Supporting Information

Daniels et al. 10.1073/pnas.1203021109

SI Text

Empirical Methods. ***Ethics statement.*** 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.

Study system. The data were collected from a large group of captive pigtailed macaques (*Macaca nemestrina*) socially housed at the Yerkes National Primate Center in Lawrenceville, Georgia. Pigtailed macaques are indigenous to south East Asia and live in multi-male, multi-female societies characterized by female matrilines and male group transfer upon onset of puberty (1). Pigtailed macaques breed all year. Females develop swellings when in Œstrus. Macaque societies more generally are characterized by social learning at the individual level, social structures that arise from nonlinear processes and feedback to influence individual behavior, frequent non-kin interactions and multiplayer conflict interactions, the cost and benefits of which can be quantified at the individual and social network levels (2–7).

The study group contained 48 socially mature individuals and 84 individuals in total. Our analyses used only 47 of the 48 socially mature individuals. One individual (Ud) was excluded because she died during the study period. Socially mature males were at least 48 mo and socially mature females were at least 36 mo by study start. These thresholds correspond to approximate onset of social maturity in pigtailed macaques. The study group had a demographic structure approximating wild populations and subadult and adult males were regularly removed to mimic emigration occurring in wild populations. All individuals, except eight (four males, four females), were either natal to the group or had been in the group since formation. The group was housed in an indoor-outdoor facility, the outdoor compound of which was 125 × 65 ft.

Pigtailed macaques have frequent conflict and employ targeted intervention and repair strategies for managing conflict (6). Data on social dynamics and conflict were collected from this group over a stable, four-month period. Operational definitions are provided below.

Operational definitions. *Fight*: includes any interaction in which one individual threatens or aggresses a second individual. A fight was considered terminated if no aggression or withdrawal response (fleeing, crouching, screaming, running away, submission signals) was exhibited by any of the fight participants for *two minutes* from the last such event. A fight can involve multiple individuals. Third parties can become involved in pair-wise fights through intervention or redirection, or when a family member of a fight participant attacks a fourth-party. Fights in this data set ranged in size from 1 to 30 individuals, counting only the socially-mature animals (excluding the animal who died; see above). Fights can be represented as small networks that grow and shrink as pair-wise and triadic interactions become active or terminate, until there are no more individuals fighting under the above described two-minute criterion. In addition to aggressors, a fight can include individuals who show no aggression or submission (e.g., third-parties who simply approach the fight or show affiliative/submissive behavior upon approaching and recipients of aggression—typically threats—who show no response). Because fights involve multiple actors, two or more individuals can participate in the same fight but not interact directly.

In this study only information about fight composition (which individuals were involved) is used. We do not consider any internal aspects of the fight, such as who does what to whom. Time data were collected on fight onset and termination but are not used in these analyses.

Alpha Female: the female with the highest power score (see below) in the group. This female (Fp) did not give subordination signals to any other female (see below). She was the second most powerful individual (males and females) in the study group.

Kin: Maternal and paternal relatedness data were available for all subadult and adult animals.

Matriline: an adult female and her daughters. In our study group, all females in a matriline were biologically related through the maternal line.

Power: the degree of consensus among individuals in the group about whether an individual is capable of using force successfully. We quantified consensus by taking into account the total number of subordination signals an individual receives and multiplying this quantity by a measure of the diversity of signals received from its population of signalers (quantified by computing the Shannon entropy of the vector of signals received by individual *i*) (3). In pigtailed macaque societies, the subordination signal is the silent bared teeth display (2) emitted outside the conflict context during pass-byes and affiliative interactions. The distribution of power in our study group is heavy tailed.

Policing: A policing intervention is an impartial intervention performed by a third party into an ongoing conflict (6). Three males and one female preform the majority of effective policing interventions but only the three males (Eo, Qs, Fo) specialize on policing (8). These four individuals occupy the top four spots in the power structure and sit towards the tail of the distribution.

