Functional Compression through Graph Coloring

Vishal Doshi, *Student Member, IEEE*, Devavrat Shah, *Member, IEEE*, Muriel Médard, *Fellow, IEEE*,
Michelle Effros, *Member, IEEE*

Abstract—We consider the problem of functional compression. The objective is to separately compress possibly correlated discrete sources such that a fixed and arbitrary deterministic function of those sources can be computed given the compressed data from each source. This is motivated by problems in sensor networks and database privacy. Our architecture gives a quantitative definition of privacy for database statistics. Further, we show that this method can provide significant coding gains over traditional Slepian-Wolf coding in sensor networks.

We consider both the lossless and lossy computation of a function. Specifically, we present results of the rate regions for three instances of the problem where there are two sources: 1) lossless computation where one source is available at the decoder, 2) under a special condition, lossless computation where both sources are separately encoded, and 3) lossy computation where one source is available at the decoder.

Our general strategy extends an approach developed by Orlitsky and Roche (2001) for the first case. We extend that result by providing an achievability scheme that is based on the coloring of the characteristic graph. This suggests a layered architecture where the functional layer controls the coloring scheme, and the data layer uses existing distributed source coding schemes. We extend this graph coloring method to provide algorithms and rates for all three problems.

Index Terms—Distributed source coding, functional compression, distributed computing

I. INTRODUCTION

Generally speaking, data compression considers the compression of a source (sources) and its (their) recovery via a decoding algorithm. Functional compression considers the recovery not of the sources, but of a function of the sources. It is a method for reducing the number of bits required to convey relevant information from disparate sources to a third party. The key contributions of this article are to provide meaning to the word "relevant" in this context. We will derive the information theoretic limits for a selection of functional compression problems and give novel algorithms to achieve these rates.

A. Motivations and Applications

We are motivated to study this problem mainly by two applications. First, consider medical records databases. The data is located in several different locations. There are enormous

amounts of private data in the databases. Some government agency wants to release certain statistics, or functions, of the data useful to researchers. Thinking of the data as a bit-string, we provide a way for the agency to release a minimal set of bits to compute a set of allowable functions. Thus, our architecture allows for a minimal loss of privacy, given the need to compute certain statistics.

Next, consider a network of wireless sensors measuring temperature in a building. There are bandwidth and power constraints for each sensor, and the sensors communicate only with a central receiver, not with each other. The receiver wishes only to compute the average temperature in the building. We want to determine if it is possible to compress beyond the traditional distributed data compression rate bounds given by Slepian and Wolf.

We can frame both of the above questions as functional compression problems. In each case, we wish to minimize the source description rates either to guarantee privacy or to achieve higher compression rates (thus conserving bandwidth and power).

We demonstrate the possible rate gains by example.

Example 1. *Consider two sources uniformly and independently producing k-bit integers X and Y; assume $k \geq 2$. We assume independence to bring to focus the compression gains from using knowledge of the function. First suppose $f(X,Y) = (X,Y)$ is to be perfectly reconstructed at the decoder. Then, the rate at which X can encode its information is k bits per symbol (bps); the same holds for Y. Thus the sum rate is $2k$ bits per function-value (bpf).*

Next, suppose $f(X,Y) = X + Y \mod 4$. The value of $f(X,Y)$ depends only upon the final two bits of both X and Y. Thus, at most (and in fact, exactly) 2 bps is the encoding rate, for a sum rate of 4 bpf. Note that the rate advantage, $2k - 4$ is unbounded because we are reducing a possibly huge alphabet to one of size 4.

Finally, suppose $f(X,Y) = X + Y \mod 4$ as before, but the decoder is allowed to recover f up to some distortion. We consider the Hamming distortion function on f. Consider recovering f up to a Hamming distortion of 1. One possible coding scheme would simply encode the single least significant bit for both X and Y. Then one could recover the least significant bit of $f(X,Y)$, thus achieving an encoding rate of 1 bps per source or 2 bpf. □

Using knowledge of the decoder's final objective helps achieve better compression rates. Example 1 is relatively simple because the function is separable. When the function is not separable (e.g. $f(X,Y) = |X - Y|$), it is much less obvious how to separately encode X and Y.

