# Maximizing Multivariate Information with Error-Correcting Codes 

Kyle Reing*, Greg Ver Steeg, and Aram Galstyan


#### Abstract

Multivariate mutual information provides a conceptual framework for characterizing higher-order interactions in complex systems. Two well-known measures of multivariate information-total correlation and dual total correlationadmit a spectrum of measures with varying sensitivity to intermediate orders of dependence. Unfortunately, these intermediate measures have not received much attention due to their opaque representation of information. Here we draw on results from matroid theory to show that these measures are closely related to error-correcting codes. This connection allows us to derive the class of global maximizers for each measure, which coincide with maximum distance separable codes of order $k$. In addition to deepening the understanding of these measures and multivariate information more generally, we use these results to show that previously proposed bounds on information geometric quantities are tight at the extremes.


## I. INTRODUCTION

Characterizing the interactions in a system by its pairwise correlations is a common modeling assumption, but a more complete picture would also consider many-to-one, and many-to-many interactions. In systems with emergent properties, such as those studied in ecology [1], systems biology [2], neuroscience [3], and genetics [4] (to name a few), these higher-order interactions play a vital role in macroscopic behavior. The issue of defining a proper measure of higher-order correlation, at least from an information theoretic perspective, is what the study of multivariate mutual information (MMI) seeks to do. Tools based on MMI and the related subject of information decomposition have already seen diverse application, offering insights into problems like structure discovery in biological/artificial neural networks [5], [6], [7], [8], [9], [10], [11], and in determining how genes cooperate to produce a particular phenotypic effect [12], [13]. Two major challenges faced by any work attempting to utilize MMI are the selection of an appropriate measure, and its computational cost. The first concern is discussed in detail by Timme et al. [14], responding to the fact that no single measure is universally accepted by the community. Choosing from the multitude of measures, each with their own idiosyncratic definition of group information, often boils down to preference, application, or commitment to certain axiomatic principles. To add to this long list of considerations, the cost of analytically or approximately calculating the measure may be prohibitive in all but the smallest applications. The most expressive approaches (such as those based on the partial information decomposition framework [15]) are usually the most intractable, limiting their practical

[^0]use to the simplest forms of multivariate dependence [16]. All this serves to reiterate that MMI is far from solved, and could benefit greatly from continued study, especially regarding the aforementioned computational problems.

Our contribution to this ongoing dialogue is to identify the global maximizers for a class of MMI measures related to Watanabe's total correlation. Finding the extrema of a function is usually a matter of optimization, but for nonconvex and/or high-dimensional functions, there may not be a guaranteed strategy for convergence to the optimal solution. Knowledge of a closed form generative procedure for the extrema of MMI is desirable for at least two reasons: 1) these properties inform us about the types of dependence preferentially treated by the measure, which is important when comparing different approaches; 2) additionally, the scalar value of MMI for any distribution becomes more meaningful when it can be ranked according to its relative distance to the min and max.

The main theoretical results of this work connect the global maximizers for each of the $n-1$ measures ( $n$ being the number of random variables) with the class of maximum distance separable (MDS) codes [17] (equiv. ideal secret sharing schemes [18]). Through this connection, it follows that a complete description of the set of global maximizers is still unknown, as this knowledge could be used to solve an open conjecture in coding theory and finite projective geometry [19] known as the MDS conjecture. The proof of this result draws on a few disciplines, including matroid theory and coding theory. Since we expect readers will come from varying backgrounds, we have tried to make the writing expository and the material self-contained. Section $\square$ Introduces the family of measures, summarizes the main contributions in non-technical terms, and runs through an example of why an uninformed search procedure is unlikely to discover the maximizers. The preliminaries of Section III are split into two nearly disjoint subsections, with each presenting only the concepts necessary for both understanding and completion of the final proof. The proofs of the main claim appear in Section IV, building on the definitions, theorems, and intuitions of Sections $\Pi$ and III Section $V$ discusses important related work, focusing on a family of information geometric measures shown to be upper-bounded by the measures studied here. Previous results on these bounds are strengthened in light of the connection to MDS codes. Finally, Section VI concludes with a discussion of where this work fits with regards to general trends in MMI research, highlighting the appearance of coding theory and cryptography in many recent papers.

## II. COHESION MEASURES

## A. Total Correlation and Dual Total Correlation

Historically, the study of multivariate mutual information began with the introduction of two measures, one being Watanabe's total correlation (TC) [20] (equiv. multiinformation [21]), defined as

$$
T C(X)=\sum_{i=1}^{n} H\left(X_{i}\right)-H(X) .
$$

Here the vector random variable $X=\left\{X_{1}, \ldots, X_{n}\right\}$ will be taken over a finite support $\|X\|$, and entropy is defined in the standard way as $\sum_{x \in\|X\|} p(x) \log [1 / p(x)]$. We adopt the convention that each marginal $X_{i}$ has the same support $q$ without loss of generality, implying $\|X\| \leq q^{n}$. Additionally, we assume that our logarithms are taken with respect to base $q$.

A related, and equally important measure for our discussion is the dual total correlation (DTC) (equiv. excess entropy [22], binding information [23]), originating from Han's [24] work on lattice theoretic duality of information measures. Dual total correlation can be defined as

$$
\begin{aligned}
D T C(X) & =H(X)-\sum_{i=1}^{n} H\left(X_{i} \mid X_{E / i}\right) \\
& =\sum_{i=1}^{n} H\left(X_{E / i}\right)-(n-1) H(X)
\end{aligned}
$$

where $X_{E / i}$ denotes the marginal random variable of cardinality $(n-1)$ that excludes $X_{i}$. Both total correlation and its dual share a number of desirable properties [24][25], some of which are detailed below.

- Entropic: A measure is considered entropic if it is defined as a function of subset entropies $H\left(X_{A}\right)$ for any $X_{A} \subseteq X$. Both total correlation and dual total correlation are linearly entropic since they are given by linear functions $\langle\mathbf{C}, \mathbf{H}\rangle$, where $\mathbf{H}$ is a column vector of all subset entropies, $\mathbf{C}$ a row vector of $2^{n}$ constants, and $\langle\cdot, \cdot\rangle$ the inner product.
- Correlative: A measure is correlative if it equals zero when the marginals $X_{i}, \forall i \in[n]$ are mutually independent.
- Symmetric: A measure is symmetric if it remains unchanged under permutation of the marginals.
- Non-negative: The measure is always $\geq 0$.

One property not shared between the two measures is that total correlation can be represented as the KL-Divergence $D_{K L}\left(p(x) \| \prod_{i=1}^{n} p\left(x_{i}\right)\right)$, whereas dual total correlation has no such divergence expression. Another difference is the set of maximizing distributions for each measure. Previous work has characterized these distributions [23][26], shown in Table IaIb below for three variables with binary support. Maximizers of TC are exactly the distributions for which
a bijection over support $q$ exists between every pair of marginal variables. Such a relationship is often referred to as informational redundancy, making the distribution in Table Ia maximally redundan ${ }^{11}$. The counterpart to redundancy is informational synergy, which refers to the presence of relationships that only exist from a collection $X_{A}(1<|A|<$ $n$ ) to subset $X_{B} \subseteq X_{E / A}$ of the complement set. The parity distribution in Table $I b$ only has dependence when all $n$ variables are considered, mapping every $(n-1)$ variable subset $\left(X_{E / i}, \forall i \in[n]\right)$ to the remaining variable $X_{i}$. Therefore, we can say that Table Ib is maximally synergistic of order $(n-1)$.

