where fights with one or zero participants have been omitted. Shaded areas connect error bars indicating bootstrapped 95% confidence intervals.

Exploring the Space of Sparse Solutions. A detail often overlooked in sparse coding studies is that the sparse minimization problem [Eq. **1**] is nonconvex, with many local minima. It is unlikely that an organism using sparse coding will be able to find the global minimum. Still, our local minimization algorithm finds metastable minima that have similar E and good performance. To explore variability in the set of metastable solutions, we performed the minimization 500 times starting from different random initial conditions (*SI Text*). Fig. 5 displays typical sparse bases, which consist mostly of correlated subgroups of size 1 to 3. Because sparsity favors basis vectors that make independent contributions, this suggests that the predictable independent units of conflict are typically individuals, pairs, and triplets.

Though there are many local minima, they share structure, which can be seen in the high frequency of many correlated groups (Fig. 5*B*, see also Figs. S6 and S7). On average, of the groups that make up two different sparse bases, 43% appear in both.

Although the sparse coding algorithm has the freedom to represent correlations among groups of individuals that cannot be represented using pair-wise interactions (see *Materials and Methods*), the sparse structures in our dataset are mostly combinations of the high-covariance pairs used in the regularized spin-glass model (Fig. 5*C*). Specifically, 94% of the nonsingleton groups appearing in the 500 sparse bases are representable using pairs with covariance larger than one standard deviation above the mean.

The sparse groups found in our analyses indicate that predictable conflict participants in this system include co-occurring groups of related individuals (genetically related through the mother), policers (7, 25) (*SI Text*) co-occurring with many types of other individuals, the alpha female as a singleton and the alpha female co-occurring with groups of related individuals. On average, these combinations make up $72 \pm 8\%$ of each basis' groups, a significantly larger fraction than for bases with the same group sizes in which individuals have been randomly substituted according to their frequency in fights ($30 \pm 13\%$; Fig. S8).

Discussion

Sparse coding provides a cognitively parsimonious model for compressing strategic correlations found in a conflict time series, effectively detecting only the most important correlations and thereby providing good performance under memory constraints. Sparse groups can capture regularities as well as a spin-glass model that fits all individual frequencies and important pair-wise correlations. Sparse bases provide an intuitive way to express regularities using less memory. Varying the amount of compression with a single parameter (in our case, the sparseness penalty) is a practical means of model selection, exploring the tradeoff between over- and under-fitting data.
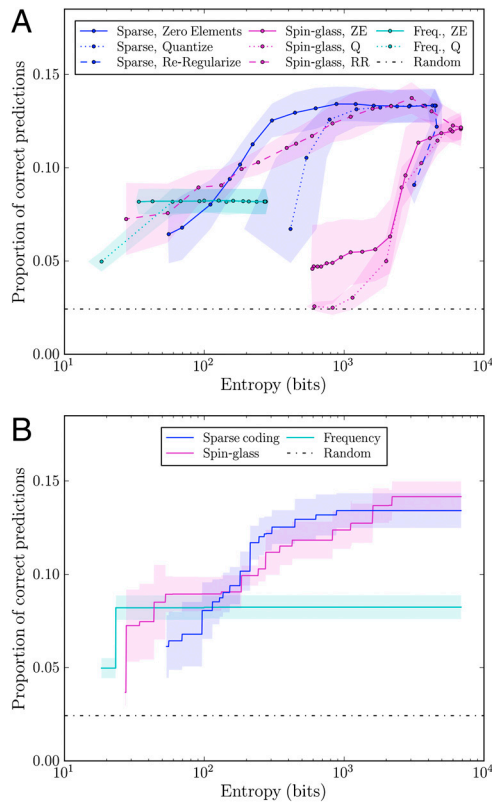
Fig. 4. Predictive power versus entropy of memory storage. (*A*) Smaller entropy is achieved in each model in three different ways: (dotted lines) by quantizing each model's parameters (the elements of h_i, J_{ij} ($j \geq i$), and B_{ij}, for frequency, spin-glass, and sparse coding, respectively), mapping them to a restricted set of allowed values to simulate diminished precision; (dashed lines) by regularizing and refitting the model (in the case of spin-glass, reducing the size of the model by limiting it to fit only individuals who appear in high-covariance pairs, and in the case of sparse coding, increasing λ to reduce the number of basis vectors); (solid lines) by zeroing the smallest stored elements. Sparse coding is quite robust to zeroing elements, performing as well as spin-glass when it is regularized and refitted. Shaded regions connect error bars that represent +/- one standard deviation in both entropy and performance. (*B*) The Pareto front for each model, defined as the maximum average performance achievable using less than a given amount of memory (including trials not depicted in *A*; *SI Text*). Shaded regions represent +/- one standard deviation in performance for the best average performer.

