Searching for structure in collective systems

Colin R. Twomey[1]*, Andrew T. Hartnett[2], Matthew M. Grobis[3], & Pawel Romanczuk[4,5]

[1] Department of Biology, University of Pennsylvania, Philadelphia, PA, USA
[2] Zipline Scientific Consulting, LLC., USA,
[3] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA,
[4] Institute for Theoretical Biology, Department of Biology, Humboldt Universität zu Berlin, Germany
[5] Bernstein Center for Computational Neuroscience, Berlin, Germany

July 4, 2018

Abstract

Collective systems such as fish schools, bird flocks, and neural networks are comprised of many mutually-influencing individuals, often without long-term leaders, well-defined hierarchies, or persistent relationships. The remarkably organized group-level behaviors readily observable in these systems contrast with the ad hoc, often difficult to observe, and complex interactions among their constituents. While these complex individual-level dynamics are ultimately the drivers of group-level coordination, they do not necessarily offer the most parsimonious description of a group's macroscopic properties. Rather, the factors underlying group organization may be better described at some intermediate, mesoscopic scale. We introduce a novel method from information-theoretic first principles to find a compressed description of a system based on the actions and mutual dependencies of its constituents, thus revealing the natural structure of the collective. We emphasize that this method is computationally tractable and requires neither pairwise nor Gaussian assumptions about individual interactions.

1 Introduction

Collective behavior is an emergent property of the actions and interactions of a system's constituents. Typically, these individual actions are readily observable, while interactions are hidden: they can only be inferred indirectly, and in many cases only with great difficulty. A growing body of research on collective systems is devoted to exactly this problem (see e.g. Ballerini et al., 2008; Lukeman et al., 2010; Nagy et al., 2010; Katz et al., 2011; Herbert-Read et al., 2011; Bialek et al., 2012; Strandburg-Peshkin et al., 2013; Rosenthal et al., 2015; Harpaz et al., 2017; Torney et al., 2018). In general, methods for inferring interactions fall into one of two categories. Many make strong, system-specific simplifying assumptions about the nature of the interactions, e.g. requiring linearity and/or pairwise interaction topologies, thus limiting their widespread applicability. Other

*Corresponding author email crtwomey@sas.upenn.edu

Author contributions: C.R.T. formulated the problem and solution, and wrote the paper; A.T.H. and P.R. assisted with the solution and writing; P.R. contributed simulation model and data; M.M.G. contributed empirical tracking data and writing. **Data and materials availability**: Empirical video data for schooling fish provided by Iain D. Couzin, from the work of Katz et al. (2011). I.D.C. contributed directly to the filming and collection of this data. **Funding**: P.R. acknowledges support from the German Science Foundation (DFG) grant RO 47766/2-1; M.M.G. was supported by an NSF Graduate Research Fellowship. Funding and support for the empirical data used in this paper was provided by I.D.C. (see Katz et al., 2011, for additional information). Early funding and support for C.R.T.'s work with rate-distortion theory was provided by I.D.C. and an NSF Graduate Research Fellowship; current funding and support is provided by Joshua B. Plotkin and a mindCORE Postdoctoral Research Fellowship.

solutions relax these constraints but at the cost of tractability as systems scale in size to more than a few individuals. Even once accomplished, the problem of inferring all dependencies between all elements of a system at every moment is only the first step in the analysis of collective behavior. The overarching goal is to then understand how the dependencies between elements determine group-level coordination.

We will call this focus on characterizing the moment-to-moment interactions of a group the ‘bottom-up’ approach to quantifying collectivity. At the other end of the scale, the ‘top-down’ approach is to simply measure one or more bulk properties of the system, such as average alignment (when such a property makes sense; e.g. for locusts or fish in Buhl et al., 2006; Tunstrøm et al., 2013, respectively). However, for nest-site selection in honeybees (Seeley & Visscher, 2004), bridge formation (Reid et al., 2015) or foraging decisions (Greene & Gordon, 2007) in ants, social conflict policing in Macaques (Flack et al., 2006), quorum sensing in bacteria (Papenfort & Bassler, 2016), or neuronal avalanches in slices of neocortex (Beggs & Plenz, 2003), average alignment would not be the most meaningful aggregate measure of collectivity. In general, the choice of what bulk property to measure, and even what bulk properties may be sensible to measure, is system dependent.

The top-down and bottom-up views of collectivity are not mutually exclusive. On the contrary, ideally they are complementary, and it may be necessary to employ either or both depending on the system and the question asked. Here, we introduce a third approach to the problem of quantifying collectivity with the aim of unifying these two views, while addressing some of their practical and fundamental issues. First, building from information-theoretic first principles, we introduce a measure of aggregate collectivity based directly on the observable actions of a system’s individual elements. This measure quantifies the relative degree of statistical dependence shared by a set of elements and in principle is valid for any system of any size. The degree of macroscopic collectivity and its variation over time can thus be productively quantified and compared across systems. Second, we show that this measure allows us to find a natural decomposition of a system into simpler components. This decomposition provides a mesoscale description of a system that may offer a simpler basis on which to make inferences about the causal system-wide interactions that underly group-level organization. Finally, in addition to a rigorous theoretical foundation, we show that this approach is readily applicable to both observed and simulated experimental data in practice.

1.1 Redundancy

Let $S = \{1, 2, \dots, n\}$ be the indices of a set of random variables, $\{X_i\}_{i\in S}$, which in general may be neither identically distributed nor independent. In the context of a fish school or a bird flock, this could be the set of all the velocity vectors of the individuals in the group; for neurons, this could be the state of each neuron (firing or silent). In general, it could be any heterogeneous assemblage of the microscopic observables of a system. If we were asked to faithfully record the current state of the whole group, one strategy would be to simply write down a description of each element separately. One of the foundational results from information theory is that no lossless description of a random variable can be shorter on average than the tight lower bound given by its entropy (Shannon, 1948). Thus a description of the system given by recording every element separately would require on average a minimum of $\sum_{i\in S} H(X_i)$ bits, where $H(X_i)$ is the entropy of X_i.

Alternatively, another strategy would be to instead write down a shared (or ‘joint’) description of all elements at once. A joint description can capitalize on the dependencies among a set of variables to reduce the overall description length needed. For example, to characterize the state of both a lamp and the light switch that controls it, one could simply record the on/off state of one of the two components. Knowing the state of either the switch or the lamp automatically tells us the state of the other, under perfect operating conditions. For less than perfect operating conditions it will be necessary to include additional information about the state of the other component, but only as frequently as the light switch fails to determine the state of the lamp. In either case, the joint entropy of the lamp and the light switch together determines the lower bound on the lossless joint description of the system. Thus the smallest lossless joint description requires $H(\{X_i\}_{i\in S})$ bits on average, where we are guaranteed that $H(\{X_i\}_{i\in S}) \leq \sum_{i\in S} H(X_i)$.

In fact, the only way in which the joint description is as costly as the sum of the individual (or 'marginal') descriptions is if all X_i's are independent. The difference between the marginal and joint descriptions, given by

$$I(\{X_i\}_{i\in S}) = \sum_{i\in S} H(X_i) - H(\{X_i\}_{i\in S}), \tag{1.1}$$

gives us a natural measure of how much we reduce the fundamental representation cost by using a joint, rather than a marginal, description. Another way to think about Eq. 1.1 is as a measure of redundancy: the amount of information that is made redundant (unnecessary) when describing $\{X_i\}_{i\in S}$ as a whole rather than by parts. A similar interpretation can be found in Watanabe (1960)'s original investigation of Eq. 1.1 as a general measure of multivariate correlation.[1]

Notably, redundancy in the absolute sense given by Eq. 1.1 scales in magnitude with the size of the system. For example, if we take n identical copies[2] of the same random variable, X, then we have $I(\{X_i\}_{i\in S}) \propto H(X)$ with the constant of proportionality equal to $n-1$. If $H(X) > 0$, then in the limit as $n \to \infty$, $I(\{X_i\}_{i\in S}) \to \infty$. To compare redundancies between systems or subsystems of different sizes, it can be useful to instead consider the relative redundancy, i.e.

$$r = \frac{I(\{X_i\}_{i\in S})}{\sum_{i\in S} H(X_i)} = 1 - \frac{H(\{X_i\}_{i\in S})}{\sum_{i\in S} H(X_i)} = 1 - s, \tag{1.2}$$

where s is then the proportion of non-redundant, or incompressible, information in the set. Taking the same example as before, for n identical copies of X, as $n \to \infty$, $r \to 1$, and $s \to 0$ (see Fig. 1). At the other extreme, if instead of n identical copies we have n mutually independent X_i's, then as $n \to \infty$, $r = 0$ and $s = 1$. In general, $0 \leq r < 1$ for any finite set of random variables with at least one variable having non-zero entropy, and, correspondingly, $0 < s \leq 1$.