TABLE I: Maximizing Distributions for Three Variables
(a) Binary Maximizers TC

| $\boldsymbol{X}_{\mathbf{0}}$ | $\boldsymbol{X}_{\mathbf{1}}$ | $\boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{P r}$ |
| :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | $1 / 2$ |
| 1 | 1 | 1 | $1 / 2$ |


| $\boldsymbol{X}_{\mathbf{0}}$ | $\boldsymbol{X}_{\mathbf{1}}$ | $\boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{P r}$ |
| :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | $1 / 4$ |
| 0 | 1 | 1 | $1 / 4$ |
| 1 | 0 | 1 | $1 / 4$ |
| 1 | 1 | 0 | $1 / 4$ |

## B. Fujishige's Total Correlation(s) and Cohesion

If TC is maximized by information contained in every pair of variables, and DTC by information present only in the $n$ variable joint state, what about information in intermediate collections of variables? Do any measures exist whose maximizers prefer dependence among subsets with cardinality between 2 and $n$ ? A promising place to start in answering this question is a family of measures introduced by Fujishige [28]. The main content of his paper focuses on connecting entropy and polymatroids (introduced in Section III-A, but a small portion is dedicated to applying these insights to extend TC and DTC. His observation was that summing over all marginals of cardinality 1 causes each member of the set to be covered 1 time (or respectively, sums of cardinality $(n-1)$ cover each element $(n-1)$ times for DTC). These cover relations are reflected in the constant multiples of joint entropy appearing in both measures. From this, one can generalize to correlation measures which sum over marginals of cardinality $k$ and cover each element $\binom{n-1}{k-1}$ times. These generalized measures, and their duals, are given by

$$
\begin{gathered}
\mathcal{C}^{(k)}(X)=\sum_{X_{A} \in \mathcal{E}_{k}} H\left(X_{A}\right)-\binom{n-1}{k-1} H(X) \\
\mathcal{C}^{(n-k)}(X)=\sum_{X_{B} \in \mathcal{E}_{n-k}} H\left(X_{B}\right)-\binom{n-1}{n-k-1} H(X) .
\end{gathered}
$$

Here, $k$ (and $(n-k)$ respectively) is called the interaction order, which dictates the cardinality of subsets appearing in the first term

$$
\mathcal{E}_{k}=\{A: A \subseteq X,|A|=k\}
$$

[^1]For $n$ variables, $\mathcal{C}^{(1)}$ is clearly equivalent to the total correlation, and $\mathcal{C}^{(n-1)}$ to the dual total correlation. For those more comfortable with the conditional entropy definition of dual total correlation, each of the measures can be written as

$$
\begin{aligned}
\mathcal{C}^{(k)}(X) & =\sum_{X_{A} \in \mathcal{E}_{k}} H\left(X_{A}\right)-\binom{n-1}{k-1} H(X) \\
& =\sum_{X_{A} \in \mathcal{E}_{k}} H\left(X_{A}\right)+\left(\binom{n-1}{k}-\binom{n}{k}\right) H(X) \\
& =\binom{n-1}{k} H(X)-\sum_{X_{B} \in \mathcal{E}_{n-k}} H\left(X_{B} \mid X_{A}\right)
\end{aligned}
$$

The desirable properties satisfied by $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(n-1)}$, including being linearly entropic, continue to be satisfied for any choice of parameters $n$ and $k$. Because of this, the measures can be related according to the following linear inequalities [28], which we call polymatroid bounds.

$$
\begin{array}{r}
(n-k) \cdot \mathcal{C}^{(k)}(X) \geq k \cdot \mathcal{C}^{(k+1)}(X) \\
(n-k) \cdot \mathcal{C}^{(n-k)}(X) \geq k \cdot \mathcal{C}^{(n-k-1)}(X) \tag{1}
\end{array}
$$

Fujishige referred to each measure as a (dual) total correlation, but we believe such overlapping nomenclature might be confusing. To maintain some link to total correlation without trying to re-brand existing terminology, we refer to the family of measures as Cohesion measures, after one of Watanabe's (and subsequently, Han's) original names for the total correlation [24], [29]. For interaction order $k$, the associated measure is referred to as Cohesion-k, or adopting the shorthand of [30], $\mathcal{C}^{(k)}$.

## C. Summary of main results

The maximizers for Cohesion- $k$ have only been studied for the special cases of $k=1$ and $(n-1)$. This leaves the question open regarding each Cohesion measures' sensitivity to certain orders of dependence. We will show through the main result of the paper (Theorem 3) that each intermediate measure can be upper-bounded by a constant, and that this bound is always achieved for a particular class of distributions with marginal support $q$. The exact value of $q$, if known in general, would provide an answer to the MDS conjecture (discussed further in Section III). Intuitively, the MDS conjecture is a claim about how large $q$ must be ( $q \geq n-1$ ) in order for certain types of linear dependence structures to occur. Despite the uncertainty surrounding $q$, it is still possible to prove that such an integer must exist for arbitrary values of $n$ and $1 \leq k \leq n-1$. The proof relies on a subfield of combinatorics, called matroid theory, to exploit the dependence structure of distributions that achieve the bound with equality. Matroids are objects that generalize the concepts of linear independence and provide a common language for talking about dependence, both in codes and probability distributions.

While matroids are required to make the most general claims, for special cases, such as when the number of variables is a prime power and $q=n$, straightforward methods
for constructing MDS codes are well established. Because these special cases provide concrete examples of distributions that achieve the bound, the details of their construction is worth introducing. Thus, the preliminaries of Section III attempts to balance high level abstractions required for the proof (Section III-A) with motivated examples from coding theory (Section III-B). Before jumping straight into these preliminaries, we wish to solidify intuition for the problem, and further demonstrate why the theoretical approach is necessary. To do this, we look at the simplest Cohesion measure whose maximizers have not been characterized, which is Cohesion- 2 over four random variables.

## D. Cohesion-2 for 4 Variables

A naive but immediately accessible way of exploring the maximizing distributions, given the material presented so far, is empirical evaluation. This might take the form of a simplex search using gradient descent, or even a brute force evaluation of every point on a discretized (alt. uniformly sampled) simplex. Due to the low dimensionality of this special case, we have the luxury of performing the latter without much computational concern. Assuming for now that each marginal variable is binary, and given a sample from the 16 -simplex, we can calculate all three Cohesion measures for four variables and use the results as coordinates to a three-dimensional vector space.

Through the polymatroid upper and lower bounds of Equation 1, we can define a convex polytope over feasible solutions. Normally the region is unbounded, but constraints on the support of each variable lead to three additional inequalities
$\mathcal{C}^{(1)}+\mathcal{C}^{(3)} \leq 4, \quad \mathcal{C}^{(2)}+3 \cdot \mathcal{C}^{(1)} \leq 12, \quad \mathcal{C}^{(2)}+3 \cdot \mathcal{C}^{(3)} \leq 12$.
The proofs for these inequalities appear in Appendix A. Figure 1 shows the two-dimensional projections of this space, alongside the bounds. Given these results, we are interested in two questions: what are the empirically maximal distributions, and do they achieve the upper bound? As a sanity check, we see Cohesion-1 (TC) and Cohesion-3 (DTC) achieve their maximum of 3 bits for the four variable versions of distributions in Table 1, which meet the bounds with equality. Cohesion-2, on the other hand, peaks for the distribution in Table $\Pi$, which has been called a redundant-synergy distribution [31], owing to the combination of third-order (first three variables) and second-order (first and last variable) information. This answers the first question, but when comparing the value of this distribution ( 5 bits) with the upper bound ( 6 bits), we observe a large gap in the feasible region. The appearance of this gap unfortunately means that there is no verification this distribution is actually maximal. Perhaps the sampling over the simplex was too course, or the bound is too loose and additional (possibly nonlinear) inequalities are required. Suppose we are fortunate in not having to check these alternatives due to our access to an oracle $\mathcal{O}(\mathcal{C}, n)$. This oracle takes as input our Cohesion measure $\mathcal{C}$, some


Fig. 1: Plots of each Cohesion measure for four binary variables, alongside the polymatriod bounds (orange lines). The color gradient represents the value of the third measure, with blue for low values, and yellow for high. The dashed red lines correspond to the constant upper bounds, introduced in Section IV

TABLE II: Binary Maximizers for Cohesion-2

| $\boldsymbol{X}_{\mathbf{0}}$ | $\boldsymbol{X}_{\mathbf{1}}$ | $\boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{X}_{\mathbf{3}}$ | $\boldsymbol{P r}$ |
| :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 0 | $1 / 4$ |
| 0 | 1 | 1 | 0 | $1 / 4$ |
| 1 | 0 | 1 | 1 | $1 / 4$ |
| 1 | 1 | 0 | 1 | $1 / 4$ |

integer number of variables $n$, and returns a distribution of cardinality $n$ in the set of global maximizers for $\mathcal{C}$. On entering Cohesion-2 and $n=4$, the following distribution is revealed (Table III).