Performance for the compression schemes we test saturates at about 1,000 bits of memory, a small fraction of the total information in the 500 fights to which they have access ($500 \times 47 = 23{,}500$ bits). The saturation point roughly corresponds to the quantity others have called predictive information (26), the amount of information contained in past data that is useful for predicting the future. We stress that this quantity is defined with respect to our dataset and the number of fights observed during the study period. Important questions for future work include how this quantity scales with the number of observed fights, the size of the study group, fight frequency, and fight composition uniqueness.*

The results reported in this paper raise several interesting cognitive issues. The time series is serving as a "social scene" roughly analogous to the visual scenes used in neural sparse coding studies. Hence we are treating the conflict dataset as though it is static and instantaneously available in its complete form to our study subjects. If we are interested in extracting its regularities purely from the perspective of an experimenter, this makes sense. However, we are also interested in whether the monkeys in our

---

*In particular, saturation with respect to the number of observed fights is connected with the timescale over which conflict strategies evolve.

system might use sparse coding to identify predictable participants in conflict. An important issue is then the character of the sample they are using to perform the compression. In particular, because the sample is accumulating over time, the monkeys will have to repeatedly update the output of the optimization. How might this occur?

Assuming the monkeys are compressing the time series each individual will be performing the computation independently. This means it is likely that that each monkey will find a different basis (see Fig. S6). An important question is how much consensus there is across these bases. If the monkeys are using the bases to make predictions about who will be involved in a fight, the more consensus the more orderly and useful the predictions are likely to be (27). Each monkey may have access to a different sample of fights and to samples of different sizes. This additional heterogeneity is likely to reduce consensus.

A critical question for students of social evolution is how individuals use knowledge of collective dynamics to make strategic decisions. Although we do not yet know whether the animals are using any of the models we have explored in this paper, our results clearly illustrate that useful regularities exist in the time series and, furthermore, different algorithms that can be employed to extract these regularities make different cognitive demands on decoders. Now needed are studies explicitly addressing individual perception of collective dynamics. Important issues include how many bits of memory for "who and what" primates (14, 20, 28) and other organisms can store, how the capacity to encode collective dynamics relates to primate general intelligence (29), and the implications of this capacity for behavioral and social complexity (27, 30).

Materials and Methods

Data Collection. The data were collected over approximately 150 h during a four-month period from a large group of captive pigtailed macaques (*Macaca nemestrina*) socially housed at the Yerkes National Primate Research Center in Lawrenceville, Georgia. Our analyses focus on 47 socially mature individuals of 84 individuals in total. The analyses take as input a conflict event time series collected by J. C. Flack using an all-occurrence sampling procedure. A conflict or fight in this study includes any interaction in which one individual threatens or aggresses a second individual. A conflict was considered terminated if no aggression or withdrawal response (fleeing, crouching, screaming, running away, submission signals) was exhibited by any of the conflict participants for 2 min from the last such event. A fight can involve multiple individuals. Third parties can become involved in pair-wise conflict through intervention or redirection or when a family member of a conflict participant attacks a fourth party. Fights in this dataset ranged in size from 1 to 30 individuals, counting only the socially mature animals (excluding one animal who died during data collection; see *SI Text*). Full details on the data collection protocol, operational definitions, and study system details are provided in *SI Text*.

The data collection protocol was approved by the Emory University Institutional Animal Care and Use Committee and all data were collected in accordance with its guidelines for the ethical treatment of nonhuman study subjects.

The dataset includes 1,078 fights with one or more mature individuals. In-sample data consists of a random sampling of 500 fights, leaving the remaining 578 fights as out-of-sample data. Although other analyses indicate that the conflict patterns in the data set are roughly stationary, this sampling procedure buffers against temporal biases that might confound prediction of the out-of-sample data if we divided the data temporally. Performance values in Figs. 2 and 4 and Figs. S3*B*, S5 and S9 represent averages over 10 different choices of in-sample data; in the case of sparse coding, we also use 10 realizations of random starting bases for each set of in-sample data.