2 Method

While relative redundancy, or equivalently incompressibility, can be used to compare the degree of collectivity exhibited by very different systems, it can also be used to characterize the dependency structure within a given system. Writing the relative redundancy as a function of a subset of the system, $A \subseteq S$, we have

$$r(A) = 1 - \frac{H(\{X_i\}_{i\in A})}{\sum_{i\in A} H(X_i)}. \tag{2.1}$$

[1] As noted by Watanabe (1960), its significance as a potential measure of organization stretches back still further, to at least Rothstein (1952).

[2] Meaning that they share the same outcome.

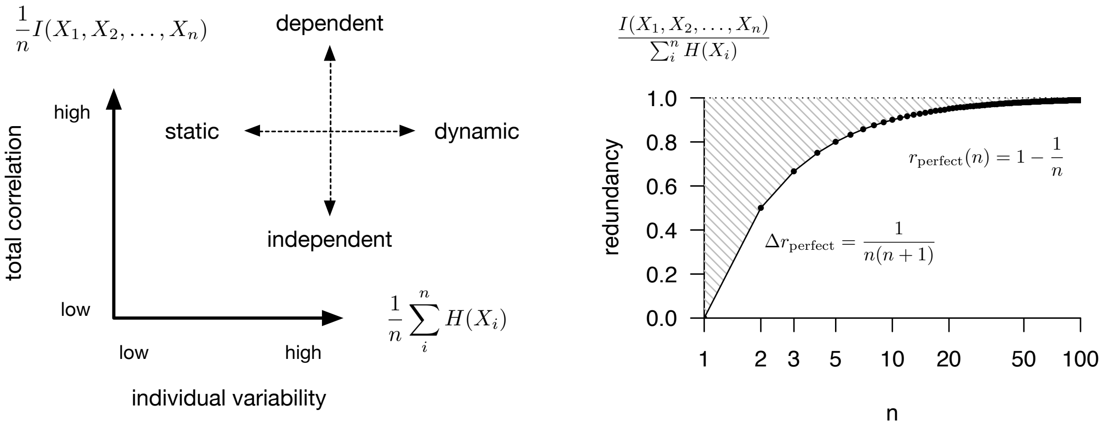


Figure 1. (*Left*) Schematic description of a system, $\{X_i\}_{i \in S}$, by its average total correlation (y-axis), measuring dependence, and the average marginal entropy of its elements (x-axis). (*Right*) Feasible (white) and infeasible (shaded) redundancies for systems of a given size, n. The upper bound is given by a system in which every element is perfectly dependent on every other element (so knowing the state of one element is as good as knowing the state of every element in the system). The lower bound is zero, which occurs when every element is independent of every other element.

What divisions of a system maximize the relative redundancy of each subset? Can the subdivisions of a system achieve a higher average relative redundancy than the system as a whole?

To begin making these questions concrete, let $\widehat{S}$ be a set of indices for a collection of subsets of S, which we will refer to as the *components* of system S. That is, let $\widehat{S} = \{1, 2, \ldots, m\}$, where typically[3] $m \leq n$, and introduce a probabilistic assignment $p(j|i)$, $\forall (i, j) \in (S, \widehat{S})$.[4] Then the expected quality of an assignment to a given component is

$$\mathbb{E}\left[r(A)|j\right] = \sum_{A \in \mathcal{P}(S)} r(A)p(A|j), \tag{2.2}$$

where $\mathcal{P}(S)$ is the power set (set of all subsets) of S, and

$$p(A|j) = \prod_{i \in A} p(j|i) \prod_{i \in A^{\mathrm{c}}} \left[1 - p(j|i)\right], \tag{2.3}$$

is the probability of subset A given the assignments of elements to component j, by a simple counting argument.[5] Treating the quality of each component equally, the expected quality over all components is then

$$\mathbb{E}\left[r(A), j\right] = \frac{1}{m} \sum_{j \in \widehat{S}} \mathbb{E}\left[r(A)|j\right]. \tag{2.4}$$

[3] If $m > n$ then some components will necessarily be empty.

[4] The use of i and j as elements of S and $\widehat{S}$, respectively, will follow this convention in the rest of the paper.

[5] Unless stated otherwise, the complement of a set is taken with respect to S, i.e. $A^{\mathrm{c}} = \{k \in S : k \notin A\}$.

2.1 Rate-distortion theory

While this gives us a natural way to evaluate the quality of a given assignment, it does not immediately provide us with a way to find such an assignment. Instead, we draw inspiration from the information-theoretic treatment of compression given by rate-distortion theory (see Shannon, 1959; Cover & Thomas, 2006). Classical rate-distortion theory addresses the following problem: given a source (random variable) X, a measure of distortion, d, and an allowable level of average distortion D, determine the minimum rate necessary for a compressed description of X that introduces an average distortion no more than D. I.e.,

$$R(D) = \min_{p(\hat{x}|x)\,:\,\mathbb{E}d(x,\hat{x})\,\leq\, D} I(X;\widehat{X}). \tag{2.5}$$

In this case, the rate measures the information, $I(X;\widehat{X})$, that the compressed representation, $\widehat{X}$, needs to keep about the source, X, where

$$I(X;\widehat{X}) = \sum_{x,\hat{x}} p(x,\hat{x}) \log \frac{p(x,\hat{x})}{p(x)p(\hat{x})} \tag{2.6}$$

is the mutual information between X and $\widehat{X}$. The lower the rate, the better the compression, but (depending on the source and the distortion measure) the higher the average distortion introduced. Surprisingly, not only can the rate-distortion curve be characterized numerically in general, the minimal compressed representation of X can be found via a simple, iterative, alternating minimization algorithm (Blahut, 1972; Arimoto, 1972).

2.2 Redundancy compression

Though there are important differences from rate-distortion theory (discussed in Appendix A), we can similarly frame the problem of finding structure based on redundancy as a compression problem. Here, we wish to find the assignment of elements of S to components of $\widehat{S}$ that achieves an average redundancy no less than r^*, and otherwise preserves as little about the original identities of the elements as possible. I.e.,

$$R(r^*) = \min_{p(j|i)\,:\,\mathbb{E}[r(A),j]\,\geq\, r^*} I(S;\widehat{S}), \tag{2.7}$$

where $p(j|i)$ is further required to be non-negative and sum to one. This is not a standard rate-distortion problem,[6] but we can use much of the same ideas developed by Blahut (1972) and Arimoto (1972) in their original numerical algorithms for the channel capacity and rate-distortion problems for deriving a practical solution. We give a brief account of this derivation here; see Appendix A for a complete account.

Introducing Lagrange multipliers to constrain the $\sum_{i\in S} p(j|i) = 1$ (non-negativity will be enforced by the form of the solution), the variational problem becomes

$$L\left[p(j|i)\right] = I(S;\widehat{S}) - \beta \sum_{j\in\widehat{S},A\in\mathcal{P}(S)} r(A)p(A|j) + \sum_{j\in\widehat{S}} \lambda(j) \sum_{i\in S} p(j|i), \tag{2.8}$$

[6] A consequence of the differences with the standard rate-distortion formulation is that we should not expect $R(r^*)$ to necessarily behave similarly to $R(D)$ as we vary r^* and D, respectively.

where β, the lagrange multiplier for the average redundancy constraint, absorbs the $1/m$ term. Taking the derivative with respect to a particular j' and i', we have

$$\frac{\partial}{\partial p(j'|i')} L\left[p(j|i)\right] = p(i') \log \frac{p(j'|i')}{p(j')} - \beta \sum_{j \in \widehat{S}, A \in \mathcal{P}(S)} r(A) \frac{\partial p(A|j)}{\partial p(j'|i')} + \lambda(i'), \tag{2.9}$$

where

$$\frac{\partial p(A|j)}{\partial p(j'|i')} = \begin{cases} 0 & \text{if } j \neq j', \\ f_{i'}(A|j') & \text{if } j = j', i' \in A, \\ -f_{i'}(A|j') & \text{if } j = j', i' \in A^{\mathsf{c}}, \end{cases} \tag{2.10}$$

and

$$f_i(A|j) = \prod_{k \in A \setminus \{i\}} p(j|k) \prod_{k \in A^{\mathsf{c}} \setminus \{i\}} \left[1 - p(j|i)\right], \tag{2.11}$$

where $A \setminus \{i\}$ is the relative complement of $\{i\}$ with respect to A.