This article provides a general framework that, in certain

TABLE I
RESEARCH ARTICLES ON ZERO-DISTORTION SOURCE CODING PROBLEMS

Problem types	$f(x,y) = (x,y)$	General $f(x,y)$
Side information	Shannon [1]	Orlitsky and Roche [2]
Distributed	Slepian and Wolf [3]	Ahlswede and Körner [4] Körner and Marton [5]

cases, allows us to solve both the problem of finding the best possible rates as well as coding schemes that allow for approximations of these rates.

Example 1 showcases the three specific scenarios considered in this article: side information with zero distortion, distributed compression with zero distortion, and side information with nonzero distortion. We proceed by placing our results in their historical context.

B. Historical Context

We can categorize compression problems with two sources along three dimensions. First, whether one source is locally available at the receiver (call this "side information") or whether both sources are communicating separately (call this "distributed"). Second, whether $f(x,y) = (x,y)$ or is more general. Finally, whether there is zero-distortion or non-zero-distortion. In all cases, the goal is to determine the rates ((R_x, R_y) for distributed coding or R_x for side information coding) at which X and Y must be encoded in order for the decoder to compute $f(X,Y)$ within distortion $D \geq 0$ with high probability.

1) Zero Distortion: First, consider zero distortion. Shannon [1] considers the side information problem where $f(x,y) = x$. Slepian and Wolf [3] consider the distributed problem where $f(x,y) = (x,y)$. Many practical and near-optimal coding schemes have been developed for both of the above problems such as DISCUS codes by Pradhan and Ramchandran [6] and source-splitting techniques by Coleman et al. [7]. We provide the precise theorems in a later section.

Orlitsky and Roche provide a single-letter characterization for the side information problem for a general function $f(x,y)$. Ahlswede and Körner [4] determine the rate region for the distributed problem for $f(x,y) = x$. Körner and Marton [5] consider zero-distortion with both sources separately encoded for the function $f(x,y) = x + y \mod 2$. There has been little work on a general function $f(x,y)$ in the distributed zero-distortion case.

A sampling of these contributions are summarized in Table I. The * indicates where this article contributes. Specifically, for zero distortion, we provide a framework that leads to an optimal modular coding scheme for the side information problem for general functions. We give conditions under which this framework can be extended to the distributed problem for general functions.

2) Nonzero Distortion: Next, consider nonzero distortion problems. Wyner and Ziv [8] considered the side information problem for $f(x,y) = x$. The rate region for the case of nonzero-distortion with both sources separately encoded is unknown but bounds have been given by Berger and Yeung [9],
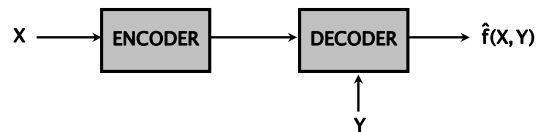

Fig. 1. The functional compression problem with side information.

Barros and Servetto [10], and Wagner, Tavildar, and Viswanath [11]. Wagner et al. considered a specific distortion function for their results (quadratic). In the context of functional compression, all of these theorems are specific to $f(x,y) = (x,y)$.

Feng, Effros, and Savari [12] solved the side information problem where the encoder and decoder have noisy information about the sources. Yamamoto solved the side information problem for a general function $f(x,y)$.

A sampling of these contributions are summarized in Table II. Specifically, for nonzero distortion, we extend the framework derived for zero distortion and apply it in this more general setting. As indicated above, the distributed setting with nonzero distortion and a general function is quite difficult (even the special case $f(x,y) = (x,y)$ is not completely solved).

C. Overview of Results

Now, we describe the three problems considered in this article along with an overview of the results.

1) Functional Compression with Side Information: The side information problem is depicted in Figure 1. We describe a scheme for encoding X such that $f(X,Y)$ can be computed within expected distortion D at a receiver that knows Y.