Data collection protocol. During observations all individuals were confined to the outdoor portion of the compound and were visible to the observer, J.F. The approximately 150 h of observations occurred for up to eight hours daily between 1,100 and 2,000 h over a 20-wk period, comprising roughly 122 d from June through October 1998, and were evenly distributed over the day. Conflict and signaling data were collected using all-occurrence sampling.

Provisioning occurred before observations and once during observations at the same time each day. The group was stable during the data collection period (defined as no reversals in status signaling interactions resulting in a change to an individual's power score; see ref. 3).

The data were collected by a trained observer (J. C. Flack). The observer spent roughly 100 h prior to data collection learning to recognize individuals and accurately code their behavior from the observation tower above the monkey compound. Accuracy was validated by a second trained observer (F.B.M. de Waal). J.C.F. further refined data collection skills and evaluated accuracy using video playback of macaque behavior. Coding accuracy was nearly 100%. We note, however, that there is no way to be sure we have the "true time series," because there is no way to get around the fact that the data were collected by a human observer—automated collection is not yet possible for this type of data. In theory, if the observers are incorporating their own sparse descriptions on the data, then the observations could already be biased toward regular and predictable individuals. Although this is highly unlikely given the observer training protocol, we feel it should nonetheless be acknowledged.

Descriptive conflict data. The data set includes 1,078 fights with one or more mature individuals. Within these, there are 888 distinct fights involving unique sets of individuals. The average

number of mature participants in a fight is 3.3, the median is 3 mature participants, and the largest fight involved 30 mature participants. We remove fights involving none of the 47, and each remaining fight involves between 1 and 30 of the selected individuals. Fig. S1 illustrates the data we analyze, displaying the first 30 fights in the time series. Fig. S2 shows the distribution of appearance frequency over individuals; the least frequent individual (Mv) appears 31 times, and the most frequent (Eo) appears 227 times.

Paper Figures: Additional Details. ***Figure 2.*** As shown in Fig. S5*B*, we tested two variants of regularization for the spin-glass model: retaining individuals who appeared in pairs with either largest covariance or most significant covariance. The covariance for any pair in our data is equal to the deviation of observed co-occurrences from that predicted by the frequency model:

$$\text{covariance}(x_i, x_j) = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle, \quad \textbf{[S1]}$$

where $\langle \cdot \rangle$ indicates average value. We measure the significance of each covariance using a bootstrapped p value, the probability that the covariance would change sign in another sample given the number of observed co-occurrences, assuming a binomial distribution. We find that large covariances tend to be more significant (as expected since large covariances come from frequently occurring pairs, implying smaller fluctuations), such that the two regularization methods are similar in performance. The best average performance (used in Fig. 2) was found using significance ordering, and the best performance using small numbers of individuals (used in Fig. 4*A*) was found using covariance ordering.

R_0^2 indicates the reconstruction error with the sparse reconstruction set to zero: $R_0^2 = \sum_{\{i,j\}} X_{\{ij\}}$.

Figure 3. We compare the fight size distributions from models fit to in-sample data to the fight size distribution of the full data set (including both in- and out-of-sample data) in order to simplify the presentation of Fig. 3. Results do not change significantly when comparing to only out-of-sample data.

The goodness-of-fit for the fight size distribution is measured using the Kullback-Leibler divergence D_{KL}, listed in the legend of Fig. 3. Because some large fight sizes were observed in the data but not observed in the models (most notably in the frequency model, which produced no fights of size 13 or larger in 10^5 samples), we compute D_{KL} including only fights of size 12 and smaller. This excludes less than 2% of observed fights.

Figure 4. To measure the entropy required to remember a given model, we first quantize the elements that specify the model (h_i for the frequency model, J_{ij} for the spin-glass model, and B_{ij} for the sparse coding model) by splitting the range of element values into N_{levels} segments and changing each element's value to the midpoint of the segment in which it falls. The entropy of the resulting quantized distribution is estimated using the NSB algorithm described below; the total entropy is the entropy of this distribution multiplied by the number of elements that must be remembered.