TABLE III: Reed-Solomon Maximizers for Cohesion-2

| $\boldsymbol{X}_{\mathbf{0}}$ | $\boldsymbol{X}_{\mathbf{1}}$ | $\boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{X}_{\mathbf{3}}$ | $\boldsymbol{P r}$ | $\boldsymbol{X}_{\mathbf{0}}$ | $\boldsymbol{X}_{\mathbf{1}}$ | $\boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{X}_{\mathbf{3}}$ | $\boldsymbol{P r}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 0 | $1 / 16$ | 2 | 2 | 2 | 2 | $1 / 16$ |
| 0 | 1 | 2 | 3 | $1 / 16$ | 2 | 3 | 0 | 1 | $1 / 16$ |
| 0 | 2 | 3 | 1 | $1 / 16$ | 2 | 0 | 1 | 3 | $1 / 16$ |
| 0 | 3 | 1 | 2 | $1 / 16$ | 2 | 1 | 3 | 0 | $1 / 16$ |
| 1 | 1 | 1 | 1 | $1 / 16$ | 3 | 3 | 3 | 3 | $1 / 16$ |
| 1 | 0 | 3 | 2 | $1 / 16$ | 3 | 2 | 1 | 0 | $1 / 16$ |
| 1 | 3 | 2 | 0 | $1 / 16$ | 3 | 1 | 0 | 2 | $1 / 16$ |
| 1 | 2 | 0 | 3 | $1 / 16$ | 3 | 0 | 2 | 1 | $1 / 16$ |

Calculating Cohesion-2 for this distribution yields a value of 6 quaternary digits ( 12 bits), meeting the upper bound with equality. Our oracle has verified that the polymatroid bounds are indeed enough to fully characterize Cohesion, at least in this special case. However, we've only replaced one mystery with another: how does the oracle work, and does it generalize to arbitrary Cohesion measures? It turns out that each row in Table III is related to an important concept in algebraic coding theory called a Reed-Solomon code. Reed-Solomon codes have the property of being maximum distance separable (MDS), and provide a way of constructing these codes when the support is large $(q \geq n)$ and $n$ is a prime power. Notice that our restriction to binary support in Figure 1 played an important role in the existence of a gap. The dangers of empirical evaluation should now be apparent if they weren't before. Even if the criteria on the number of variables and size of the support are met, a search over the $n^{n}$ simplex would be required to find a Reed-Solomon distribution. As we will later show, if the MDS conjecture
holds true, for any value of $n$ and $k, q$ is expected to be at least $n-1$, meaning superexponential search complexity is likely unavoidable. Thankfully, since we know which vertex of the Cohesion polytope a maximizer would appear if one existed, we can abandon search in favor of investigating the structural properties at this point. The next section is dedicated to preliminaries on matroid and coding theory, including illustrative examples whenever possible.

## III. PRELIMINARIES

## A. Matroids and Polymatroids

A matroid $M$ is given by a tuple $(E, \mathcal{I})$, where $E$ is some finite set, and $\mathcal{I}$ is a collection of subsets of $E$ satisfying the following three conditions:
(M1): $\emptyset \in \mathcal{I}$
(M2): If $I \in \mathcal{I}$ and $I^{\prime} \subseteq I$, then $I^{\prime} \in \mathcal{I}$
(M3): If $I_{1}$ and $I_{2}$ are in $\mathcal{I}$ and $\left|I_{1}\right|<\left|I_{2}\right|$, then there is an element $e$ of $I_{2}-I_{1}$ such that $I_{1} \cup e \in \mathcal{I}$
$E$ and $\mathcal{I}$ are referred to as the ground set and the independent set of $M$, respectively. As an example, consider an $m \times n$ matrix $A$. Define the ground set $E(A)$ to be the set of all column labels $\{i \in[n]\}$ indexing $A$. The independent set $\mathcal{I}(A)$ consists of all collections of labels that index linearly independent columns. For the matrix $A$ given below, the independent set would be $\mathcal{I}(A)=$ $\{\emptyset,\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\}\}$.

$$
\left.A=\begin{array}{ccc}
\mathbf{1} & \mathbf{2} & \mathbf{3} \\
1 & 0 & 1 \\
0 & 1 & 1
\end{array}\right]
$$

Any matroid that can be used to describe the dependence structure of a matrix defined over some finite field $\mathbb{F}_{q}$ (as with $A$ above for $q=2$ ) is called a vector matroid. The last axiom (M3) is called the exchange axiom, and is the least intuitive of the three. To get a better sense for it, we include a proof that it is satisfied for the independent set of a vector matroid.

Proof: Let $I_{1}$ and $I_{2}$ be two members of the independent set $\mathcal{I}$ for a vector matroid $M(A)$, with $\left|I_{1}\right|<\left|I_{2}\right|$. This means both $I_{1}$ and $I_{2}$ index collections of linearly independent columns. The subspace $\mathcal{W}$ spanned by $I_{1} \cup I_{2}$
has dimension at least $\left|I_{2}\right|$ (which occurs if $I_{1} \subset I_{2}$ ). Assume for the sake of contradiction that $I_{1} \cup e$ is linearly dependent for all $e \in I_{2}-I_{1}$. Subtraction here is set subtraction, making $I_{2}-I_{1}$ the set of all elements contained in $I_{2}$, but not in $I_{1}$. By this assumption, $\mathcal{W}$ is contained in the span of $I_{1}$, implying $\left|I_{2}\right| \leq \operatorname{dim\mathcal {W}} \leq\left|I_{1}\right|<\left|I_{2}\right|$; contradiction.

For two matroids $M_{1}$ and $M_{2}$, denote the corresponding ground sets by $E\left(M_{1}\right), E\left(M_{2}\right)$, and independent sets by $\mathcal{I}\left(M_{1}\right), \mathcal{I}\left(M_{2}\right)$. These matroids are said to be isomorphic ( $M_{1} \cong M_{2}$ ) if there exists a bijection $\psi$ satisfying

$$
\begin{gathered}
\psi: E\left(M_{1}\right) \rightarrow E\left(M_{2}\right), \quad \forall X \subseteq E\left(M_{1}\right) \\
\psi(X) \in \mathcal{I}\left(M_{2}\right) \Longleftrightarrow X \in \mathcal{I}\left(M_{1}\right)
\end{gathered}
$$

If a matroid $M_{1}$ is isomorphic to a vector matroid $M_{2}(A)$, where $A$ is a matrix over the field $\mathbb{F}_{q}$, then $M_{1}$ is said to be $\mathbf{F}_{\mathbf{q}}$-representable. It's important to note that not every matroid is $\mathbb{F}_{q}$-representable for an arbitrary field. To illustrate this, let $\mathbb{F}_{2}$ be the Galois field of two elements, where element $x \in\{0,1\}$, and addition/multiplication are given by mod-2 arithmetic. Construct a matroid $M$ over $E=$ $\{1,2,3,4\}$ such that all subsets of cardinality two or less are contained in the independent set; namely,

$$
\mathcal{I}=\{S \subseteq E: 0 \leq|S| \leq 2\}
$$

Any matroid defined over a ground set of cardinality $n$, whose independent set contains all subsets of cardinality $k$ or less is called a uniform matroid $U_{k, n}$. Here, $M$ corresponds to the uniform matroid $U_{2,4}$.