Then setting $\partial L / \partial p(j'|i') = 0$ and splitting the sum over $\mathcal{P}(S)$ into terms with and without $i' \in A$, we have

$$\begin{aligned} p(i') \log \frac{p(j'|i')}{p(j')} = \beta & \sum_{\{A \in \mathcal{P}(S)\,:\, i' \in A\}} r(A) f_{i'}(A|j') \\ - \beta & \sum_{\{A \in \mathcal{P}(S)\,:\, i' \in A^{\mathsf{c}}\}} r(A) f_{i'}(A|j') \\ - \lambda&(i'). \end{aligned} \tag{2.12}$$

Let

$$d(i,j) = \frac{1}{p(i)} \sum_{\{A \in \mathcal{P}(S)\,:\, i \in A\}} r(A) f_i(A|j), \tag{2.13}$$

and define $d_{\mathsf{c}}(i,j)$ to be identical except substituting $i \in A^{\mathsf{c}}$ for $i \in A$. Lastly, let $\Delta d(i,j) = d(i,j) - d_{\mathsf{c}}(i,j)$. Then, dividing through by $p(i')$ and substituting, we have,

$$\log \frac{p(j'|i')}{p(j')} = \beta \Delta d(i',j') - \frac{\lambda(i')}{p(i')}. \tag{2.14}$$

Finally, substituting $\log \mu(i') = \lambda(i')/p(i')$ and solving for $p(j'|i')$,

$$p(j'|i') = \frac{p(j')}{\mu(i')} e^{\beta \Delta d(i',j')}. \tag{2.15}$$

Enforcing the constraint that $\sum_{j \in \widehat{S}} p(j|i') = 1$ and simplifying notation, we have

$$p(j|i) = \frac{p(j) e^{\beta \Delta d(i,j)}}{\sum_{j' \in \widehat{S}} p(j') e^{\beta \Delta d(i,j')}}. \tag{2.16}$$

Before moving on, it is worth noting that $\Delta d(i,j)$ has a simple and intuitive interpretation. It is the difference in redundancy for component j when i is included versus when it is excluded, weighted by the relative importance of i.

Note that $p(j)$ and $p(A|j)$ depend on the choice of $p(j|i)$. The final algorithm,

$$\begin{cases} p_t(j|i) &= \frac{p_t(j)e^{\beta\Delta d(i,j)}}{\sum_{j'\in\widehat{S}} p_t(j')e^{\beta\Delta d(i,j')}}, \\ p_{t+1}(j) &= \sum_{i\in S} p_t(j|i)p(i), \\ p_{t+1}(A|j) &= \prod_{i\in A} p_t(j|i)\prod_{i\in A^{\mathsf{c}}}\left[1 - p_t(j|i)\right], \end{cases} \tag{2.17}$$

follows a similar alternating minimization scheme to the one developed by Blahut and Arimoto and generalized by Csiszár & Tsunády (1984), albeit with only local optimality guarantees similar to Tishby et al. (1999); Banerjee et al. (2005). See Appendix A and Alg. A1 for a complete derivation and description of the algorithm. The two practical issues with the algorithm are (1) the 2^n scaling of the number of subsets of S as n (the number of elements of S) increases, and (2) the general difficulty of estimating the mutual information between variables, let alone among multiple variables.

For the first issue, it is worth noting that there are non-trivial collective systems of empirical interest even for small n. Current computational hardware may permit exact computation up to around $n \approx 15$ even on consumer hardware, which would be relevant for many experimental systems (as in e.g. Miller & Gerlai, 2007; Katz et al., 2011; Jolles et al., 2018). For larger systems, Monte-Carlo estimation of $\Delta d(i,j)$ can be readily employed, e.g. for K samples,

$$\begin{aligned} \widehat{d}(i,j) &= \frac{1}{p(i)K}\sum_{k=1}^{K} r\big(A_{ij}\cup\{i\}\big), \\ \widehat{d}_{\mathsf{c}}(i,j) &= \frac{1}{p(i)K}\sum_{k=1}^{K} r\big(A_{ij}\setminus\{i\}\big), \quad \text{where } A_{ij}\sim f_i(\cdot|j). \end{aligned} \tag{2.18}$$

For large systems in particular initializing near good solutions may be helpful. In many systems we may expect elements to be spatially or temporally dependent, and use that prior knowledge to initialize reasonable clusters. However the preliminary results given in the next section do not employ any such strategy; we simply run the algorithm many times beginning with many different initial conditions and select the best solution generated.

While there is no exact universal solution to the practical difficulties of the second issue, we can proceed by maximizing a lower bound on component redundancy. For continuous random variables that are marginally Gaussian, the Gaussian mutual information is a lower bound on the total mutual information (Foster & Grassberger, 2011; Kraskov et al., 2004). Thus we can use

$$r(A) \geq \frac{I_G(\{X_i\}_{i\in A})}{\sum_{i\in A} H_G(X_i)}, \tag{2.19}$$

which is simple to compute in practice. When the random variables comprising the system are not marginally Gaussian, we can still use this bound by substituting copula transformed variables G_i for X_i, for which we enforce that each G_i is Gaussian distributed. We emphasize that this transformation is valid and unique for any set of continuous random variables; this is guaranteed by Sklar's theorem (Sklar, 1959), which ensures that the lower bound on redundancy given by Eq. 2.19 is applicable in general. The preliminary results in the next section are based on maximizing this Gaussian bound on redundancy using a copula transform to enforce Gaussian marginals.

2.3 Simulation experiments

We tested the proposed algorithm on two sets of data: simulations of schooling groups, and empirical data collected from the movements of schooling fish in a lab environment. The former allow us to control the dependency structure of the system, while the latter allows us to demonstrate applicability to empirical systems. Simulations used a simple model of coordinated movement based on attraction, alignment, and repulsion social forces (based on Romanczuk et al., 2012; Romanczuk & Schimansky-Geier, 2012; a description of the model and additional information on the simulation conditions can be found in Appendix B). Position and velocity data for independent groups of size $n = 5,\ 10$, and 20 were generated for a high ($\eta = 0.2$) and low ($\eta = 0.15$) noise conditions.

2.4 Empirical experiments

Movement data of fish comes from videos originally recorded by Katz et al. (2011). In that work, groups of 10, 30, and 70 golden shiners (*Notemigonus crysoleucas*) were purchased from Anderson Farms (www.andersonminnows.com) and filmed in a 1.2×2.1 m tank with an overhead camera. Videos were then corrected for lens distortion and fish were tracked using the same custom in-house software developed by Haishan Wu and used in Rosenthal et al. (2015). The software begins by detecting all individuals in each frame, then linking individuals across frames to form tracks. All tracks were manually corrected to ensure accuracy. Individual positions and velocities were estimated from these tracks using a 3$^{\text{rd}}$ order Savitzky-Golay filter (Savitzky & Golay, 1964; similar to e.g. Harpaz et al., 2017) with a 7 frame smoothing window (videos were recorded at 30 fps). Interactions between fish are time-dependent: in results presented we chose a fixed window of ± 15 s surrounding a given time t to estimate the dependency structure of the group.

3 Results

The algorithm outlined in the previous section requires specifying a parameter, β, which controls the relative importance of maximizing the average redundancy of the components as opposed to maximally compressing the original set of system elements. While it will be interesting to investigate the 'soft-partitioning' aspect of this approach in future work, here we simply consider the hard assignment case, which requires only that β is large. Fig. 2 (*Right*) illustrates this point, showing the stabilization of average component redundancy for $\beta > 5$. We found that $\beta = 200$ was sufficient to recover hard assignments in all cases tested here.[7] Since relative redundancy ranges between 0 and 1 for any dataset, these parameter values should generalize well to other systems, and leaves the method free of parameter fine-tuning.

To validate that the Monte-Carlo estimate of $\Delta d(i,j)$ employed is effective, we compared its behavior to exact computations of $\Delta d(i,j)$ for small system sizes (simulated groups of size 5 and 10). We ran each version of the algorithm for up to 10 components and took the best (maximum)

[7] Using the simultaneous updating variant of the algorithm, see Appendix A.