The optimal rate for the functional compression with side information problem was given by Orlitsky and Roche in the zero distortion case. While the resulting characterization is complete and tight, it is difficult to calculate even for simple source distributions. For this problem, this article provides a new interpretation for that rate through a simple algorithm that can be approximated with available heuristics. Computing the Orlitsky-Roche rate requires optimizing a distribution over an auxiliary random variable W. We provide an interpretation of W that leads to a simple achievability scheme for the Orlitsky-Roche rate that is modular with each module being a well-studied problem. It can be extended to and motivates our functional distributed source coding scheme below.

As mentioned earlier, Yamamoto gave a characterization of the rate distortion function for this problem as an optimization
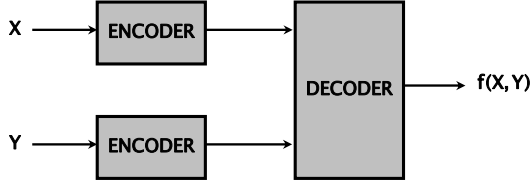
Fig. 2.   The distributed functional compression problem.

TABLE II
RESEARCH ARTICLES ON NONZERO-DISTORTION SOURCE CODING PROBLEMS

Problem types	$f(x,y) = (x,y)$	General $f(x,y)$
Side information	Wyner and Ziv [8] Feng et al. [12]	Yamamoto [14] *
Distributed	Berger and Yeung [9] Barros and Servetto [10] Wagner et al. [11]	

over an auxiliary random variable. We give a new interpretation to Yamamoto's rate distortion function for nonzero distortion. Our formulation of the rate distortion function leads to a coding scheme that extends the coding schemes for the zero distortion case. Further, we give a simple achievability scheme that achieves compression rates that are certainly at least as good as the Slepian-Wolf rates and also at least as good as the zero distortion rate.

For zero distortion, the rate is a special case of the distributed functional compression problem considered next where one source is compressed at entropy-rate, thus allowing for reconstruction at the decoder.

2) Distributed Functional Compression: The distributed functional compression problem is depicted in Figure 2. In this problem, X and Y are separately encoded such that the decoder can compute $f(X,Y)$ with zero distortion and arbitrarily small probability of error. We describe the conditions under which the coding scheme mentioned in the previous section can be extended to the distributed set up. Further, we provide a less general condition depending only upon the probability distribution of the sources under which our scheme is optimal.

Thus, we extend the Slepian-Wolf rate region to a general deterministic function $f(x,y)$. Our rate region is, in general, an inner bound to the general rate region, and we provide conditions under which it is the true rate region.

D. Organization

The rest of the article is organized as follows. Section II gives the problem statement and presents the related technical background necessary to understand the main results. Section III presents those results. Section IV gives an example application of our results to Blue Force Tracking. The proofs for our results are given in Section V. Future research directions and conclusions are given in Chapter VI.

II. FUNCTIONAL COMPRESSION BACKGROUND

We consider the three proposed problems within a common framework. We borrow much of the notation from [13, Chapter 12].

A. Problem Setup

Let $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ be discrete memoryless sources drawn from finite sets $\mathcal{X}$ and $\mathcal{Y}$ according to a joint distribution $p(x,y)$. Denote by $p(x)$ and $p(y)$ the marginals of $p(x,y)$.

We denote n-sequences of random variables X and Y as $\mathbf{X} = \{X_i\}_{i=k}^{k+n-1}$ and $\mathbf{Y} = \{Y_i\}_{i=k}^{k+n-1}$, respectively, where n and k are clear from context. We generally assume $k = 1$. Because the sequence $(\mathbf{x},\mathbf{y})$ was drawn i.i.d. according to $p(x,y)$, we can write the probability of any instance of the sequence as $p(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{n} p(x_i,y_i)$.

The sources encode their messages (at rates $R_x, R_y \in [0,\infty)$); a common decoder uses these descriptions to compute an approximation to a fixed deterministic function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ or $f : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}^n$, its vector extension, where again n will be clear from context.