Proposition 1 (Oxley): The uniform matroid $U_{2,4}$ is not $\mathbb{F}_{2}$-representable

Proof: Assume that $U_{2,4}$ is $\mathbb{F}_{2}$-representable, implying that $U_{2,4} \cong M(A)$ for some matrix $A$ defined over $\mathbb{F}_{2}$. Since the largest element in $\mathcal{I}\left(U_{2,4}\right)$ has cardinality two, the column space of $A$ must have dimension two. A twodimensional vector space over $\mathbb{F}_{2}$ has exactly four members, but only three of them are non-zero. This means that $A$ cannot have four distinct non-zero columns, so a set of two columns in $A$ must be linearly dependent. However, this contradicts the original claim, since every pair of columns must be linearly independent in order for $U_{2,4} \cong M(A)$.

A polymatroid $\mathcal{P}$ is given by a tuple $(E, f)$ of ground set $E$ and a submodular set function $f$. A submodular set function is a mapping from the power set of $E$ to the non-negative real numbers $\left(f: 2^{E} \rightarrow \mathbb{R}^{+}\right)$satisfying the following three conditions:

```
(non-negativity): }f(\emptyset)=0,f(S)\geq0\quad\forallS\subseteq
(monotonicity): }\quadf(S)\leqf(T)\LongleftrightarrowS\subseteqT\subseteq
(submodularity):}f(S\cupT)+f(S\capT)\leqf(S)+f(T
```

Fujishige [28] was the first to show that Shannon entropy acts as a submodular set function $\mathcal{H}$ for the polymatroid $(E, \mathcal{H})$ over ground set $E$ of random variables $\left\{X_{1}, X_{2}, \ldots, X_{n}\right\}$. Polymatroids are generalizations of matroids in that, if we
restrict the codomain of the set function to the non-negative integers $\left(f: 2^{E} \rightarrow \mathbb{Z}^{+}\right)$and require the additional upper bound condition $f(S) \leq|S| \quad \forall S \subseteq E$, then $f$ is called a rank function, and can be used to build a matroid $M$.

Lemma 1 (Oxley): Let $E$ be a set and $f$ a function on $2^{E}$ satisfying the above rank function conditions. If $X$ and $Y$ are subsets of $E$ such that, for all $y \in(Y-X), f(X \cup y)=$ $f(X)$, then $f(X \cup Y)=f(X)$

Proof: See [32], the argument follows by induction on the number of elements in $(Y-X)$.

Theorem 1 (Oxley): Let $E$ be a set and $f$ a function on $2^{E}$ satisfying the above rank function conditions. Let $\mathcal{I}$ be the collection of subsets $S$ of $E$ for which $f(S)=|S|$. Then $(E, \mathcal{I})$ is a matroid having rank function $f$.

Proof: We will show that the independent set $\mathcal{I}$, constructed in the manner above satisfies the matroid axioms (M1)-(M3).

- By the non-negativity and upper bound conditions on $f, 0 \leq f(\emptyset) \leq|\emptyset|=0$, so $f(\emptyset)=|\emptyset|$ and $\emptyset \in \mathcal{I}$, satisfying (M1).
- If $I \in \mathcal{I}$ then $f(I)=|I|$. For $I^{\prime} \subseteq I$, and by submodularity of $f, f(I)+f(\emptyset) \leq f\left(I^{\prime}\right)+f\left(I-I^{\prime}\right)$. Since each term on the right hand side is upper-bounded, we can simplify to $|I| \leq\left|I^{\prime}\right|+\left|I-I^{\prime}\right|=|I|$. Equality must hold throughout, implying $f\left(I^{\prime}\right)=\left|I^{\prime}\right|$ and $I^{\prime} \in \mathcal{I}$, satisfying (M2).
- Assume for the sake of contradiction $I_{1}$ and $I_{2}$ are in $\mathcal{I}$ with $\left|I_{1}\right|<\left|I_{2}\right|$, but for all $e \in I_{2}-I_{1}, f\left(I_{1} \cup e\right) \neq$ $\left|I_{1} \cup e\right|$. We have $\left|I_{1}\right|+1>f\left(I_{1} \cup e\right) \geq f\left(I_{1}\right)=\left|I_{1}\right|$, meaning $f\left(I_{1} \cup e\right)=\left|I_{1}\right|$. Now, by applying Lemma 1 with $X=I_{1}$ and $Y=I_{2}$, it follows that $f\left(I_{1}\right)=$ $f\left(I_{1} \cup I_{2}\right)$. But $f\left(I_{1} \cup I_{2}\right) \geq f\left(I_{2}\right)=\left|I_{2}\right|$, so $\left|I_{1}\right| \geq\left|I_{2}\right|$; contradiction, meaning (M3) is satisfied.

With these basic definitions and theorems established, we move on to our discussion of coding theory.

## B. Algebraic Coding Theory and MDS Codes

The fundamental object of coding theory is the message, which in the discrete case considered here is simply a string of length $k$ represented over some alphabet $\Sigma$ of cardinality $q$. Different subfields of coding theory are often interested in transformations of messages into codewords (strings of length $n \geq k$ ), with the utility of these codewords ranging from error-correction to cryptographic secret sharing [33]. For each mapping from message to codeword, the message length $k$, codeword length $n$, alphabet size $q$, and minimum distance $d$, are important parameters. The minimum distance is given by $d:=\min _{c_{i}, c_{j} \in \mathcal{C}} \Delta\left(c_{i}, c_{j}\right)$, where $\Delta(\cdot)$ is the Hamming Distance, and $c_{i} c_{j}$ are any two codewords in $\mathcal{C} \subseteq q^{n}$. This minimum distance is used as a measure of error-correction and detection capabilities of a code. In particular, if the minimum distance between any two codewords is $k+1$, the coding scheme can be used to detect
$k$ errors, while a minimum distance of $2 k+1$ allows for the correction of $k$ errors [17].

For any coding scheme with minimum distance $d$, we can upper bound the number of codewords $\mathcal{M}$ in the following way. For the codewords $c_{1}, \ldots, c_{\mathcal{M}}$ of an $(n, k, d)_{q}$ code $\mathcal{C}$, let $\bar{c}_{i}$ be the prefix of $c_{i} \in \mathcal{C}$ of length $n-d+1$. Each $\bar{c}_{i}$ must be distinct $\left(\bar{c}_{i} \neq \bar{c}_{j}\right)$, otherwise $\Delta\left(c_{i}, c_{j}\right) \leq d-1$, violating the claim that $\mathcal{C}$ has minimum distance $d$. Thus, the number of these prefixes bounds $\mathcal{M} \leq q^{n-d+1}$, which is known as the Singleton bound. For linear codes, which are any $\mathcal{C} \subseteq q^{n}$ such that each codeword $c_{i} \in \mathcal{C}$ is generated by a linear combination of codewords $c_{j} \in \mathcal{C}\left(c_{i} \neq c_{j}\right), \mathcal{M}$ is equal to $q^{k}$. Plugging this into the Singleton bound, we get $d \leq n-k+1$. Codes that meet this bound with equality are termed maximum distance separable (MDS), and are an important class of codes with optimal error-correction/detection capabilities (for large alphabets). If we treat each $c_{i}$ as a $1 \times n$ row vector, MDS codes are represented by a basis $c_{1}, \ldots, c_{k}$ of $k$ vectors, whose span produces all $q^{k}$ codewords. This $k \times n$ collection of basis vectors is called the generator matrix $\mathcal{A}_{k, n}$ of $\mathcal{C}$. An interesting property of $\mathcal{A}_{k, n}$ follows from the constraints imposed by the minimum distance $d=n-k+1$; each set of $k$ (and by extension, $\leq k$ ) columns are linearly independent. This fact falls directly from the prefix argument introduced above.

Recalling examples from the matroid preliminaries, linearly independent columns can be used to define the independent set of a vector matroid. In this case the matroid produced by $\mathcal{A}_{k, n}$ is isomorphic to the uniform matroid $U_{k, n}$. This will be the crucial fact allowing us to connect codes and probability distributions, but we must first discuss the alphabet size $q$ (or equivalently, the field $\mathbb{F}_{q}$ of $\mathcal{A}_{k, n}$ ) necessary for a code to be MDS. The MDS conjecture postulates that MDS codes only exist when $q \geq n-1$, except when $k=3$ or $k=q-1$ and $q=2^{m} m \in \mathbb{Z}^{>0}$, in which case $q \geq n-2$ [19][34]. While such conditions seem arbitrary at first glance, they stem from a meaningful equivalence with an object in projective geometry called a $k$-arc. Most progress on the MDS conjecture, such as the recent positive result for prime fields [35], proceeds by arguments and new results on these geometric objects.