Our first method for reducing the amount of entropy is to vary N_{levels}; the solid lines in Fig. 4*A* vary N_{levels} from 2 to 2^9. We find that performance saturates in each model at or below $N_{\text{levels}} = 2^9$. For the two remaining methods, we fix $N_{\text{levels}} = 2^9$.

Our second method is to regularize and refit the model. For the spin-glass model, this involves limiting the model to fit only individuals who appear in high-covariance pairs, as in Fig. S5*B* (circles). For the sparse coding model, changing the regularization corresponds to changing λ; we therefore test two other smaller λ values ($\lambda = 10^{-1}$ and $\lambda = 10^{-1.5}$), which decreases the average size of groups in the model.

Our third method is simply to change the elements smallest in magnitude to zero, "forgetting" them. This works well for the sparse coding model because many elements are already near zero. It is not surprising that this has a larger effect on the spin-glass model, because zeroing any specific J_{ij} value can change the frequencies and covariances of all individuals in a nontrivial way.

Horizontal error bars are $\pm$ one standard deviation of the estimated entropies (the uncertainties in the individual entropies due to sample size are much smaller), and vertical error bars are $\pm$ one standard deviation of the success rate.

In Fig. 4*B*, we plot the Pareto front, indicating the maximum prediction performance attainable using a given amount of entropy (or less). In the calculation of the front, we also include all other model variations we have tested. This includes the regularization of the spin-glass model using the most significant covariances (Fig. S5*B* squares). We also include variations in N_{levels} for some regularized spin-glass models and zeroed sparse models, which in some cases slightly expands the Pareto front; that is, it is sometimes slightly more efficient to both forget some elements entirely and substantially quantize the remaining elements. Specifically, we include variations of N_{levels} from 2 to 2^9 for regularized spin-glass models limited to 31, 12, and 4 individuals as well as sparse coding models limited to 20 and 10 nonzero elements.

The Space of Metastable Sparse Bases. We can look at similarities between pairs of bases by defining the group structure of the basis as the unordered set of groups present in the basis. In Fig. S6, we represent sets of metastable bases with the same group structure as nodes and link two nodes with an edge when the group structures in the two bases differ by a single group. Group structures that have large basins of attraction (and are therefore more frequently sampled) are themselves nearby to many other solutions.

We can check that the clustering seen in the basis-space network in Fig. S6 results from common structure within frequent bases and not simply from the statistics of frequent groups. We compare the original set of 500 bases to 500 random bases composed of collections of groups (basis vectors) chosen randomly from all those seen in the original 500 bases. (We do not allow a random basis to contain more than one copy of the same group, and we constrain each random basis to contain $n = 25$ nonzero elements because bases with different n would be less likely to have matching structure.) The analogous network, shown in Fig. S7, has many fewer connections and no identical group structures, showing that there is shared structure in the original bases beyond what can be explained by random collections of basis vectors.

Additional Tests for Generative Models. In this section, we compare the goodness of fit of the generative models for other coarse-grained statistics over the data. Specifically, we measure the correlation (using Pearson's coefficient and Spearman's rank coefficient) between sets of statistics observed in the data and those produced by the generative model. We analyze four statistics: the probability of observing each individual in a given fight [$P(A)$]; the probability of observing each pair of individuals in a given fight, corrected by their expected probability based on individual frequencies [$P_c(AB)$, equal to the covariance of A and B]; the average fight size for fights involving each individual [$\bar{n}(A)$]; and the average fight size for fights involving each pair [$\bar{n}(AB)$]. The results are given in Tables S1 and S2. For comparison, we include data from an inductive game theory (IGT) model (9) and correlations between in- and out-of-sample data.

Comparing rank correlations computed using out-of-sample data, the spin-glass model performs best, with significant correlations for all four statistics. Sparse coding performs slightly worse, without significant correlation for $\bar{n}(A)$.

In Fig. S4, we show the values of frequencies and covariances in each of the models plotted against their observed values, for both in- and out-of-sample observed data.