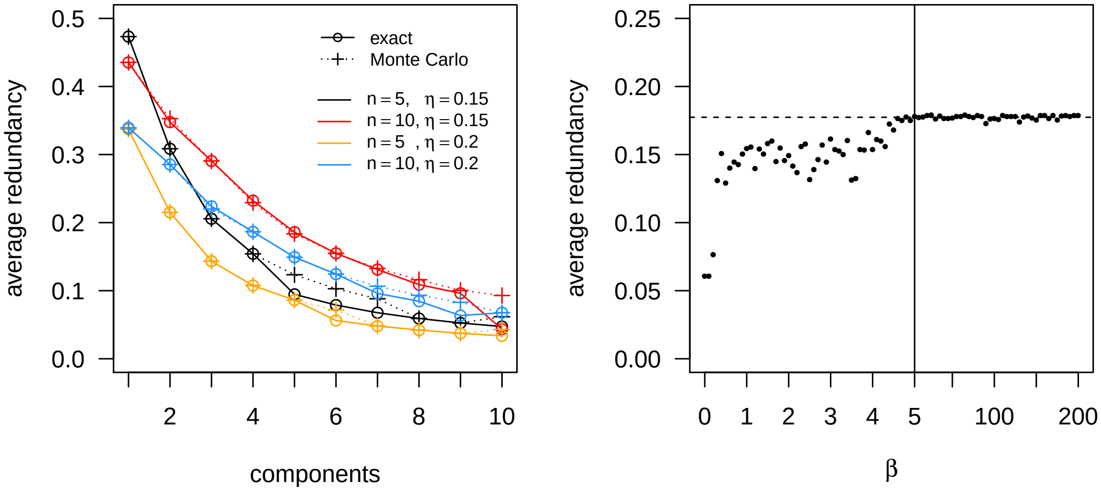


Figure 2. Algorithm implementation and parameter sensitivity. (*Left*) Comparison of exact and Monte Carlo estimates of $\Delta d(i, j)$, for single groups of size 5 and 10, for low and high noise conditions. (*Right*) Impact of the choice of β on the average redundancy of the recovered components for a single simulated group of size 10, high noise condition, searching for 5 components. Dotted line shows the mean of the solutions for $\beta > 5$.

average component redundancy achieved over 100 random initializations of the assignment matrix $p(j|i)$. Fig. 2 (*Left*) shows that the results are in good agreement, and where there are discrepancies they tend to favor the Monte Carlo method.

We tested the Monte Carlo algorithm on simulated data in which the dependency structure of the simulated groups was known. For each test, we computed the best average component redundancy recovered for up to 10 components, again using 100 random initializations of the assignment matrix for each computation. Average component redundancies for single groups of size 5, 10, and 20 (Fig. 3 *Left*) have a peak at a single component, which includes all the elements of the system. Redundancies for two non-interacting groups, in pairs of matched size groups of 5, 10, and 20, have peaks at 2 components for size 10 and 20, with a plateau or slight decline for the pair of size 5 (Fig. 3 *Center*). Finally, the redundancies for a system of three non-interacting groups of mixed sizes 5, 10, and 20 was computed, with peaks at 3 and 4 components for high and low noise conditions, respectively (Fig. 3 *Right*).[8] Taken together, this is evidence that the peaks and plateaus of the average component redundancies recovered by the algorithm do indeed reflect the dependency structure of the underlying system. It suggests that these features may be useful in identifying relevant structure in other systems, even those with less extreme dependency structures.

Fig. 4 illustrates the iterative generation of assignments for the algorithm in the mixed three group (high noise) case. Assignments change and harden until they converge on a (local) maximal average redundancy partition of the systems elements (*Left*). The assignments generated by the algorithm of system elements to components corresponds one-to-one with the original, non-interacting set of three groups (of sizes 5, 10, and 20) comprising the whole system (of total size 35). Positions of the elements of the system and their velocity vectors are shown for one time point, colored by

[8] In both noise conditions all three non-interacting groups were split into separate components. In the low noise condition, the group of 20 was further subdivided into two components.

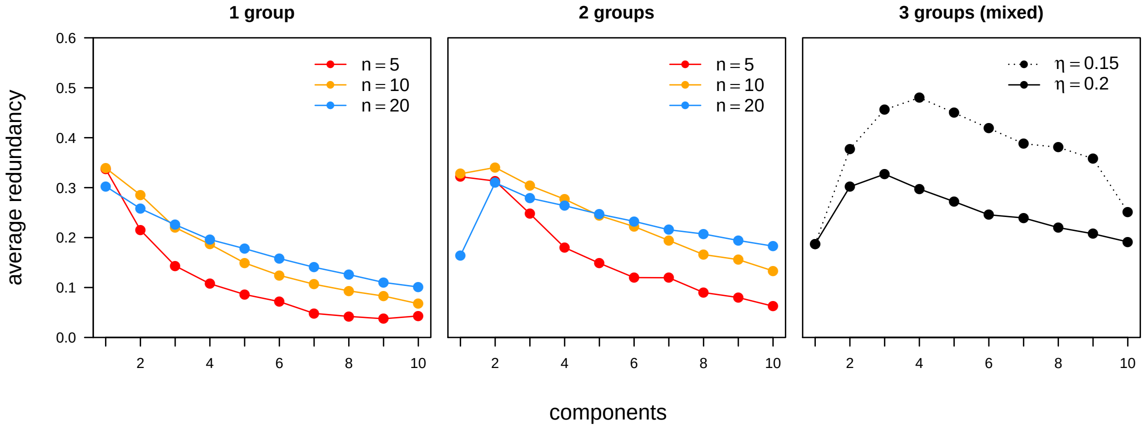


Figure 3. Partitioning results for simulations of 1, 2, or 3 independent (non-interacting) groups. (*Left*) For single cohesive groups of size (n) 5, 10, or 20, the average redundancy (y-axis) of the components decreases approximately monotonically as we increase the number of components. (*Center*) For two non-interacting groups of the same size, the average redundancy peaks, or at least plateaus, at two components. (*Right*) A mixed collection of three non-interacting groups, with sizes 5, 10, and 20, achieve peak average redundancy with three or four components, depending on the noise (η) used in the simulation. For comparison, the left two plots show results for $\eta = 0.2$ (the 'high' noise).

the component they were assigned to (which corresponds to their original group), in Fig. 4 (*Left*). Note that, while the snapshot shown in Fig. 4 was chosen to show the three distinct groups, at many points in the simulation the positions, velocities, or both, overlapped between the three groups.

Finally, we applied the algorithm to empirical data collected on fish schools to validate that the method is able to recover sensible results for strongly interacting groups and from non-simulated data. Fig. 5 shows that for fish, groups of size 10 interact strongly enough in at least one instance to be considered one coherent unit, while groups of size 30 are already large enough to have subsets that more strongly interact with one another than the rest of the group. Fish systems of size 70 do not always have a clear peak, but a broad plateau of possible subdivisions. The component assignments, positions, and velocities for a group of 30 fish is shown in Fig. 5 (*Right*) for three superimposed time points. Two of the components (shown as red and blue) show particularly coherent structure over the course of the 20 seconds of movement shown. Further work is needed to investigate the duration of substructure in fish schools, as well as the emergence and disappearance of components over time.