For any n, D, R_x, and R_y, we define a *distributed functional code* of block length n for the joint source (X,Y) and function f as two encoder maps,

$$e_x : \mathcal{X}^n \to \{1,\ldots,2^{nR_x}\},$$
$$e_y : \mathcal{Y}^n \to \{1,\ldots,2^{nR_y}\},$$

and a decoder map,

$$r : \{1,\ldots,2^{nR_x}\} \times \{1,\ldots,2^{nR_y}\} \to \mathcal{Z}^n.$$

Consider a distortion function, $d : \mathcal{Z} \times \mathcal{Z} \to [0,\infty)$, with vector extension

$$d(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{n} \sum_{i=1}^{n} d(z_{1i}, z_{2i}),$$

where $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}^n$. As in [8], we assume that the distortion function satisfies $d(z_1, z_2) = 0$ if and only if $z_1 = z_2$. (Otherwise, one can define the equivalence classes of the function values to make this condition hold.)

The *probability of error* is

$$P_e^n = \Pr[\{(\mathbf{x},\mathbf{y}) : d(f(\mathbf{x},\mathbf{y}), r(e_x(\mathbf{x}), e_y(\mathbf{y}))) > D\}].$$

A rate pair, (R_x, R_y), is *achievable* for a distortion D if there exists a sequence of distributed functional codes at those rates and distortion level such that $P_e^n \to 0$ as $n \to \infty$. The *achievable rate region* is the set closure of the set of all achievable rates. Our most general objective is to find this achievable rate region.

B. Previous Results

We begin by defining a construct useful in formulating all the results.

Definition 2. *The characteristic graph* $G_x = (V_x, E_x)$ *of* X *with respect to* Y, $p(x,y)$, *and* $f(x,y)$ *is defined as follows:* $V_x = \mathcal{X}$ *and* $(x_1, x_2) \in \mathcal{X}^2$ *is in* E_x *if there exists a* $y \in \mathcal{Y}$ *such that* $p(x_1, y)p(x_2, y) > 0$ *and* $f(x_1, y) \neq f(x_2, y)$.
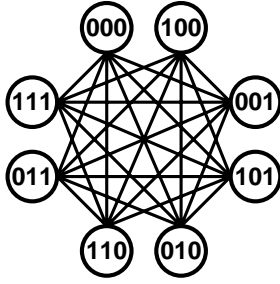
Fig. 3. Example of a characteristic graph.

Defined thus, G_x is the "confusability graph" from the perspective of the receiver. If $(x_1, x_2) \in E_x$, then the descriptions of x_1 and x_2 must be different to avoid confusion about $f(x, y)$ at the receiver. This was first defined by Shannon when studying the zero error capacity of noisy channels [15]. Witsenhausen [16] used this graph to consider our problem in the case when one source is deterministic, or equivalently, when one encodes X to compute $f(X)$ with 0 distortion. The characteristic graph of Y with respect to X, $p(x, y)$, and $f(x, y)$ is defined analogously and denoted G_y. When notationally convenient and clear, we will drop the subscript.

The importance of using the characteristic graph construct becomes clear when considering independent sets[1] of the graph. By definition of the edge set, knowledge of Y and the independent set uniquely determines $f(x, y)$.

We illustrate this with Example 3.

Example 3. *To illustrate the idea of confusability and the characteristic graph, consider again the sources from Example 1. Suppose both sources are uniformly and independently generating 3-bit integers. The function of interest is $f(x, y) = x + y \mod 4$. Then, the characteristic graph of X with respect to Y, $p(x, y) = \frac{1}{8}$ for all (x, y), and f is shown in Figure 3. There is an edge between x_1 and x_2 if and only if they differ in one or both of their final two bits. This is because, given any y, $f(x_1, y) = f(x_2, y)$ if and only if x_1 and x_2 have the same final two bits.* $\square$

Next, we define graph entropy, which we use later to more generally derive the rate required for communication for problems such as in Example 3.

Definition 4. *Given a graph $G = (V, E)$ and a distribution on the vertices V, Körner [17] defines the graph entropy as:*

$$H_G(X) = \min_{X \in W \in \Gamma(G)} I(W; X), \qquad (1)$$

where $\Gamma(G)$ is the set of all independent sets of G.

The notation $X \in W \in \Gamma(G)$ means that we are minimizing over all distributions $p(w, x)$ such that $p(w, x) > 0$ implies $x \in w$ where w is an independent set of the graph G. We now demonstrate how this can be used to solve problems like that given in Example 3.

---

[1] A subset of vertices of a graph G is an independent set if no two nodes in the subset are adjacent to each other in G. With the characteristic graph, independent sets form equivalence classes.