Calculations on Subsets of the Data. We can explore the effect of limited data and calculate more statistical properties of the data by splitting it into smaller parts; here, we split the 47 individuals into groups based on their frequencies of participation in fights.

The maximum predictability of time series data is related to the information contained in its correlations, which can be quantified using differences in entropies. The total possible entropy (47 bits/fight for 47 individuals) is significantly reduced first due to the small average fight size. Knowing that the average fight size is a, the remaining entropy is*

$$S_0 = a \log_2\left(\frac{\ell}{a}\right) + (\ell - a)\log_2\left(\frac{\ell}{\ell - a}\right), \qquad \textbf{[S2]}$$

where ℓ is the number of individuals. Considering all 47 individuals, $S_0 = 17.4$ bits/fight, such that knowing the average fight size carries $47 - 17.4 = 29.6$ bits/fight of information. Yet this information does not help in predicting individual withheld identities, because for each prediction we already know that some individual was involved. Knowing the frequencies with which individuals appear does help us, however, and so does knowing any higher order correlations. Thus to characterize the information in the data useful for predicting withheld identities, we measure reductions in entropy below S_0.

For instance, the information contained in the frequency distribution that is useful for predicting withheld identities is

$$I_1 = S_0 - S_1, \qquad \textbf{[S3]}$$

where S_i for $i > 0$ is the entropy of the maximum entropy distribution over fights that matches i-wise correlations. Analogously, $I_2 = S_0 - S_2$ is the useful information captured by the spin-glass model, and $I_N = S_0 - S_N$ represents all frequency and interaction information of all orders†.

We use the NSB method (10, 11) to estimate these entropies in our data. This method provides reliable estimates for the entropy when the number of samples is much more than $2^{\ell/2}$, where ℓ is the dimensionality of the data. With about 10^3 fight samples, we cannot calculate the entropy including all individuals ($2^{47/2} \approx 10^7$), but we can calculate entropies for subsets of individuals (eg. $2^{12/2} = 64$). We therefore divide the 47 individuals into four groups based on their frequency in fights (12 individuals in quartiles 1–3, and 11 individuals in quartile 4). See Fig. S3. Total entropy S_N is measured using all 1,078 fights, S_2 is measured using 10^5 samples from a spin-glass model fit to all fights, and S_1 and S_0 are calculated analytically using

$$S_1 = -\sum_i f_i \log_2 f_i + (1 - f_i)\log_2(1 - f_i) \qquad \textbf{[S4]}$$

and Eq. **S2**, where individual frequencies f_i and average fight size a are calculated over all fights.

Sparse Groups are Socially Meaningful. To determine to what extent our sparse bases consist of socially meaningful groups, we calculate the mean proportion of groups in each basis that match one of three criteria: The group contains a policer, the group is made entirely of one or more groups of related individuals, or the group is made entirely of the alpha female or the alpha female plus one or more groups of related individuals. As shown in Fig. S8, we find that these combinations make up, on average, $72 \pm 8\%$ of the groups in each sparse basis. Comparing to null sets of bases that retain the same distribution of group sizes but replace individual identities either at random ("random replacement") or at the same frequency at which individuals appear in fights ("frequency replacement"), we find for both nulls significantly fewer groups that match known kinship groups or contain the alpha female and kinship groups and significantly fewer groups that contain a policer for random replacement only. Thus frequency models can provide information about critical social roles played by individuals (e.g. policing), but sparse coding can additionally extract higher-order structures (e.g., kinship ties).