4 Discussion

There are a wide range of both general purpose clustering algorithms (see Jain, 2010; Xu & Tian, 2015) and network community detection methods (see Forunato, 2010), owing to a diversity of plausible clustering and community detection criteria. The justification for the average relative redundancy criterion presented here stems from its principled approach to the specific problem of quantifying collectivity, as argued at the beginning of this paper, and its demonstrated ability to identify the dependent structure of collective systems, as shown by the results obtained in the previous section. Usage of this criterion requires sufficient information for estimating the redundancy of any subset of the system, thus it cannot be used as a drop-in replacement for other clustering or com-

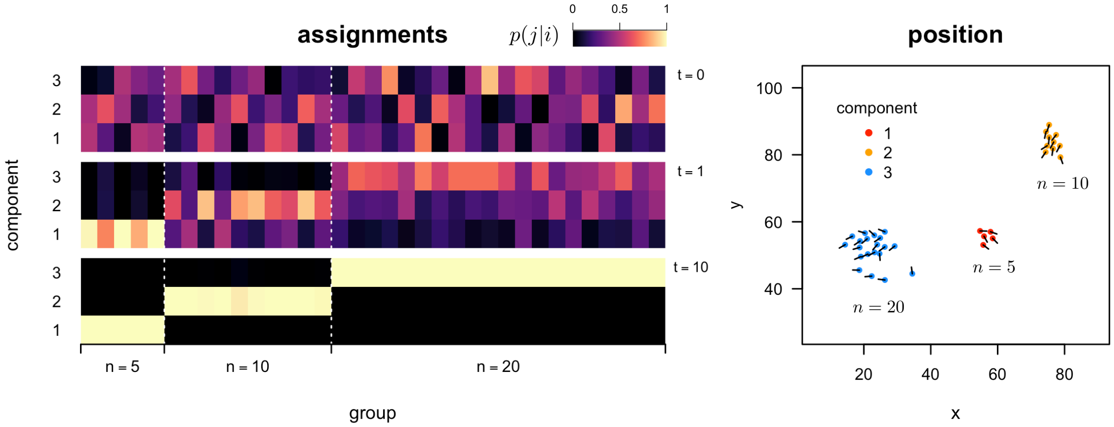


Figure 4. Generation of assignments for a mixture of three non-interacting simulated groups. (*Left*) Assignments generated by the proposed sequential algorithm for three components after initialization ($t = 0$), 1 iteration ($t = 1$), and 10 iterations ($t = 10$), at *top*, *middle*, and *bottom*, respectively. The color scale indicates the probability of assigning a member of a group (column) to a particular component (row), where low to high probability is coded dark to light (color scale top right). Original groupings of the system into its three non-interacting subsets are indicated on the x-axis. (*Right*) Two-dimensional positions (arbitrary units) of simulated system at one time point, color-coded by final component assignment; velocity vectors indicated by line segments.

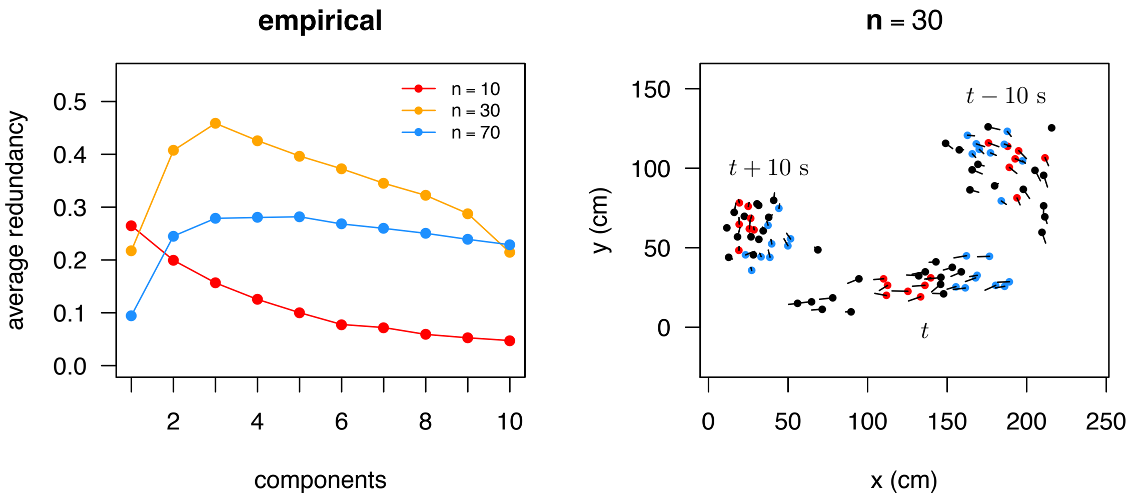


Figure 5. Redundant substructure for empirical fish schools. (*Left*) Average component redundancy as a function of the number of components, for fish groups of size 10, 30, and 70. (*Right*) Example partitioning of a group of size 30 fish into three components, colored black, red, and blue. Dots indicate the positions of the fish in a large ($1.2 \text{ m} \times 2.1 \text{ m}$) arena, while line segments indicate the velocity vector of each individual. Positions (cm) and velocities are shown superimposed for the group at times t and $t \pm 10$ s.

munity detection methods that operate on arbitrary similarity or correlation matrices. If the lower bound on redundancy given by Eq. 2.19 is to be used, then marginal normality must be enforced via copula transform or directly by the process generating the data.

In addition to its theoretical and empirical justification, the method is also computationally efficient, making it usable in practice. The proposed Monte-Carlo algorithm improves on the computational complexity of both the brute-force (check every partition) and naïve exact (sum over every subset) solutions, while achieving comparable results. It is instead limited by the cost of computing, for each element, log determinants for the Gaussian average redundancy bound. This reduces to matrix multiplication and thus scales (depending on the method) as $O\big(n^{3+1}m\big)$ or $O\big(n^{2.373+1}m\big)$, where m is the number of components, assuming a fixed number of Monte-Carlo samples and total iterations of the algorithm. In fact, this worst-case bound will almost never be achieved in practice, since it requires probabilistically sampling at least one assignment of all n elements to each of the m components when evaluating $\Delta d(i, j)$. The probability this occurs decreases exponentially in n.

There are many open questions left for future work. First, the identification of a peak in the average redundancy plot as a function of the number of components is only a heuristic. In some cases, as in e.g. the group of 70 fish studied here, there may be no obvious peak, and thus there may be more than one useful decomposition of the group. In other cases, depending on the question being asked, it may be more appropriate to divide the group into a given number of components regardless of the existence or position of a peak. Further theoretical work is needed on the significance of peaks or plateaus in the average redundancy plot; we present only empirical evidence of their utility here. Second, an investigation of these features as a function of the time-window chosen for computing the dependency structure may be important for understanding how the dependency structure of the group scales with time. It might be expected that this in fact plays a very important role, in that on short time-scales for many systems only very local interactions will matter, while at long enough time scales the system is best represented as only one component.

It may also be important to investigate the algorithm presented here in the context of generating a soft-partitioning of a system's elements into partially overlapping components. Using intermediate values of β may allow the algorithm to find better average redundancy solutions 'in-between' m and $m + 1$ components, in which the assignments for some elements are split between some set of components. At the same time, since optimal sets of components are not guaranteed to be unique, it may be important to explore the set of equally (or nearly equally) optimal solutions as an ensemble of equivalent descriptions of a system. Moreover, exploring the range of solutions as the number of components varies may reveal whether or not the system exhibits some form of hierarchical dependency structure. In hierarchical systems we would expect components to be successively subdivided as the number of components increases, while in non-hierarchical systems this would not be the case.

One of the most interesting potential applications of the method may be to long time-series data for collective systems, in which the dependency structure of the group changes over time. Characterizing the natural decomposition of a system as a function of time may reveal important time-dependent mesoscopic features. How do the natural number of components of a system fluctuate in time, and how long do components persist? How do the components of a system interact as a function of time? These questions are central to the study of collective systems and can now be addressed quantitatively via the method introduced here.

References

Arimoto, S. (1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, **18**(1):14–20.

Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cisbani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., Viale, M., & Zdravkovic, V. (2008) Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study. *PNAS*, **105**(4):1232–1237.

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005) Clustering with bregman divergences. *Journal of Machine Learning Research*, **6**:1705–1749.

Beggs, J. M. & Plenz, D. (2003) Neuronal avalanches in neocortical circuits. *Journal of Neuroscience*, **23**(35):11167–11177.

Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., & Walczak, A. M. (2012) Statistical mechanics for natural flocks of birds. *PNAS*, **109**(13):4786–4791.

Blahut, R. (1972) Computation of channel capacity and rate-distortion function. *IEEE Transactions on Information Theory*, **18**(4):460–473.

Buhl, J., Sumpter, D. J. T., Couzin, I. D., Hale, J. J., Despland, E., Miller, E. R., & Simpson, S. J. (2006) From disorder to order in marching locusts. *Science*, **312**(5778):1402–1406.

Couzin, I. D., Krause, J., James, R., Ruxton, G. D., & Franks, N. R. (2002) Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, **218**:1–11.

Cover, T. M. & Thomas, J. A. (2006) *Elements of information theory*. Wiley-Interscience, 2$^{\text{nd}}$ edition.

Csiszár, I. & Tsunády, G. (1984) Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplement Issue*, **1**:205–237.

Flack, J. C., Girvan, M., de Waal, F. B. M., & Krakauer, D. C. (2006) Policing stabilizes construction of social niches in primates. *Nature*, **439**:426–429.

Forunato, S. (2010) Community detection in graphs. *Physics Reports*, **486**:75–174.

Foster, D. V. & Grassberger, P. (2011) Lower bounds on mutual information. *Physical Revew E*, **83**:010101.