Example 5. *Consider again the scenario in Example 1 as presented in Example 3. For the graph in Figure 3, the maximally independent sets are the sets with the same final two bits. To minimize $I(X; W) = H(X) - H(X|W)$, we must maximize $H(X|W)$; this occurs when $p(w)$ is nonzero only over the 4 maximally independent sets of the graph. This is because $H(X|w) = 1$ for all maximally independent w. Therefore, we get $H_G(X) = 3 - 1 = 2$.* $\square$

Witsenhausen [16] considered a graph with vertices equal to the support of the random variable X and the edge set defined such that x and x' have an edge when $f(x) \neq f(x')$; he showed that the graph entropy is the minimal rate at which a single source can be encoded such that a desired function can be computed with zero distortion. Witsenhausen's graph equals the characteristic graph of X with respect to Y, $p(x, y)$, and $f(x, y) = f(x)$ when Y is constant.

Orlitsky and Roche [2] defined an extension of Körner's graph entropy, the *conditional graph entropy*.

Definition 6. *The conditional graph entropy is*

$$H_G(X|Y) = \min_{\substack{X \in W \in \Gamma(G) \\ W - X - Y}} I(W; X|Y). \qquad (2)$$

The additional constraint that $W - X - Y$ forms a Markov chain formally enforces the constraint that W should not contain any information about Y that is not available through X. If X and Y are independent, $H_G(X|Y) = H_G(X)$.

Theorem 7 (Orlitsky-Roche Theorem, 2001 [2]). *When G is the characteristic graph of X with respect to Y, $p(x, y)$, and $f(x, y)$, $R_x \geq H_G(X|Y)$ is the rate region for reliable computation of the function $f(X, Y)$ with zero distortion and arbitrarily small probability of error when Y is available as side information.*

A natural extension of this problem is the functional compression with side information problem for nonzero distortion. Yamamoto gives a full characterization of the rate-distortion function for the side information functional compression problem [14] as a generalization of the Wyner-Ziv side-information rate-distortion function [8]. Specifically, Yamamoto gives the rate distortion function as follows.

Theorem 8 (Yamamoto Theorem, 1982). *The rate distortion function for the functional compression with side information problem is*

$$R(D) = \min_{p \in \mathcal{P}(D)} I(W; X|Y)$$

where $\mathcal{P}(D)$ is the collection of all distributions on W given X such that $W - X - Y$ forms a Markov chain and there exists a $g : \mathcal{W} \times \mathcal{Y} \to \mathcal{Z}$ satisfying $E[d(f(\mathbf{X}, \mathbf{Y}), g(\mathbf{W}, \mathbf{Y}))] \leq D$.

This is a natural extension of the Wyner-Ziv rate-distortion result [8]. The constraint $X \in W \in \Gamma(G)$ in the definition of the Orlitsky-Roche rate (Definition 6) specifies a subset of distributions in $\mathcal{P}(D)$ which retain optimality when $D = 0$.

C. Graph Entropies

Our results depend on the use of more graph tools, which we now describe. Alon and Orlitsky [18] defined the OR-

power graph of G as $G^n = (V_n, E_n)$ where $V_n = V^n$ and two vertices $(\mathbf{x}_1, \mathbf{x}_2) \in E_n \subseteq V_n \times V_n$ if any component $(x_{1i}, x_{2i}) \in E$. Thus, two blocks of source observations are confusable if any pair of symbols in those blocks are confusable.

A vertex coloring of a graph is any function $c : V \to \mathbb{N}$ of a graph $G = (V, E)$ such that $(x_1, x_2) \in E$ implies $c(x_1) \neq c(x_2)$. The entropy of a coloring is the entropy of the induced distribution on colors $p(c(x)) = p(c^{-1}(c(x)))$ where $c^{-1}(x) = \{\bar{x} : c(\bar{x}) = c(x)\}$ and is called a *color class*.