1. Caldecott J (1986) An Ecological and Behavioural Study of the Pig-Tailed Macaque. (S. Karger, Basel).
2. Flack JC, de Waal F (2007) Context modulates signal meaning in primate communication. *Proc Natl Acad Sci USA* 104:1581–1586.
3. Flack JC, Krakauer DC (2006) Encoding power in communication networks. *Am Nat* 168:E87–E102.
4. Flack J, Girvan M, De Waal F, Krakauer D (2006) Policing stabilizes construction of social niches in primates. *Nature* 439:426–429.
5. 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.
6. Flack JC, de Waal FBM, Krakauer DC (2005) Social structure, robustness, and policing cost in a cognitively sophisticated species. *Am Nat* 165:E126–E139.
7. Thierry B, Singh M, (2004) *Macaque Societies: A Model for the Study of Social Organization* (Cambridge Univ Press, Cambridge, UK).
8. Krakauer D, Page K, Flack J (2011) The immuno-dynamics of conflict intervention in social systems. *PloS One* 6:e22709.
9. DeDeo S, Krakauer DC, Flack JC (2010) Inductive game theory and the dynamics of animal conflict. *PLoS Comput Biol* 6:e1000782.
10. Nemenman I, Bialek W, de Ruyter van Steveninck R (2004) Entropy and information in neural spike trains: Progress on the sampling problem. *Phys Rev E* 69:056111.
11. Nemenman I, Shafee F, Bialek W (2002) Entropy and inference, revisited. *Adv. Neural Inf. Proc. Syst.* eds Dietterich TG, Becker S, Ghahramani Z (MIT Press, Cambridge, MA).

*The maximum entropy distribution constrained to have average fight size a has each individual appearing at random with the same probability a/ℓ.

†Note that this information measure is related to, but not equal to, the multi-information $MI = S_1 - S_N$.

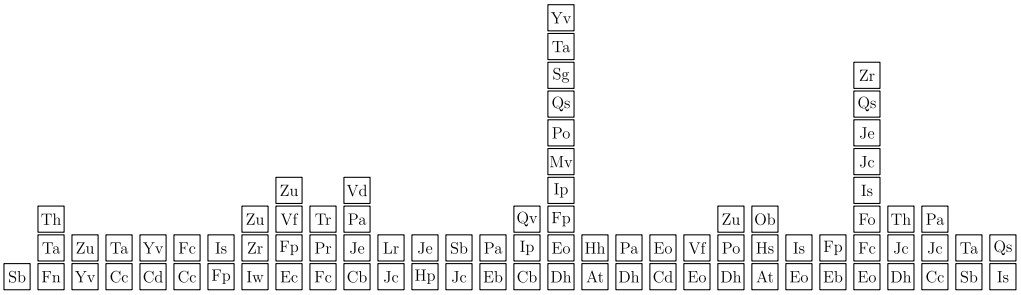

Fig. S1. The first 30 fights in the dataset, including only the 47 mature individuals (without Ud).

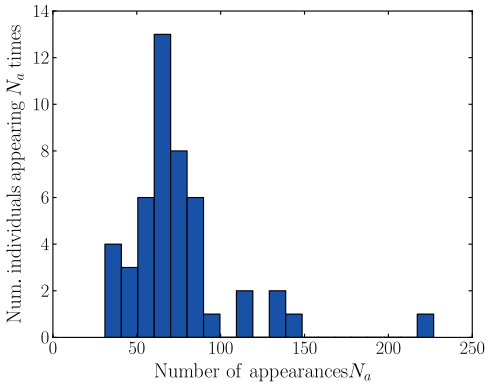

Fig. S2. Distribution of appearance frequency for the 47 mature adults in 1,078 fights.