Greene, M. J. & Gordon, D. M. (2007) Interaction rate informs harvester ant task decisions. *Behavioral Ecology*, **18**(2): 451–455.

Harpaz, R., Tkačik, G., & Schneidman, E. (2017) Discrete modes of social information processing predict individual behavior of fish in a group. *PNAS*. doi: 10.1073/pnas.1703817114.

Herbert-Read, J. E., Perna, A., Mann, R. P., Schaerf, T. M., Sumpter, D. J. T., & Ward, A. J. W. (2011) Inferring the rules of interaction of shoaling fish. *Proceedings of the National Academey of Sciences*, **108**(46):18726–18731.

Jain, A. K. (2010) Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, **31**:651–666.

Jolles, J. W., Laskowski, K. L., Boogert, N. J., & Manica, A. (2018) Repeatable group differences in the collective behaviour of stickleback shoals across ecological contexts. *Proceedings of the Royal Society B*, **285**(1872):20172629.

Katz, Y., Tunstrøm, K., Ioannou, C. C., Huepe, C., & Couzin, I. D. (2011) Inferring the structure and dynamics of interactions in schooling fish. *PNAS*, **108**(46):18720–18725.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004) Estimating mutual information. *Physical Revew E*, **69**:066138.

Lukeman, R., Li, Y.-X., & Edelstein-Keshet, L. (2010) Inferring individual rules from collective behavior. *Proceedings of the National Academey of Sciences*, **107**(28):12576–12580.

Miller, N. & Gerlai, R. (2007) Quantification of shoaling behaviour in zebrafish (*Danio rerio*). *Behavioural Brain Research*, **184**(2):157–166.

Nagy, M., Ákos, Z., Biro, D., & Vicsek, T. (2010) Hierarchical group dynamics in pigeon flocks. *Nature*, **464**:890–893.

Papenfort, K. & Bassler, B. (2016) Quorum sensing signal-response systems in gram-negative bacteria. *Nature Reviews Microbiology*, **14**:576–588.

Reid, C. R., Lutz, M. J., Powell, S., Kao, A. B., Couzin, I. D., & Garnier, S. (2015) Army ants dynamically adjust living bridges in response to a cost-benefit trade-off. *PNAS*, **112**(49):15113–15118.

Romanczuk, P. & Schimansky-Geier, L. (2012) Swarming and pattern formation due to selective attraction and repulsion.

Interface Focus, **2**(6):746–756.

Romanczuk, P., Bär, M., Ebeling, W., Lindner, B., & Schimansky-Geier, L. (2012) Active brownian particles. *The European Physical Journal Special Topics*, **202**(1):1–162.

Rosenthal, S. B., Twomey, C. R., Hartnett, A. T., Wu, H. S., & Couzin, I. D. (2015) Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *PNAS*, **112**(15):4690–4695.

Rothstein, J. (1952) Organization and entropy. *Journal of Applied Physics*, **23**:1281–1282.

Savitzky, A. & Golay, M. J. E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36**(8):1627–1639.

Seeley, T. D. & Visscher, P. K. (2004) Quorum sensing during nest-site selection by honeybee swarms. *Behavioral Ecology and Sociobiology*, **56**(6):594–601.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**(3):379–423.

Shannon, C. E. (1959) Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, **7**(4):142–163.

Sklar, A. (1959) Fonctions de répartition à n-dimensions et leurs marges. *Publications de l'Institute de Statistique de l'Université de Paris*, **8**:229–231.

Slonim, N., Atwal, G. S., Tkačik, G., & Bialek, W. (2005) Information based clustering. *PNAS*, **102**(51):18297–18302.

Strandburg-Peshkin, A., Twomey, C. R., Bode, N. W. F., Kao, A. B., Katz, Y., Ioannou, C. C., Rosenthal, S. B., Torney, C. J., Wu, H. S., Levin, S. A., & Couzin, I. D. (2013) Visual sensory networks and effective information transfer in animal groups. *Current Biology*, **23**:R709–R711.

Tishby, N., Pereira, F. C., & Bialek, W. (1999) The information bottleneck method. In Hajek, B. & Sreenivas, R. S., editors, *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377. University of Illinois Press.

Torney, C. J., Lamont, M., Debell, L., Angohiatok, R. J., Leclerc, L.-M., & Berdahl, A. M. (2018) Inferring the rules of social interaction in migrating caribou. *Philosophical Transactions of the Royal Society B*, **373**(1746):20170385.

Tunstrøm, K., Katz, Y., Ioannou, C. C., Huepe, C., Lutz, M. J., & Couzin, I. D. (2013) Collective states, multistability and transitional behavior in schooling fish. *PLOS Computational Biology*, **9**(2):e1002915.

Watanabe, S. (1960) Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, **4**:66–82.

Xu, D. & Tian, Y. (2015) A comprehensive survey of clustering algorithms. *Annals of Data Science*, **2**(2):165–193.

A Algorithm

Here we give an expanded account of the redundancy compression algorithm.

A.1 Rate-Distortion Compression

Classical rate-distortion theory treats the following optimization problem:

$$\begin{array}{lrcll} \underset{p(\hat{x}|x)}{\text{minimize}} & I(X;\widehat{X}) & & & \\ \text{subject to} & \mathbb{E}[d(x,\hat{x})] & \leq & D & \\ & p(\hat{x}|x) & \geq & 0 & \forall (x,\hat{x}) \in (X,\widehat{X}) \\ & \sum_j p(\hat{x}|x) & = & 1 & \forall x \in X, \end{array} \tag{A.1}$$

where

$$\mathbb{E}\left[d(x,\hat{x})\right] = \sum_{\hat{x}\in\widehat{X}} \sum_{x\in X} p(\hat{x}|x)p(x)d(x,\hat{x}), \tag{A.2}$$

and $p(x)$ is given. The problem as stated is not convex due to the form of $I(X;\hat{X})$. However, writing the objective as

$$I(X;\widehat{X}) = \sum_{x,\hat{x}} p(\hat{x}|x)p(x)\log p(\hat{x}|x) - \sum_{x,\hat{x}} p(\hat{x}|x)p(x)\log p(\hat{x}), \tag{A.3}$$

it is clear that the problem is convex when varying $p(\hat{x}|x)$ or $p(\hat{x})$ separately, holding the other constant. Since the distortion constraint, $\mathbb{E}\left[d(x,\hat{x})\right]$ is convex in $p(\hat{x}|x)$, the problem can be restated as a convex double minimization of the form

$$\min_{p(\hat{x}|x)} \min_{p(\hat{x})} I(X;\widehat{X}), \tag{A.4}$$

which is minimized for fixed $p(\hat{x}|x)$ by

$$p(\hat{x}) = \sum_{x} p(\hat{x}|x)p(x), \tag{A.5}$$

and for fixed $p(\hat{x})$ by

$$p(\hat{x}|x) = \frac{p(\hat{x})\exp\left[-\beta d(x,\hat{x})\right]}{\sum_{\hat{x}'} p(\hat{x}')\exp\left[-\beta d(x,\hat{x}')\right]}, \tag{A.6}$$

(see Blahut, 1972; Arimoto, 1972; Cover & Thomas, 2006). This leads to the classic Blahut-Arimoto algorithm, which, by iterative application of these two self-consistent equations for a given β, converges to an optimal solution point on the rate-distortion curve with tangent slope equal to β.

A.2 Redundancy Compression

In this paper, we are interested in a similar problem:

$$\begin{array}{llll} \underset{p(j|i)}{\text{minimize}} & I(S;\widehat{S}) & & \\ \text{subject to} & \mathbb{E}[r(A,j)] \geq r^* & \forall j \in \widehat{S} & \\ & p(j|i) \geq 0 & \forall (i,j) \in (S,\widehat{S}) & \\ & \sum_j p(j|i) = 1 & \forall i \in S, & \end{array} \tag{A.7}$$

where

$$\mathbb{E}\left[r(A,j)\right] = \frac{1}{m}\sum_{j\in\widehat{S}} r(A,j) \tag{A.8}$$

and

$$r(A,j) = \sum_{A\in\mathcal{P}(S)} r_A \prod_{i\in A} p(j|i) \prod_{i\in A^{\mathsf{c}}} \left[1-p(j|i)\right]. \tag{A.9}$$

The fixed $1/m$ weighting of the marginal importance of each component, j, in the redundancy constraint, $\mathbb{E}\left[r(A,j)\right]$, is a minor variation from the classical rate-distortion problem. The important difference is that the $r(A,j)$ inequality constraint is not convex with respect to $p(j|i)$. However, with change of variables $b_A = \log r_A$, $y_{ij} = \log p(j|i)$, and $\bar{y}_{ij} = \log\left[1-p(j|i)\right]$, we can define

$$g(A,j) = \sum_{A\in\mathcal{P}(S)} \exp\left[\sum_{i\in A} y_{ij} + \sum_{i\in A^{\mathsf{c}}} \bar{y}_{ij} + b_A\right], \tag{A.10}$$

where $r(A,j) = g(A,j)$, with $g(A,j)$ convex with respect to y_{ij} and $\bar{y}_{ij}$ and invariant with respect to $p(j|i)$ or $p(j)$.