As previously mentioned, when $q$ is large, strategies for constructing certain MDS codes are known, with the most well established corresponding to Reed-Solomon (RS) codes [36]. In explaining what Reed-Solomon codes are, and how they are generated, we will only consider the classical case when $q=n$. Note that this limits the possible values of $n$ significantly, since finite fields $\mathbb{F}_{q}$ only exist when $q=p^{m}$ is a prime power. In these cases, the elements of $\mathbb{F}_{p^{m}}$ are the $p^{m}$ roots of the polynomial $x^{\left(p^{m}\right)}-x$ [37]. In the simplest case when $q$ is a prime number $(m=1)$, the elements of $\mathbb{F}_{p}$ are just $\mathbb{Z}_{p}$, the integers $\bmod p$. Given values for $k$, and
$q=n$, start by defining a polynomial

$$
f(z)=\sum_{j=0}^{k-1} f_{j} z^{j} \quad \forall_{j} f_{j} \in \mathbb{F}_{q}
$$

of degree less than $k$ for some indeterminate $z$. Since the coefficients range over all $k$-tuples in $\left(\mathbb{F}_{q}\right)^{k}, f(z)$ is one of $q^{k}$ polynomials of this form. Next, define a $q$-tuple $\mathcal{B}=\left(\beta_{1}, \beta_{2}, \ldots, \beta_{q}\right)$ such that $\beta_{1}=0, \beta_{2}=1, \beta_{3}=\alpha, \ldots, \beta_{q}=\alpha^{q-2}$, for a primitive element $\alpha$ over $\mathbb{F}_{q}$. A primitive element is any element of the field that forms a multiplicative cyclic group, meaning raising $\alpha$ to some power generates all other elements (except 0 ). This implies $\mathcal{B}$ is just an ordered list of distinct elements in $\mathbb{F}_{q}$ (ex: a permutation of $[q]-1$ for $q$ a prime number). The idea behind introducing $\mathcal{B}$ is that we wish to construct another polynomial of degree less than $n$ by evaluating the polynomials $f(z)$ at $n$ points. This procedure is referred to as a valuation map from $\left(\mathbb{F}_{q}\right)^{k}$ to $\left(\mathbb{F}_{q}\right)^{n}$. By evaluating $f(z)$ at each element in $\mathcal{B},\left(f\left(\beta_{1}\right), \ldots, f\left(\beta_{q}\right)\right)$, with each $f\left(\beta_{i}\right)=\sum_{j=0}^{k-1} f_{j} \beta_{i}^{j}$, we can produce the $k$ basis vectors for our generator matrix $\mathcal{A}_{k, n}$. These vectors form a Vandermonde matrix, where each row is $\mathcal{B}$ raised to some power.

$$
\mathcal{A}_{k, n}=\left[\begin{array}{ccccc}
1 & 1 & 1 & \cdots & 1 \\
0 & 1 & \alpha & \cdots & \alpha^{-1} \\
0 & 1 & \alpha^{2} & \ldots & \alpha^{-2} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & \alpha^{k-1} & \ldots & \alpha^{-(k-1)}
\end{array}\right]
$$

Here the negative exponents are the same as a reverse indexing of $\mathcal{B}-\{0\}$ (ex: $\alpha^{-1}=\alpha^{q-2}$ ). To build intuition as to why/how this procedure works, let's look at the ReedSolomon code underlying the distribution in Table III In this case, $n=4$ with field $\mathbb{F}_{4}=\mathbb{F}_{2^{2}}$. The elements of $\mathbb{F}_{4}$ (the roots of $x^{4}-x$ ) are $\{0,1, z, z+1\}$ for indeterminate $z$. Addition and multiplication are given by the following tables, where $z^{2}=z+1$.
(a) Addition over $\mathbb{F}_{4}$

| $\oplus$ | $\mathbf{0}$ | $\mathbf{1}$ | $\boldsymbol{z}$ | $\boldsymbol{z}^{\mathbf{2}}$ |
| :---: | :---: | :---: | :---: | :---: |
| $\mathbf{0}$ | 0 | 1 | $z$ | $z^{2}$ |
| $\mathbf{1}$ | 1 | 0 | $z^{2}$ | $z$ |
| $\boldsymbol{z}$ | $z$ | $z^{2}$ | 0 | 1 |
| $\boldsymbol{z}^{\mathbf{2}}$ | $z^{2}$ | $z$ | 1 | 0 |

(b) Multiplication over $\mathbb{F}_{4}$

| $\otimes$ | $\mathbf{0}$ | $\mathbf{1}$ | $\boldsymbol{z}$ | $\boldsymbol{z}^{\mathbf{2}}$ |
| :---: | :---: | :---: | :---: | :---: |
| $\mathbf{0}$ | 0 | 0 | 0 | 0 |
| $\mathbf{1}$ | 0 | 1 | $z$ | $z^{2}$ |
| $\boldsymbol{z}$ | 0 | $z$ | $z^{2}$ | 1 |
| $\boldsymbol{z}^{\mathbf{2}}$ | 0 | $z^{2}$ | 1 | $z$ |

Since 4 is a prime power, multiplication over the field obeys $\bmod -g(z)$ arithmetic, for a prime polynomial $g(z)$ of degree $m=2$ (in particular, $g(z)=z^{2}+z+1$ ). We have chosen not to include a detailed background on nonprime finite fields (which can be found here [37]), but only knowledge of the addition/multiplication tables is required to move forward. For $k=2$, the polynomials $f(z)$ of the valuation map take the form $f_{0}+f_{1} z,\left\{f_{0}, f_{1}\right\} \in \mathbb{F}_{4}$, and $\mathcal{B}=\{0,1, z, z+1\}$. Evaluating $f(z)$ at each $\beta \in \mathcal{B}$ results in the vector $\left(f_{0}, f_{0}+f_{1}, f_{0}+f_{1} z, f_{0}+f_{1}(z+1)\right)$. Choosing
the standard basis, i.e $f(z)=1$ and $f(z)=z$, the vector simplifies to $(1,1,1,1)$ and $(0,1, z, z+1)$ respectively. The span of these produces the following 16 row vectors.

TABLE V: Reed-Solomon codes over $\mathbb{F}_{4}$

$$
\left|\begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 1 & z & z+1 \\
0 & z & z+1 & 1 \\
0 & z+1 & 1 & z \\
1 & 1 & 1 & 1 \\
1 & 0 & z+1 & z \\
1 & z+1 & z & 0 \\
1 & z & 0 & z+1
\end{array}\right|\left|\begin{array}{cccc}
z & z & z & z \\
z & z+1 & 0 & 1 \\
z & 0 & 1 & z+1 \\
z & 1 & z+1 & 0 \\
z+1 & z+1 & z+1 & z+1 \\
z+1 & z & 1 & 0 \\
z+1 & 1 & 0 & z \\
z+1 & 0 & z & 1
\end{array}\right|
$$

Since all fields with $p^{m}$ elements are isomorphic to $\mathbb{F}_{p^{m}}$ up to relabeling of the elements, we can map $z$ to 2 and $z+1$ to 3. Doing so produces the rows of Table III