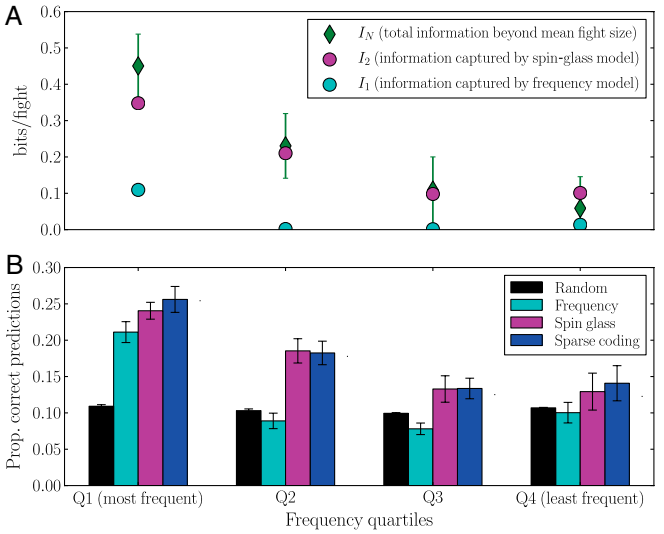

Fig. S3. Analysis of data split into four quartiles of individuals based on their frequencies of appearance in fights. (*A*) Using data only within each quartile, dimensions are small enough to accurately estimate entropies and thus the information captured by each maximum entropy model (I_1 and I_2) as well as all possible interaction information (I_N). Higher frequency quartiles carry more interaction information and in each case pair-wise interactions capture most or all possible information. Error bars are uncertainties due to finite sample size estimated using the NSB method. (*B*) Our intuition holds that more captured information corresponds to better predictions. (Note that the frequency models sometimes perform worse than random due to overfitting.)

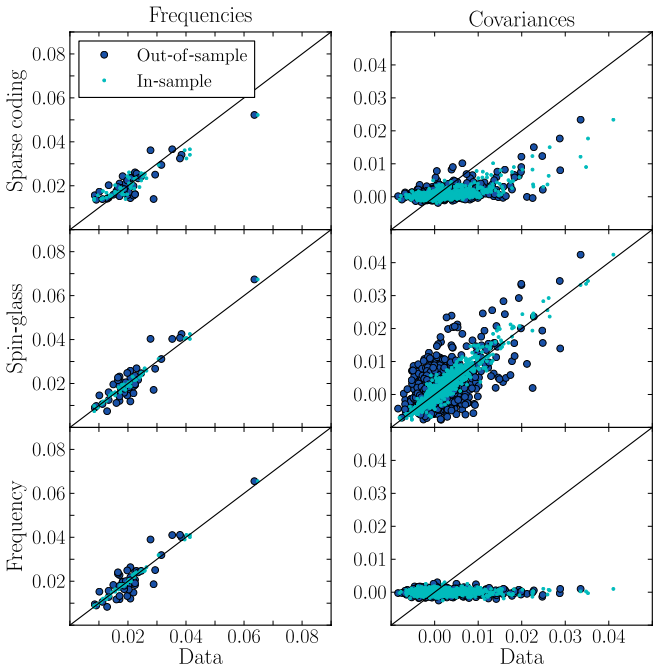

Fig. S4. Comparing frequencies and covariances in the empirical data to those found in data generated from the sparse coding model at $T = 0.3018$ (*Top*), the spin-glass model (*Middle*), and the frequency model (*Bottom*), using 10^5 fight samples. Sparse coding fits individual frequencies about as well as other methods and roughly fits only the largest covariances.