This gives the equivalent minimization problem:

$$\begin{array}{llll} \underset{p(j|i)}{\text{minimize}} & I(S;\widehat{S}) & & \\ \text{subject to} & \mathbb{E}[g(A,j)] \geq r^* & \forall j \in \widehat{S} & \\ & p(j|i) \geq 0 & \forall (i,j) \in (S,\widehat{S}) & \\ & \sum_j p(j|i) = 1 & \forall i \in S & \\ & e^{y_{ij}} \leq p(j|i) & \forall (i,j) \in (S,\widehat{S}) & \\ & e^{\bar{y}_{ij}} \leq 1-p(j|i) & \forall (i,j) \in (S,\widehat{S}). & \end{array} \tag{A.11}$$

Setting aside non-negativity constraints on $p(j|i)$ (these will be enforced by the form of the solution), we have the functional

$$L\left[p(j|i);p(j);y_{ij},\bar{y}_{ij}\right] = \sum_{i,j} p(j|i)\log\frac{p(j|i)}{p(j)} + \sum_i \lambda(i)\sum_j p(j|i) \tag{A.12}$$

$$- \beta \sum_{j, A \in \mathcal{P}(S)} \exp \left[\sum_{i \in A} y_{ij} + \sum_{i \in A^{\mathsf{c}}} \bar{y}_{ij} + b_A \right] \tag{A.13}$$

$$+ \sum_{i,j} \lambda(i,j) \left[e^{y_{ij}} - p(j|i) \right] \tag{A.14}$$

$$+ \sum_{i,j} \bar{\lambda}(i,j) \left[e^{\bar{y}_{ij}} + p(j|i) \right] . \tag{A.15}$$

We can then restate the original non-convex problem in terms of two convex minimizations and one quasiconvex minimization,

$$\min_{p(j|i)} \min_{p(j)} \min_{y_{ij}, \bar{y}_{ij}} L \left[p(j|i); p(j); y_{ij}, \bar{y}_{ij} \right] . \tag{A.16}$$

Note that, similar to Tishby et al. (1999), the problem is not jointly convex and thus there is no guarantee of a unique global solution as in the rate-distortion case. Nevertheless, the marginal (quasi-)convexity admits an efficient iterative algorithm for identifying (locally) optimal solutions, similar to Tishby et al. (1999).

Taking the derivative of L with respect to $p(j|i)$ and setting equal to zero, we arrive at

$$p(j|i) = \frac{p(j)}{\mu(i)} \exp \left[p(i)^{-1} \left[\lambda(i,j) - \bar{\lambda}(i,j) \right] \right] , \tag{A.17}$$

where $\mu(i)$ just normalizes the distribution over j for a given i. Taking the derivative of L with respect to y_{ij} and setting equal to zero, we have

$$\lambda(i,j) = \beta e^{-y_{ij}} \sum_{\{A \in \mathcal{P}(S) \,:\, i \in A\}} \exp \left[\sum_{k \in A} y_{kj} + \sum_{k \in A^{\mathsf{c}}} \bar{y}_{kj} + b_A \right] . \tag{A.18}$$

Doing the same for $\bar{y}_{ij}$ gives

$$\bar{\lambda}(i,j) = \beta e^{-\bar{y}_{ij}} \sum_{\{A \in \mathcal{P}(S) \,:\, i \in A^{\mathsf{c}}\}} \exp \left[\sum_{k \in A} y_{kj} + \sum_{i \in A^{\mathsf{c}}} \bar{y}_{kj} + b_A \right] . \tag{A.19}$$

Subtracting the two equations, we have

$$\beta \Delta d(i,j) = \lambda(i,j) - \bar{\lambda}(i,j), \tag{A.20}$$

which is equivalent to the definition of $\Delta d(i,j)$ in the main text. Substituting into Eq. A.17 produces

$$p(j|i) = \frac{p(j)}{\mu(i)} \exp \left[\frac{\beta}{p(i)} \Delta d(i,j) \right] . \tag{A.21}$$

This gives the minimizing values of L with respect to $p(j|i)$ for fixed $p(j)$, y_{ij}, and $\bar{y}_{ij}$, as in Blahut (1972); Arimoto (1972); Tishby et al. (1999); Banerjee et al. (2005). The minimizing values of L with respect to $p(j)$ are the same as in classical rate-distortion theory and are given by

$$p(j) = \sum_{i} p(j|i) p(i). \tag{A.22}$$

The minimizing value of L with respect to y_{ij} and $\bar{y}_{ij}$ under the constraints that $e^{y_{ij}} \leq p(j|i)$, and $e^{\bar{y}_{ij}} \leq [1 - p(j|i)]$, is simply

$$y_{ij} = \log p(j|i), \tag{A.23}$$

$$\bar{y}_{ij} = \log \left[1 - p(j|i)\right], \tag{A.24}$$

since the monotonically decreasing A.13 will achieve its minimum for the least negative values of y_{ij} and $\bar{y}_{ij}$, which puts them up against their constraints.

A.3 Generalization

It is clear from the form of $g(A, j)$ that the only requirement of the measured property, b_A, of any set, $A \in S$, is that it is non-negative. Thus this same method may be employed for measures on sets other than redundancy, in the same way that rate-distortion theory treats generic measures of distortion. On the other hand, when the measured property offers certain kinds of additional structure, as in e.g. the case of an average similarity (Slonim et al., 2005) measure, then other efficient solutions may be possible.

One variant to the sequential update of $p(j|i)$ as listed in Alg. A1 is to modify every $p(j|i)$ in parallel, which may be advantageous for some multiprocessor configurations. In practice, for convergence with simultaneous updating it appears to be important to introduce a slowdown factor, α, to control the update of $p_t(j|i)$, i.e. using

$$p_t(j|i) = \alpha \frac{p_t(j) e^{\beta \Delta d(i,j)}}{\sum_{j' \in \widehat{S}} p_t(j') e^{\beta \Delta d(i,j')}} + (1 - \alpha) p_{t-1}(j|i), \tag{A.25}$$

where t is the current iteration of the algorithm. The slowdown operates in a manner analogous to the learning rate in gradient descent optimization problems.

Like β, α does not require fine-tuning. It just needs to be small enough to allow for convergence, without being too small so as to allow the algorithm to converge in a reasonable number of iterations. While a more systematic investigation may be useful in identifying an efficient α, we found that $\alpha = 0.1$ and $t = 200$ iterations was sufficient to ensure convergence for all the numerical results presented in the main text. In many cases a stable assignment is reached much earlier than after 200 iterations, and in general a stopping criteria based on the difference between assignments from one iteration to the next could be employed, though we did not do so here.

Algorithm 1: Alternating minimization

```
input       :                  features  $X_1, X_2, \ldots X_n$
output      :               assignments  $p(j|i) \in [0,1],\ \forall (i,j) \in (S, \widehat{S})$

parameters  : number of components  $m \in \mathbb{N}_0^+$
            :     assignment hardness  $\beta \in \mathbb{R}_0^+$
            :         total iterations  $t_{\max} \in \mathbb{N}^+$

constraints :               normalized  $\sum_j p(j|i) = 1,\ \forall i \in S$
            :             non-negative  $p(j|i) \geq 0,\ \forall (i,j) \in (S, \widehat{S})$

   initialization with flat Dirichlet prior
1  foreach $i \in S$ do
2      $p(j|i) \sim \mathrm{Dir}(m, \mathbf{1})$

   iteratively improve assignments
3  foreach $t \in 1, \ldots, t_{\max}$ do
4      foreach $i \in S$ do
           minimization with respect to $p(j)$
5          foreach $j \in \widehat{S}$ do
6              $p(j) \leftarrow \sum_{i \in S} p(j|i) p(i)$

           minimization with respect to $y_{ij}$ and $\bar{y}_{ij}$
7          foreach $(i,j) \in (S, \widehat{S})$ do
8              $y_{ij} \leftarrow \log p(j|i)$
9              $\bar{y}_{ij} \leftarrow \log\,[1 - p(j|i)]$

           minimization with respect to $p(j|i)$
10         $p(j|i) \leftarrow \frac{p(j) e^{\beta \Delta d(i,j)}}{\sum_{j' \in \widehat{S}} p(j') e^{\beta \Delta d(i,j')}}$
```

Algorithm A1. $\mathbb{N}^+$ are the positive integers, while $\mathbb{N}_0^+$, $\mathbb{R}_0^+$, are the non-negative (positive including zero) integers and real numbers, respectively. For hard clustering, β just needs to be large. Parameter $t_{\max}$ needs to be large enough for convergence; alternatively, it can be replaced by a criterion based on a minimum difference in improvement between iterations. Lines 2 and 10 are to be understood as vector operations over the set $j \in \widehat{S}$.