## IV. MAXIMIZING DISTRIBUTIONS

To recapitulate the last two sections, we introduced matroids and demonstrated their connection to MDS codes (through the generator matrix $\mathcal{A}_{k, n}$ ) and to entropy (through polymatroids). What's left to show is that the maximizers of Cohesion-k have a matroidal structure isomorphic to that of MDS codes, and that $\mathbb{F}_{q}$-representability of an MDS code implies achievability of the polymatroid bound for distributions with support $q$. To prove the first claim, we introduce a constant upper bound on each Cohesion measure that meets the polymatroid bound at a single point. Given the standard expression for Cohesion- $k$, namely

$$
\mathcal{C}^{(k)}(X)=\sum_{X_{A} \in \mathcal{E}_{k}} H\left(X_{A}\right)-\binom{n-1}{k-1} H(X),
$$

we can upper bound each term in the sum, and lower bound the joint entropy. For any $S \subseteq X$, the entropy $H(S)$ is maximal when $S$ is uniform over its support. Since each $X_{i}$ has support $q, S=\left\{X_{1}, \ldots, X_{m}\right\},\|S\| \leq q^{m}$, it follows that $H(S) \leq \log _{q}\left(q^{m}\right)=m$, the cardinality of $S$. Thus, each term in the sum of Cohesion- $k$ is at most $k$. The joint entropy can be lower bounded by noticing $X_{A} \subseteq X$ and $H\left(X_{A}\right) \leq H(X)$ by monotonicity, implying

$$
H(X) \geq \max _{X_{A} \in \mathcal{E}_{k}} H\left(X_{A}\right)=k
$$

Substituting these results into the equation, we get the following constant bound

$$
\mathcal{C}^{(k)}(X) \leq k\binom{n}{k}-k\binom{n-1}{k-1}=k\binom{n-1}{k}
$$

Lemma 2: The entropy function at the constant bound maps the power set of random variables in $X$ to the nonnegative integers $\left(\mathcal{H}: 2^{X} \rightarrow \mathbb{Z}^{+}\right)$

Proof: Split up the elements of the power set into different cases according to the value of k . Consider some subset $S \subseteq X$ such that the cardinality of $S$ is

- $\mathbf{0}: H(\emptyset)=0$, which is integer valued.
- $\mathbf{j}, \mathbf{1} \leq \mathbf{j} \leq \mathbf{k}$ : Since the upper bound requires all subsets $X_{A}$ of cardinality $k$ to have $H\left(X_{A}\right)=k$, every collection of $k$ random variables is independent and
uniform over $q^{k}$, implying that every collection of $j$ variables between 1 and $k$ is uniform over $q^{j}$. Thus, $H(S)=j$, which is integer valued.
- $\mathbf{l}, \mathbf{k} \leq \mathbf{l} \leq \mathbf{n}$ : Both the joint entropy $H(X)$ of cardinality $n$ and all subset entropies $H\left(X_{A}\right)$ of cardinality $k$ are equal to $k$. By monotonicity, every element of the power set with cardinality $l$ between $k$ and $n$ must also have $H(S)=k$, which is integer valued.

By Lemma 2, the upper bound on subset entropies, and the fact that entropy is a submodular set function, entropy satisfies the requirements of a rank function for any distribution meeting the constant bound with equality. By Theorem 1, we can construct the independent set of a matroid $\mathcal{C}_{k, n}$ by looking at all subsets $S$ of $X$ where $H(S)=|S|$. When the bound is met, the independent set consists of all subsets of $X$ with cardinality $\leq k$, making $\mathcal{C}_{k, n} \cong U_{k, n}$ (and consequently $\mathcal{C}_{k, n} \cong M\left(\mathcal{A}_{k, n}\right)$ for the vector matroid of an MDS generator matrix). This takes care of the first claim, but says nothing about the form of the distributions meeting these criteria. To prove the second claim, we must explore the connection between independence in vector and probability spaces.

Theorem 2 (Matúśs): If a matroid $M$ is $\mathbb{F}_{q}$-representable, then $M$ also describes the statistical independence relationships for a distribution with marginal support $q$ [38].

Proof: For a vector matroid $M(A)$ with rank function $f$, recall that $0 \leq f(I) \leq|I|, \quad I \subseteq E$, where $E$ is a collection of column labels for $A$. The dual space (collection of rows) for any $I$ columns must also have rank $f(I)$. The linear space spanned by these rows over $\mathbb{F}_{q}$ has $q^{f(I)}$ points. If we take each of these coordinate vectors and assign them a probability of $q^{-f(I)}$ uniformly, the entropy of this collection will be $H\left(X_{I}\right)=f(I) \cdot \log _{q} q=f(I), \forall I \subseteq[n]$. Since under these conditions the entropy is a rank function, we have a one-to-one correspondence between the rank of a vector space and a probability space. Since the ground sets are also the same, this implies $M(A)$ is isomorphic to the matroid describing the resulting distribution.

By Theorem 2, the maximizing distributions for Cohesion-k, whose matroid structure is the same as $M\left(\mathcal{A}_{k, n}\right)$, can be generated by $\mathcal{A}_{k, n}$ (as evidenced by Table III and the Reed-Solomon code from the Section III-B example). This result is almost enough to conclude the form of Cohesion maximizers in full generality, but one final piece is needed. Construction via Reed-Solomon code generalizes to any $n$ and $k$, as long as $n=p^{m}$, but what about $n \neq p^{m}$ ?

Theorem 3: There exists an integer $q$ such that a distribution with marginal support $q$ achieves the upper bound on Cohesion- $k$ for $n$ variables.

Proof: Since the matroid structures are the same, it suffices to show that $U_{k, n}$ is $\mathbb{F}_{q}$-representable for some integer $q$. A family of subsets of $X$ is a finite sequence $\left(A_{1}, A_{2}, \ldots, A_{m}\right)$ such that each member $A_{j}, j \in$
$\{1,2, \ldots, m\}, A_{j} \subseteq X$ need not be distinct. A transversal is a subset $\left\{e_{1}, e_{2}, \ldots, e_{m}\right\}$ of $X$ such that each $e_{j} \in A_{j}, \forall j \in J$. If $S \subseteq X$, then $S$ is a partial transversal if for some subset $K$ of $J, S$ is a transversal of $\left(A_{j}: j \in K\right)$. For a family $\mathcal{A}$ of subsets, let $\mathcal{I}$ be the set of all partial transversals of $\mathcal{A}$. Then $\mathcal{I}$ defines the independent set of a transversal matroid on $X$. The class of uniform matroids is a subclass of transversal matroids (ex: family $\mathcal{A}=\left(A_{1}, A_{2}, \ldots, A_{k}\right)$ such that each $A_{i}=[n]$ defines a transversal matroid that is isomorphic to $U_{k, n}$ ). By Corollary 12.2 .17 in [32], for all transversal matroids $M$, there exists an integer $n(M)$ such that $M$ is representable over every extension field of $\mathbb{F}$ having at least $n(M)$ elements.

This concludes the proof in full generality, and solidifies the connection between matroids, MDS codes, and maximizers of Cohesion. Having accomplished what we set out to do by establishing this connection, what's left is to show from a broader perspective where these findings may be applicable.

## V. RELATED WORK

With the intention of quantifying emergence in complex systems, or how the whole deviates from the sum of its parts, Ay et al. [30] developed an approach based on the principles of information geometry. In this framework, the parts of a system are represented by the parameters of a hierarchical model $\mathcal{E}_{k}$. For a specified value of $k$, each parameter corresponds to a $k$ th order marginal of $p$. The deviation from this hierarchical model is given by the KLDivergence $D_{K L}\left(p \| \mathcal{E}_{k}\right)$. Due to the closure property of hierarchical models, this deviation is equivalent to $D_{K L}\left(p \| p^{(k)}\right)$, where $p^{(k)}$ is the maximum entropy projection for $\mathcal{E}_{k}$. This distribution is restricted to have the same $k$ th order marginals as $p$. Each of these divergence measures $D_{K L}\left(p \| p^{(k)}\right)$ was shown to be upper-bounded by a normalized version of Cohesion-k

$$
\begin{equation*}
D_{K L}\left(p \| p^{(k)}\right) \leq \frac{1}{\binom{n-1}{k-1}} \mathcal{C}^{(k)} \tag{2}
\end{equation*}
$$