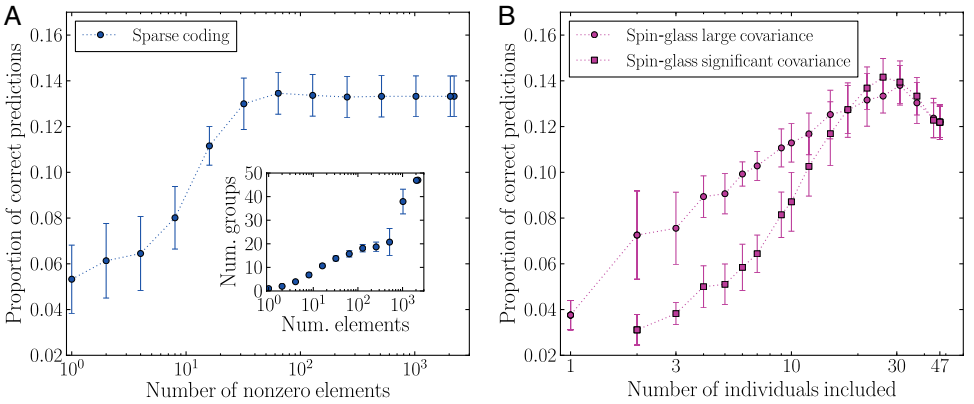

Fig. S5. (*A*) The performance of the sparse coding model (with $\lambda = 10^{-1/2}$) as the basis elements with magnitudes smaller than a varying cutoff are set to zero. The number of remaining nonzero elements is varied by changing the cutoff. Because the performance drops off significantly only below about $n = 30$, we use the largest 30 elements to represent each basis in Fig. 5. The inset shows the number of groups remaining with one or more nonzero elements. (*B*) Regularizing the spin-glass model: The performance of the spin-glass model as the size of the model is varied by including only individuals belonging to pairs with (circles) largest covariance or (squares) most significant covariance. Overfitting is evident, because predictions are best for the regularized model restricted to 26 individuals (using the most significant covariances) instead of the full model which includes all 47 individuals.

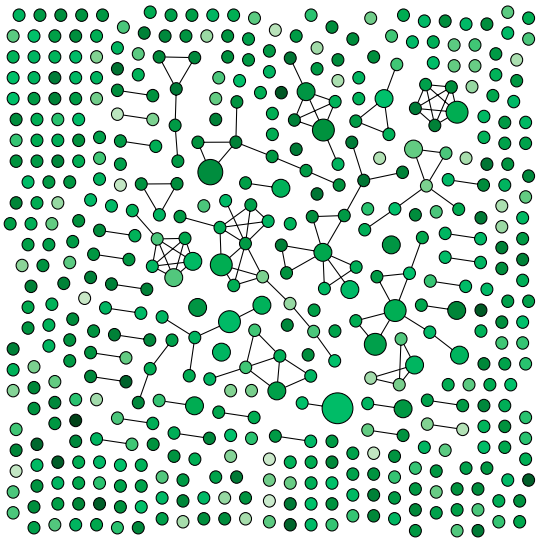

Fig. S6. A network representing the space of sparse solutions. Each node represents one group structure found in 500 randomly sampled local minima of Eq. **1** (using $n = 25$). A node's size is proportional to the number of minima that share that group structure (ranging from 1 to 6), and its color is darker if it has a larger predictive success rate (ranging from white = 10.4% to dark green = 14.5%). Structures are connected with an edge when they differ by only a single group. The best performers do not share highly similar structure, but common group structures do tend to cluster near each other.

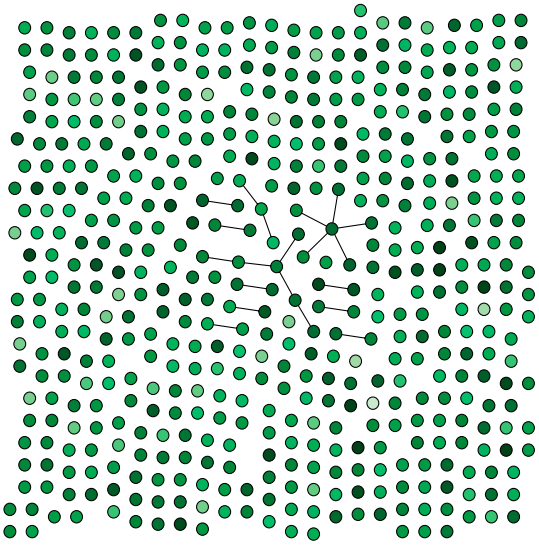

Fig. S7. Network analogous to Fig. S6 for bases consisting of random combinations of groups (see text for details). Because there is much smaller similarity between bases (smaller consensus) when bases are formed from random combinations of groups, the original sparse bases must contain structure more than random collections of groups.