Models and Computational Methods. The frequency model is analytically fit to the individual in-sample frequencies f_i by setting

$$h_i = -\ln(f_i/(1 - f_i)). \qquad [2]$$

There is no closed-form solution for J_{ij} of the spin-glass model in terms of the observed frequencies and covariances; we must instead numerically vary J_{ij} to produce the observed frequencies and covariances. Instead of
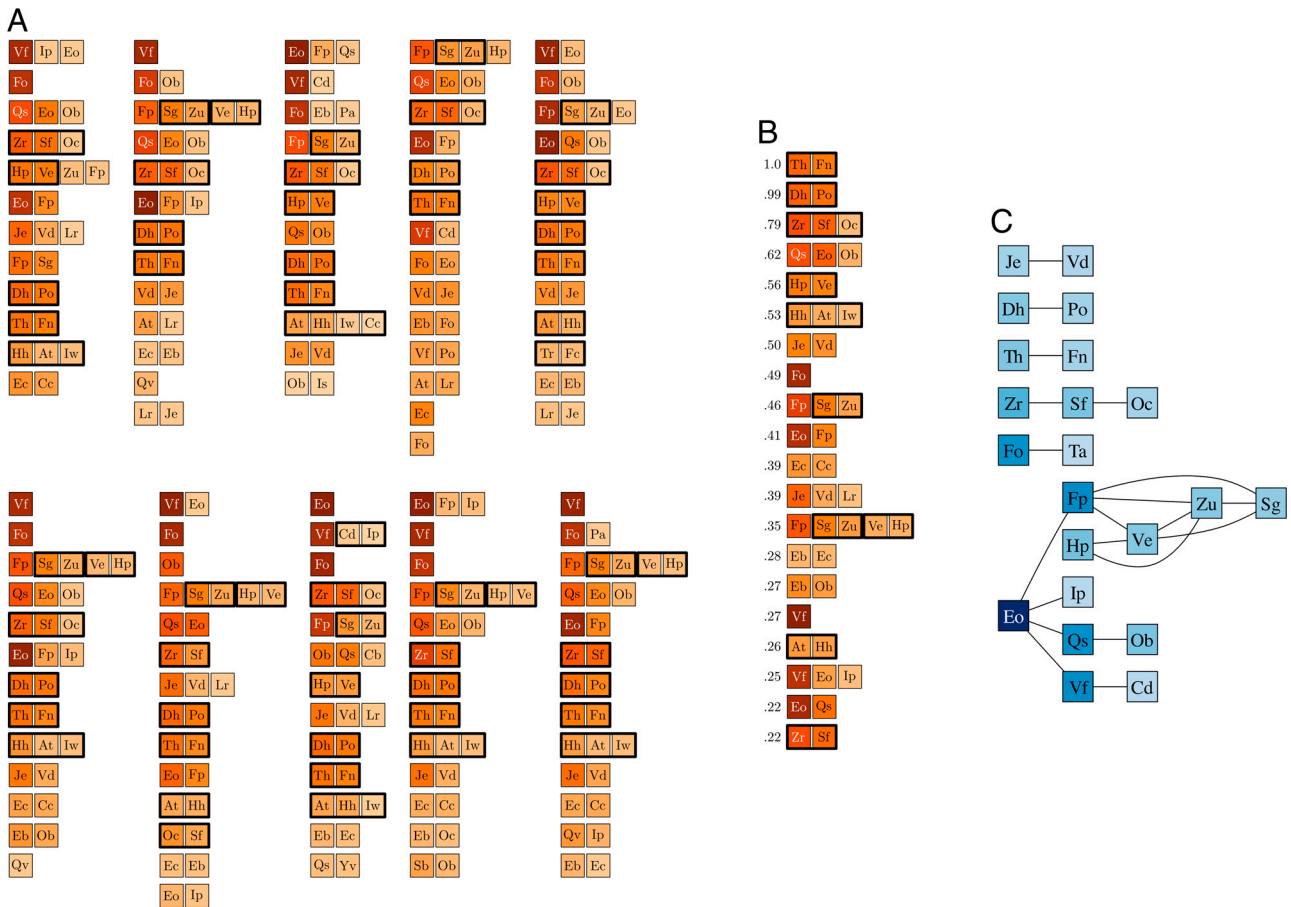
Fig. 5. (*A*) Ten example sparse bases, each a local minimum of the sparseness problem [Eq. **1**]. We represent each basis by its 30 largest elements (the approximate number at which the predictive power saturates; see Fig. S5*A*). Darker colors indicate larger magnitude, and each individual is represented by a two-letter name abbreviation. Sparse groups effectively identify sets of individuals that are likely to be involved in the same fight. Most sparse groups can be explained as combinations of genetically related individuals (indicated by thick black rectangles), policers (Eo, Qs, Fo) co-occurring with many types of other individuals, and the alpha female (Fp) occurring as a singleton or with groups of related individuals (see text). (*B*) The top 20 groups ranked by frequency in sparse bases, where a group is defined as the unordered set of individuals appearing in a basis vector. (*C*) A network representing the twenty pairs of individuals with largest absolute covariance. Many sparse groups consist of connected components of this small network. In the covariance network, darker colors indicate higher frequency in fights.

performing brute-force Monte Carlo sampling and gradient descent, we make use of a faster method known as minimum probability flow, which is known to efficiently solve spin-glass problems of our type (31).