B Simulation

The agent-based model used in this paper for generating schooling motion with known dependency structure is based on the three-zone-model introduced by Couzin et al. (2002). Each agent moves at a constant speed s_0 and responds to its conspecifics by changing its direction of motion. The interactions between individuals are governed by three basic social forces: long-range attraction, short-range repulsion, and intermediate-range alignment. However, there are two main differences from the original Couzin model: (1) the model is formulated in terms of stochastic differential equations with effective social forces (see Romanczuk et al., 2012; Romanczuk & Schimansky-Geier, 2012); and (2) instead of discrete zones, we use overlapping social forces, whereby repulsion dominates at short distances ($r_{ij} < r_{\text{rep}}$), attraction dominates at long distances $r_{ij} < r_{\text{att}}$, and the alignment contribution overlaps with attraction and repulsion up to intermediate ranges ($r_{ij} < r_{\text{alg}}$), whereby $r_{\text{rep}} < r_{\text{alg}} < r_{\text{att}}$.

B.1 Model formulation

We simulate the movement of a group of n agents via a set of $2n$ (stochastic) differential equations. The agents move in a quadratic domain of size $L \times L$ with periodic boundary conditions. The dynamics of each agent (in 2d) are described by the following equations of motion ($i = 1, \dots, n$):

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i(t), \qquad \text{with} \quad \mathbf{v}_i(t) = \begin{pmatrix} s_0 \cos(\varphi_i(t)) \\ s_0 \sin(\varphi_i(t)) \end{pmatrix}, \tag{B.1}$$

$$\frac{d\varphi_i}{dt} = \frac{1}{s_0}\left(F_{i,\varphi} + \eta_{i,\varphi}\right). \tag{B.2}$$

Here $\mathbf{r}_i$, and $\mathbf{v}_i$ are the Cartesian position and velocity vectors of each agent, with s_0 being the (constant) speed of agent i. Furthermore, $\eta_{i,\varphi}$ are Gaussian white noise terms accounting for randomness in the turning motion of individuals, and $\mathbf{f}_{i,\varphi}$ are the projections of the total social forces inducing turning behavior, where

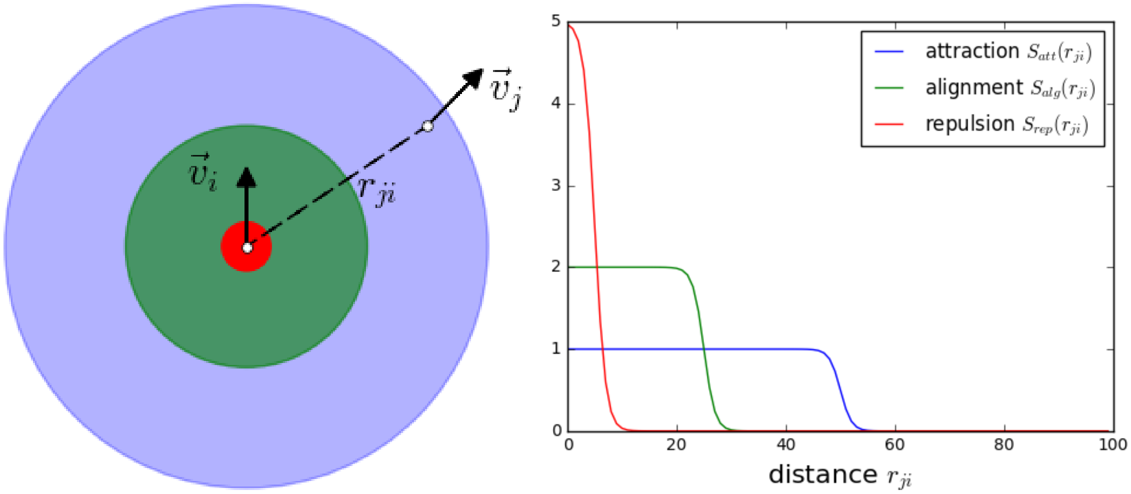
$$F_{i,\varphi} = \mathbf{f}_i \cdot \mathbf{u}_{\varphi,i} = \mathbf{f}_i \begin{pmatrix} -s_0 \sin\varphi_i \\ s_0 \cos\varphi_i \end{pmatrix}. \tag{B.3}$$

The total effective social force is a sum of three components, $\mathbf{f}_i = \mathbf{f}_{i,\text{rep}} + \mathbf{f}_{i,\text{alg}} + \mathbf{f}_{i,\text{att}}$,

$$\begin{aligned} &\text{Attraction} \qquad \mathbf{f}_{i,\text{rep}} = \sum_{j\in\text{Neigh}} +\mu_{\text{att}} S_{\text{att}}(r_{ji})\hat{r}_{ji}, \\ &\text{Repulsion} \qquad \mathbf{f}_{i,\text{rep}} = \sum_{j\in\text{Neigh}} -\mu_{\text{rep}} S_{\text{rep}}(r_{ji})\hat{r}_{ji}, \\ &\text{Alignment} \qquad \mathbf{f}_{i,\text{alg}} = \sum_{j\in\text{Neigh}} \mu_{\text{alg}} S_{\text{alg}}(r_{ji})(\mathbf{v}_j - \mathbf{v}_i), \end{aligned} \tag{B.4}$$

with $\hat{r} = \mathbf{r}/|r|$. The strength of the different interactions is set by a constant μ_X and a sigmoid function of distance, which goes from 1 to 0, with the transition point at r_X and steepness a_X:

$$S_X(r) = \frac{1}{2}\left(\tanh(-a(r - r_X) + 1\right)$$



Algorithm B1. (*Left*) Schematic of the effective social interactions, with repulsion dominating at short distances (red zone), attraction dominating at large distances (green zone) and main contribution of alignment at intermediate ranges (blue zone). (*Right*) The strength of the different social forces versus distance for the different interactions.

(see Fig B1).

The stochastic differential equations for the direction of motion of individual agents are solved by a simple Euler-Maruyama method:

$$\varphi(t+1) = \varphi(t) + \frac{1}{s_0}\left(F_{i,\varphi}(t)\Delta t + \sqrt{2D_\varphi \Delta t}\,\text{GRN(t)}\right), \tag{B.5}$$

$$\mathbf{r}(t+1) = \mathbf{r}(t) + \begin{pmatrix} s_0 \cos(\varphi_i(t)) \\ s_0 \sin(\varphi_i(t)) \end{pmatrix} \Delta t. \tag{B.6}$$

B.2 Numerical experiments

We simulated independent groups of three different sizes, $n = 5$, 10, and 15, wherein it was possible for each agent to interact with the distance dependent effective forces with all other agents within the group. The initial conditions were always a random distribution of agents in the simulation domain with random initial direction of motion. In order to ensure formation of a single cohesive group we set the attraction range to be larger then the domain size $r_{\text{att}} > L$. In all simulation runs considered here, we obtained for the used parameters (see Tab. 1) a single polarized group after a transient time of $t < 400$. Thus for our analyses we used only data for $t > 400$.

Parameter	Symbol	Value
domain size	L	100
repulsion range	r_{rep}	1.0
attraction range	r_{att}	100.0
alignment range	r_{alg}	5.0
repulsion strength	μ_{rep}	2.0
attraction strength	μ_{att}	0.3
alignment strength	μ_{alg}	0.8
steepness of interaction function	a	10
speed of individuals	s_0	1.0

Table 1. Parameter values used in the simulations.