This bound is loose in general, as evidenced by results on a tighter upper bound:

$$
D\left(p \| p^{(k)}\right) \leq H_{p}(k)+(N-k) h_{p}(k)-H_{p}(N)
$$

For a better understanding of the notation used above (what $H_{p}(\cdot)$ and $h_{p}(\cdot)$ represent), see section 3 of [30]. Other than these bounds, no results on the global maximizers of $D\left(p \| p^{(k)}\right)$ were given for general $k$ until a recent paper by Matus [39]. An upper bound on the divergence to hierarchical log-linear models was introduced for any family of subsets $\mathcal{A}$ over $X$, with an analysis of its tightness and achievability. When $\mathcal{A}$ is the family of all cardinality $k$ subsets, this divergence is the same as $D\left(p \| p^{(k)}\right)$. This special case was studied as an example of achievability, and a subset of the global maximizers were arrived at by matroid-based arguments. These distributions, referred to as partition representable distributions by Matus [40], are
exactly the global maximizers of Cohesion-k. These findings imply that the upper bound of Equation 2 is tight for the extrema of both measures over a sufficiently large field, but what about more generally? By the nature of $D\left(p \| p^{(k)}\right)$ as a KL-Divergence, any distribution that can be represented solely by its $k$ th order marginals receives zero weight, meaning no lower-order information ( $\leq k$ ) contributes to the measure. For example, the maximizers of $D\left(p \| p^{(1)}\right)$ (the maximally redundant distributions of Table Ia) would be in the set of global minimizers for $D\left(p \| p^{(2)}\right)$. This is not the case for Cohesion-k, which only assigns zero weight when all variables are independent. As a result, Cohesion has preference for linear dependence structures, since $(k-1)$ th order interactions receive almost as much weight as $k$ th order. The gap in the bound between each measure can thus be accounted for by their different treatment of lower order information. To illustrate these differences further, we compare Cohesion- $k$ and $D\left(p \| p^{(k)}\right)$ empirically for four binary variables.

Directing our attention to the second plot in Figure 2, recall that the maximizing distributions for Cohesion-2 over four binary variables was given in Table III By the result of Matus, we know that over a sufficiently large field, $D\left(p \| p^{(2)}\right)$ and Cohesion-2 have the same global maximizers, but what about when the support is restricted? When optimizing over a discretized simplex of bin size $1 / 12$, the maximizing distribution for $D\left(p \| p^{(2)}\right)$ over four binary variables is

TABLE VI: Binary (local) Maximizers for $D\left(p \| p^{(2)}\right)$

| $\boldsymbol{X}_{\mathbf{0}}$ | $\boldsymbol{X}_{\mathbf{1}}$ | $\boldsymbol{X}_{\mathbf{2}}$ | $\boldsymbol{X}_{\mathbf{3}}$ | $\boldsymbol{P r}$ |
| :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 0 | $1 / 4$ |
| 0 | 1 | 1 | 1 | $1 / 4$ |
| 1 | 0 | 1 | 1 | $1 / 6$ |
| 1 | 1 | 0 | 1 | $1 / 6$ |
| 1 | 1 | 1 | 0 | $1 / 6$ |

While this might not be the global maximizer, we can still gain insight from comparing it to maximizers of Cohesion-2. This distribution does not have a linear dependence structure, since no marginal variables have entropy of 1 bit. It does, however, have more third order information (1.44 bits) than the normalized redundant-synergy maximizer of Cohesion2 (1 bit). When the support is large, it is possible to have both linear dependence and maximal information of a particular order (Table $\Pi I)$, which is the main reason why these measures converge at the extremes.

## VI. DISCUSSION AND FUTURE WORK

In addition to its appearance in this work, coding theory has been referenced in a few recent papers on MMI and information decomposition, hinting that a meaningful connection may exist between the two fields. Secret sharing, as alluded to in Section III-B is a sub-discipline of coding theory and cryptography concerned with the secure transmission of information. Each message is transformed into a codeword such that no information about the message


Fig. 2: Plots of normalized Cohesion-k vs $D\left(p \| p^{(k)}\right)$ (labeled IGC-k along the y-axis) for four binary variables. The orange line represents the upper bound of Equation 2. Each point appearing below this line represents lower order dependence that was accounted for by Cohesion, but not the divergence.
or its encoding process can be recovered by an eavesdropper via repeated observation (and under certain computational assumptions). Ideas from secret sharing have been used to argue for an expanded definition of local positivity in the partial information decomposition [41], and to define new measures of MMI-based capacity bounds in secret key agreement [42]. MDS codes, under the secret sharing interpretation, correspond to perfect/ ideal secret sharing schemes where the share is the same size as the secret and no collection of $(k-1)$ users can recover it (assuming the secret is size $k$ ). Although this is the first time MDS codes and perfect/ideal secret sharing schemes have been affiliated with Cohesion measures, it is not the first time they have appeared in the study of MMI.

Since many potential real world applications of these measures will have data with fixed support much smaller than $q=n$, it will be interesting to further study the maximizers of Cohesion over small/ restricted fields. From the limited experimental evidence we gathered in comparing to $D\left(p \| p^{(k)}\right)$ (and some additional simplex search), it seems that Cohesion maximizers are always related to linear dependence structures over $\mathbb{F}_{q}$, with the matroid structure as close to uniform as allowable over the fixed field. Such structures seem conceptually close to near-MDS (NMDS) and almostMDS (AMDS) codes over small fields [43][44].

## APPENDIX

## Proofs for additional polymatroid bounds

Each proof has similar construction to the constant upper bound of Section IV Each term in the sum is upper-bounded by its cardinality, and the joint entropy is lower bounded through monotonicity.

$$
\begin{aligned}
\mathcal{C}^{(1)}+\mathcal{C}^{(3)} & =\sum_{i=1}^{n} H\left(X_{i}\right)+\sum_{i=1}^{n} H\left(X_{E / i}\right)-4 \cdot H(X) \\
& \leq n(1)+n(n-1)-4(n-1) \\
& \leq 4
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{C}^{(2)}+3 \cdot \mathcal{C}^{(1)} & =\sum_{A=1}^{\binom{n}{2}} H\left(X_{A}\right)+\sum_{i=1}^{n} 3 \cdot H\left(X_{i}\right)-6 \cdot H(X) \\
& \leq 2\binom{n}{2}+n(3)-6(2) \\
& \leq 12 \\
\mathcal{C}^{(2)}+3 \cdot \mathcal{C}^{(3)} & =\sum_{A=1}^{\binom{n}{2}} H\left(X_{A}\right)+\sum_{i=1}^{n} 3 \cdot H\left(X_{E / i}\right)-12 \cdot H(X) \\
& \leq 2\binom{n}{2}+n(n-1)(3)-12(n-1) \\
& \leq 12
\end{aligned}
$$