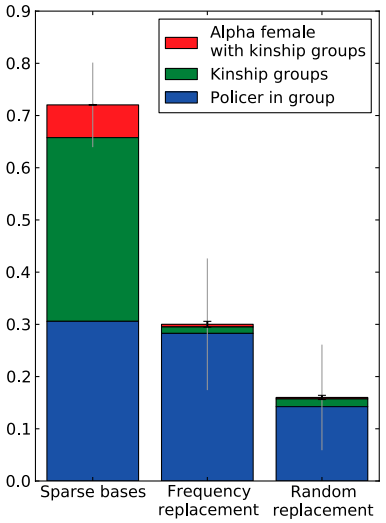

Fig. S8. Sparse bases consist of socially meaningful groups significantly more often than two null models. The total height of each bar corresponds to the mean proportion of groups in each basis that fit our definition of a socially meaningful group. Colors correspond to different types of social structures (see text). Large gray error bars indicate standard deviation of the proportion of groups per basis, and small black error bars indicate standard deviation of the mean for 500 bases over 100 null model instances.

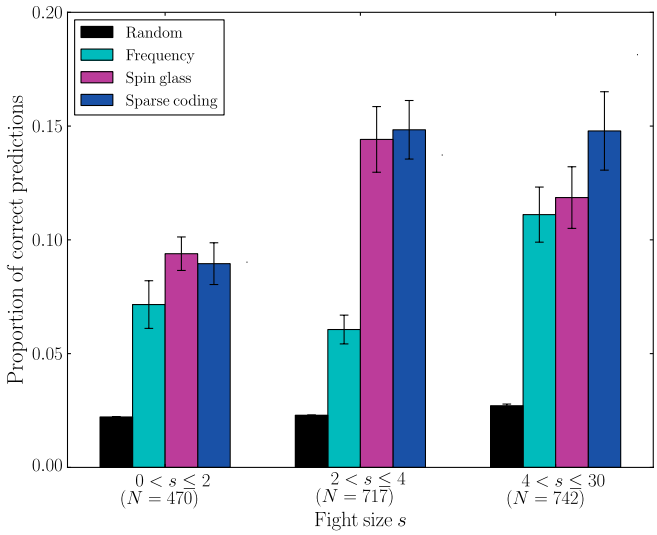

Fig. S9. For models fit using data from all 47 individuals, the proportion of correct predictions as a function of the number of participants in the fight. With more participants in a fight, it is easier for the sparse coding method to predict which other individuals are likely to be present.

Table S1. Pearson's correlation coefficients as a measure of the generative models' abilities to capture various statistics in the data

	$P(A)$	$P_c(AB)$	$\overline{n}(A)$	$\overline{n}(AB)$
In-sample				
IGT [9]	0.62*	0.51*	0.80*	0.37*
Frequency	1.00*	0.01	−0.04	0.04
Spin-glass	1.00*	0.97*	0.94*	0.50*
Sparse coding	0.95*	0.76*	0.55*	0.27*
Out-of-sample				
In-sample data	0.91*	0.60*	0.38*	0.15*
Frequency	0.91*	0.01	−0.02	0.05
Spin-glass	0.91*	0.57*	0.37*	0.07*
Sparse coding	0.84*	0.57*	0.25	0.05

$P(A)$ are individual frequencies, $P_c(AB)$ are pairwise correlations, $\overline{n}(A)$ are average fight sizes conditional on individual appearances, and $\overline{n}(AB)$ are average fight sizes conditional on pair appearances (*$p < 0.05$). Focusing on out-of-sample tests, spin-glass performs better than sparse coding on conditional average fight sizes, and the two methods perform comparably in predicting frequencies and pairwise correlations.

Table S2. Same as Table S1, using Spearman's rank correlation coefficients

	$P(A)$	$P_c(AB)$	$\overline{n}(A)$	$\overline{n}(AB)$
In-sample				
Frequency	1.00*	0.03	−0.07	0.04
Spin-glass	0.99*	0.94*	0.94*	0.57*
Sparse coding	0.90*	0.53*	0.47*	0.36*
Out-of-sample				
In-sample data	0.74*	0.35*	0.38*	0.17*
Frequency	0.74*	0.02	0.02	0.06
Spin-glass	0.74*	0.32*	0.39*	0.09*
Sparse coding	0.65*	0.27*	0.17	0.08*