The sparse coding model is defined by Eq. **1**, where X is the $\ell \times m$ matrix of m fights, each having dimension equal to the number of individuals ℓ; B is the $\ell \times N$ matrix of N basis vectors; and A is the $N \times m$ matrix of reconstruction coefficients. We use a "square" basis, with $N = \ell$; this is the smallest basis that is guaranteed to be able to perfectly fit any data when $\lambda = 0$. In the goodness of fit (R^2) term, we make one change to the usual sparse coding setup, using the saturating function tanh to map large values of the real-valued linear combination BA to 1. This allows for the combination of basis vectors with some overlapping members, effectively making the combination of basis vectors an OR function.

The sparseness function $S(a)$ is chosen to be an even function that grows monotonically (typically sublinearly) as $|a|$ increases; we use $S(a) = \ln(1 + a^2)$.

The multiplication of A_{kj} in the sparseness term by $|B| \equiv \sqrt{\sum_{i,k} B_{i,k}^2}$ effectively removes the degree of freedom corresponding to the overall size of basis elements, disallowing trivially small sparseness coming from large basis elements. This form also effectively imposes sparseness on the basis vectors, because the sparseness penalty term can be independently decreased by zeroing unused basis vectors.

A potential advantage of sparse coding beyond its ability to encapsulate fine-grained correlations is its ability to represent higher-order interactions that are not representable using pair-wise models. For instance, while the spin-glass model can represent a correlated triplet (one that appears more or less often than predicted by the independent frequency model), it must do so using nonzero interactions among pairs within the triplet. Sparse coding can represent such a triplet interaction without interactions among the pairs. Though sparse coding has this ability, we do not find strong evidence for such irreducible higher-order correlations in our dataset (see *Results*).

For minimization in the sparse coding model, we use the Newton conjugate gradient method (`scipy.optimize.fmin_ncg`), making use of analytic formulae for the Jacobian and Hessian of E to expedite the search. For the initial basis, we use a matrix with elements chosen from a uniform distribution between −1 and 1, which is then transformed to an orthonormal basis using the Gram-Schmidt procedure. Because we use a square basis, we can then define the initial coefficients as the unique set that perfectly reconstructs the data. We stop minimization when the average relative error falls below 10^{-6}.

Finally, to evaluate the likelihood of a fight given a remembered fit basis B, we treat E as an energy, such that the negative log-likelihood function becomes

$$L_{\text{sparse}}(x) = \min_A [E(A, B, x)/T], \quad [\mathbf{3}]$$

where the scaling factor T is a "temperature" that can be set to 1 for evaluating relative likelihoods but must be fit to data to create a generative model.

While we do not constrain the sign of basis elements, the significant contributors within any given basis vector nearly always have the same sign[†]. Every basis vector represented in Figs. 1 and 5 has this property, and this allows us to represent each basis vector by the magnitudes of its elements

[†]Analyzing the 30 largest elements in each basis, there was a single exception in the 500 bases we tested. This basis vector consisted of two individuals (Eo and Qs) with opposite sign and roughly equal magnitude.

without loss of generality. The fact that large contributors of opposite sign are very rare agrees qualitatively with previous work that found many more excitatory interactions (in which the appearance of an individual in a fight makes another individual more likely to join) than inhibitory interactions (in which it makes the other individual less likely to join) (2). Similarly, none of the largest covariances shown in Fig. 5C is negative. Bases are normalized such that $\sum B_{ij}^2 = 1$.

Generative Models. We test generative models that sample from the probability densities $P(x) = \exp(-L(x))$ defined by the frequency, (full) spin-glass, and sparse coding models. We do not use regularized spin-glass models to test generative model statistics because they omit some individuals, confounding the analysis of coarse grained statistics that include all individuals. To generate fights, we use metropolis Monte Carlo sampling. Note that, to calculate L for the sparse coding model, we need a value for the "temperature" parameter T, which effectively sets the probability that an individual will appear at random instead of as part of a sparse basis vector (or that an individual who is part of a basis vector will randomly fail to appear). We set T by matching the average E of generated fights to that found in fitting in-sample data

(using a Brent root finding algorithm). For the basis chosen to generate sparse coding data, we find $T = 0.3018$. For testing each coarse-grained statistic, we use 10^5 samples generated from each model fit to a single random choice of in-sample data.