## REFERENCES

[1] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina, "Higher-order interactions stabilize dynamics in competitive network models," Nature, vol. 548, pp. 210 EP -, 07 2017. [Online]. Available: https://doi.org/10.1038/nature23273
[2] E. Tekin, C. White, T. M. Kang, N. Singh, M. Cruz-Loya, R. Damoiseaux, V. M. Savage, and P. J. Yeh, "Prevalence and patterns of higher-order drug interactions in escherichia coli," npj Systems Biology and Applications, vol. 4, no. 1, p. 31, 2018. [Online]. Available: https://doi.org/10.1038/s41540-018-0069-9
[3] P. E. Latham and S. Nirenberg, "Synergy, redundancy, and independence in population codes, revisited," Journal of Neuroscience, vol. 25, no. 21, pp. 5195-5206, 2005. [Online]. Available: http://www.jneurosci.org/content/25/21/5195
[4] S. Chen and J. C. Mar, "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data," BMC Bioinformatics, vol. 19, no. 1, p. 232, Jun 2018. [Online]. Available: https://doi.org/10.1186/s12859-018-2217-z
[5] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, "Quantifying information modification in developing neural networks via partial information decomposition," Entropy, vol. 19, no. 9, 2017. [Online]. Available: http://www.mdpi.com/1099-4300/19/9/494
[6] J. W. Kay, R. A. A. Ince, B. Dering, and W. A. Phillips, "Partial and entropic information decompositions of a neuronal modulatory interaction," Entropy, vol. 19, no. 11, 2017. [Online]. Available: http://www.mdpi.com/1099-4300/19/11/560
[7] N. M. Timme, S. Ito, M. Myroshnychenko, S. Nigam, M. Shimono, F.-C. Yeh, P. Hottowy, A. M. Litke, and J. M. Beggs, "Highdegree neurons feed cortical computations," PLOS Computational Biology, vol. 12, no. 5, pp. 1-31, 05 2016. [Online]. Available: https://doi.org/10.1371/journal.pcbi. 1004858
[8] A. Erramuzpe, G. J. Ortega, J. Pastor, R. G. de Sola, D. Marinazzo, S. Stramaglia, and J. M. Cortes, "Identification of redundant and synergetic circuits in triplets of electrophysiological data," Journal of Neural Engineering, vol. 12, no. 6, p. 066007, 2015. [Online]. Available: http://stacks.iop.org/1741-2552/12/i=6/a=066007
[9] M. Wibral, J. T. Lizier, and V. Priesemann, "Bits from brains for biologically inspired computing," Frontiers in Robotics and AI, vol. 2, p. 5, 2015. [Online]. Available: https://www.frontiersin.org/article/10. 3389/frobt.2015.00005
[10] T. M. Tax, P. A. Mediano, and M. Shanahan, "The partial information decomposition of generative neural network models," Entropy, vol. 19, no. 9, 2017. [Online]. Available: http://www.mdpi.com/1099-4300/ 19/9/474
[11] S. Yu, R. Jenssen, and J. C. Príncipe, "Understanding convolutional neural network training with information theory," CoRR, vol. abs/1804.06537, 2018. [Online]. Available: http://arxiv.org/abs/1804. 06537
[12] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene regulatory network inference from single-cell data using multivariate information measures," Cell Systems, vol. 5, no. 3, pp. 251 - 267.e3, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/ S2405471217303861
[13] P. Chatterjee and N. R. Pal, "Construction of synergy networks from gene expression data related to disease," Gene, vol. 590, no. 2, pp. $250-262,2016$. [Online]. Available: http://www.sciencedirect.com/ science/article/pii/S037811191630395X
[14] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, "Synergy, redundancy, and multivariate information measures: an experimentalist's perspective," Journal of Computational Neuroscience, vol. 36, no. 2, pp. 119-140, Apr 2014. [Online]. Available: https://doi.org/10.1007/s10827-013-0458-4
[15] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," $\operatorname{CoRR}$, vol. abs/1004.2515, 2010. [Online]. Available: http://arxiv.org/abs/1004.2515
[16] A. Makkeh, D. O. Theis, and R. Vicente, "Bivariate partial information decomposition: The optimization perspective," Entropy, vol. 19, no. 10, 2017. [Online]. Available: http://www.mdpi.com/ 1099-4300/19/10/530
[17] F. MacWilliams and N. Sloane, The Theory of Error-Correcting Codes, 2nd ed. North-holland Publishing Company, 1978.
[18] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, no. 11, pp. 612-613, Nov. 1979. [Online]. Available: http: //doi.acm.org/10.1145/359168.359176
[19] J. Hirschfeld and J. Thas, "Open problems in finite projective spaces," Finite Fields and Their Applications, vol. 32, pp. $44-81$, 2015, special Issue : Second Decade of FFA. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1071579714001191
[20] S. Watanabe, "Information theoretical analysis of multivariate correlation," IBM Journal of Research and Development, vol. 4, no. 1, pp. 66 - 82, Jan 1960.
[21] M. Studeny and J. Vejnarova, "The multiinformation function as a tool for measuring stochastic dependence," in Learning in Graphical Models. MIT Press, 1998, pp. 261-297.
[22] J. P. Crutchfield, "The calculi of emergence: computation, dynamics and induction," Physica D: Nonlinear Phenomena, vol. 75, no. 1-3, pp. 11-54, 1994.
[23] S. A. Abdallah and M. D. Plumbley, "A measure of statistical complexity based on predictive information with application to finite spin systems," Physics Letters A, vol. 376, no. 4, pp. 275-281, 2012.
[24] T. S. Han, "Linear dependence structure of the entropy space," Information and Control, vol. 29, pp. 337-368, 1975.
[25] -_, "Nonnegative entropy measures of multivariate symmetric correlations," Information and Control, vol. 36, no. 2, pp. 133 -

156, 1978. [Online]. Available: http://www.sciencedirect.com/science/ article/pii/S0019995878902759
[26] N. Ay and A. Knauf, "Maximizing multi-information," arXiv:mathph/0702002, 2007.
[27] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," CoRR, vol. abs/1311.2852, 2013. [Online]. Available: http://arxiv.org/abs/1311.2852
[28] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," Information and Control, vol. 39, no. 1, pp. 55-72, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/ S001999587891063X
[29] S. Watanabe, Knowing and guessing : a quantitative study of inference and information. John Wiley and Sons Inc, 1969.
[30] N. Ay, E. Olbrich, N. Bertschinger, and J. Jost, "A geometric approach to complexity," Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 21, no. 3, p. 037103, 2011. [Online]. Available: https://doi.org/10.1063/1.3638446
[31] V. Griffith and C. Koch, "Quantifying synergistic mutual information," CoRR, vol. abs/1205.4265, 2012. [Online]. Available: http://arxiv.org/ abs/1205.4265
[32] J. Oxley, Matroid Theory, ser. Oxford graduate texts in mathematics. Oxford University Press, 2006. [Online]. Available: https://books. google.com/books?id=puKta1Hdz-8C
[33] A. Beimel, "Secret-sharing schemes: A survey," in Coding and Cryptology, Y. M. Chee, Z. Guo, S. Ling, F. Shao, Y. Tang, H. Wang, and C. Xing, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 11-46.
[34] S. Ball and A. Chowdhury, "Inclusion matrices and the mds conjecture," 112015.
[35] S. Ball and J. De Beule, "On sets of vectors of a finite vector space in which every subset of basis size is a basis II," ArXiv e-prints, Jan. 2012.
[36] I. S. Reed and G. Solomon, "Polynomial Codes Over Certain Finite Fields," Journal of the Society for Industrial and Applied Mathematics, vol. 8, no. 2, pp. 300-304, 1960. [Online]. Available: http://dx.doi.org/10.1137/0108018
[37] R. Lidl and H. Niederreiter, Introduction to Finite Fields and Their Applications. New York, NY, USA: Cambridge University Press, 1986.
[38] F. Matus, "Ascending and descending conditional independence relations," 1992.
[39] -_, "Divergence from factorizable distributions and matroid representations by partitions," IEEE Transactions on Information Theory, vol. 55, no. 12, pp. 5375-5381, Dec 2009.
[40] -_, "Matroid representations by partitions," Discrete Mathematics, vol. 203, no. 1, pp. $169-194$, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0012365X99000047
[41] J. Rauh, "Secret sharing and shared information," CoRR, vol. abs/1706.06998, 2017. [Online]. Available: http://arxiv.org/abs/1706. 06998
[42] C. Chan, A. Al-Bashabsheh, J. B. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," Proceedings of the IEEE, vol. 103, no. 10, pp. 1883-1913, Oct 2015.
[43] S. M. Dodunekov and I. N. Landgev, "On near-mds codes," in Proceedings of 1994 IEEE International Symposium on Information Theory, June 1994, pp. 427-.
[44] M. A. De Boer, "Almost mds codes," Designs, Codes and Cryptography, vol. 9, no. 2, pp. 143-155, Oct 1996. [Online]. Available: https://doi.org/10.1007/BF00124590


[^0]:    *Correspondence: reing@usc.edu
    University of Southern California, Information Sciences Institute, reing@usc.edu, gregv@isi.edu, galstyan@isi.edu

[^1]:    ${ }^{1}$ We don't make the distinction here between pairwise and multivariate redundancy. Such a distinction is important in information decomposition, where shared and unique information [27] are distinguished with respect to a target variable