See *SI Text* for full details on figures in the main text, spin-glass regularization, the space of metastable sparse bases, information measures on subsets of the data, model performance as a function of fight size (Fig. S9), and determining whether sparse groups are socially meaningful.

1. DeDeo S, Krakauer D, Flack J (2011) Evidence of strategic periodicities in collective conflict dynamics. *J R Soc Interface* 8:1260–1273.
2. DeDeo S, Krakauer DC, Flack JC (2010) Inductive game theory and the dynamics of animal conflict. *PLoS Comput Biol* 6:e1000782.
3. Harcourt AH, de Waal FBM, eds. (1992) *Coalitions and Alliances in Humans and Other Animals* (Oxford University Press, Oxford).
4. Noë R, Hammerstein P (1995) Biological markets. *Trends Ecol Evol* 10:336–339.
5. Silk J (1999) Male bonnet macaques use information about third-party rank relationships to recruit allies. *Anim Behav* 58:45–51.
6. de Waal FB (2000) Primates-a natural heritage of conflict resolution. *Science* 289:586–590.
7. Flack JC, de Waal FBM, Krakauer DC (2005) Social structure, robustness, and policing cost in a cognitively sophisticated species. *Am Nat* 165:E126–39.
8. Flack JC, Krakauer DC, de Waal FBM (2005) Robustness mechanisms in primate societies: A perturbation study. *Proc R Soc Lond B Biol Sci* 272:1091–1099.
9. Flack JC, Krakauer DC (2006) Encoding power in communication networks. *Am Nat* 168:E87–E102.
10. Judge P (2005) Dyadic and triadic reconciliation in pigtail macaques (*Macaca nemestrina*). *Am J Primatol* 23:225–237.
11. Flack JC, Jeannotte L, de Waal FBM (2004) Play signaling and the perception of social rules by juvenile chimpanzees. *J Comp Psychol* 118:49–59.
12. Krakauer D, Flack J, DeDeo S, Farmer D (2010) Intelligent data analysis of intelligent systems. *IDA* 8–17 LNCS 6065.
13. Griffiths TL, Tenenbaum JB (2009) Theory-based causal induction. *Psychol Rev* 116:661–716.
14. Hoffman ML, Beran M, Washburn D (2009) Memory for "what", "where", and "when" information in rhesus monkeys (*Macaca mulatta*). *J Exp Psychol Anim Behav Process* 35:143–152.
15. Schacter DL, Addis DR, Buckner RL (2007) Remembering the past to imagine the future: The prospective brain. *Nat Rev Neurosci* 8:657–661.
16. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–609.
17. Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res* 37:3311–3325.
18. Tulving E (1993) What is episodic memory? *Curr Dir Psychol Sci* 2:67–70.
19. Griffiths D, Dickinson A, Clayton N (1993) Episodic memory: What can animals remember about their past? *Trends Cogn Sci* 3:74–80.
20. Schwartz B (2001) Episodic memory in primates. *Am J Primatol* 55:71–85.
21. Jaynes ET (2003) *Probability Theory* (Cambridge Univ Press, Cambridge, UK).
22. Schneidman E, Berry MJ, II, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007–1012.
23. Jaynes E (1957) Information theory and statistical mechanics. *Phys Rev* 106:620–630.
24. Eichhorn J, Sinz F, Bethge M (2009) Natural image coding in V1: How much use is orientation selectivity? *PLoS Comput Biol* 5:e1000336.
25. Flack J, Girvan M, De Waal F, Krakauer D (2006) Policing stabilizes construction of social niches in primates. *Nature* 439:426–429.
26. Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. *Neural Comput* 13:2409–2463.
27. Flack J, Krakauer D (2011) Challenges for complexity measures: A perspective from social dynamics and collective social computation. *Chaos* 3:037108.
28. Nyberg L, et al. (1996) General and specific brain regions involved in encoding and retrieval of events: What, where, and when. *Proc Natl Acad Sci USA* 93:11280–11285.
29. Reader SM, Hager Y, Laland KN (2011) The evolution of primate general and cultural intelligence. *Philos Trans R Soc Lond B Biol Sci* 366:1017–1027.
30. De Waal F, Tyack P, eds. (2011) *Animal Social Complexity* (Harvard University Press, Cambridge, MA).
31. Sohl-Dickstein J, Battaglino P, DeWeese MR (2010) Minimum probability flow learning. arXiv:09064779v3.