Definition 9. *Let $\mathcal{A} \subset \mathcal{X} \times \mathcal{Y}$ be a high probability set (for our purposes, any subset such that $p(\mathcal{A}) \geq 1 - \varepsilon$). Define $\hat{p}(x, y) = p(x, y)/p(\mathcal{A})$. In other words, $\hat{p}$ is the distribution over (x, y) conditioned on $(x, y) \in \mathcal{A}$. Denote the characteristic graph of X with respect to Y, $\hat{p}$, and f as $\hat{G}_x = (\hat{V}_x, \hat{E}_x)$ and the characteristic graph of Y with respect to X, $\hat{p}$, and f as $\hat{G}_y = (\hat{V}_y, \hat{E}_y)$. Note that $\hat{E}_x \subseteq E_x$ and $\hat{E}_y \subseteq E_y$. Finally, we say c_x and c_y are ε-colorings of G_x and G_y if they are valid colorings of $\hat{G}_x$ and $\hat{G}_y$ defined with respect to some high probability set $\mathcal{A}$.*

Alon and Orlitsky [18] defined the *chromatic entropy* of a graph G.

Definition 10.

$$H_G^\chi(X) = \min_{c \text{ is an } \varepsilon\text{-coloring of } G} H(c(X)).$$

Well-known typicality results (e.g., [13]) imply that there exists a high probability set for which the graph vertices are roughly equiprobable. Thus, the chromatic entropy is a representation of the chromatic number of high probability subgraphs of the characteristic graph. We define a natural extension, the *conditional chromatic entropy*, as follows.

Definition 11.

$$H_G^\chi(X|Y) = \min_{c \text{ is an } \varepsilon\text{-coloring of } G} H(c(X)|Y).$$

The above optimizations are minima and not infima because there are only finitely many subgraphs of any fixed G, and thus only finitely many ε-colorings regardless of ε. Later, in order to use typicality results, we allow the block length n to grow without bound in order to drive the error probability to zero and therefore use G^n and study the infimum over all n.

These optimizations are NP-hard [19], but they can be approximated using existing heuristics [20], [21]. With these definitions and results, we can now formally describe our results.

III. MAIN RESULTS

The proofs of the results described in this section appear in Section V.

A. Functional Compression with Side Information $(D = 0)$

We begin by describing the zero distortion problem for a single source X and function $f(X)$. There is no side information at the decoder. Witsenhausen [16] tells us that the optimal rate is the graph entropy $H_G(X)$ defined earlier
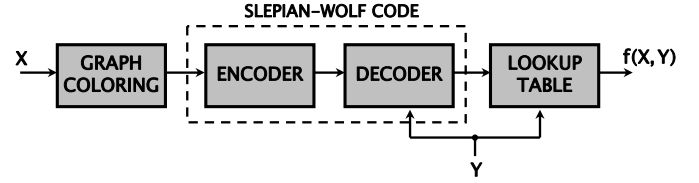

Fig. 4. Source coding scheme for the zero distortion functional compression problem with side information.

in Definition 4 where G is the characteristic graph of X with respect to the function $f(X)$. As stated earlier, the chromatic entropy is a representation of the chromatic number of a high probability subgraph of the characteristic graph. Körner proved [17] that the chromatic entropy approaches the graph entropy as block length n grows without bound.

Theorem 12 (Körner Theorem, 1973)**.**

$$\lim_{n \to \infty} \frac{1}{n} H_{G^n}^\chi(\mathbf{X}) = H_G(X). \tag{3}$$

The implications of this result are that we can compute a function of a discrete memoryless source with vanishing probability of error by first coloring a sufficiently large power graph of the characteristic graph of the source with respect to the function, and then, encoding the colors using any code that achieves the entropy bound on the colored source. The previous approach for achieving rates close to the bound $H_G(X)$ was to optimize with respect to a distribution over W as in the definition of $H_G(X)$. This theorem allows us to move the optimization from finding the optimal distribution to finding the optimal colorings. Thus, our solution modularizes the coding by first creating a graph coloring problem (for which heuristics exist), and then transmitting the colors using any existing entropy-rate code. Moreover, we can extend this technique to the functional side information case.

Next, consider the problem of lossless functional source coding with side information. Orlitsky and Roche proved the optimal rate for the zero distortion functional compression problem with side information equals $H_G(X|Y)$. Recall from Definition 6 that $H_G(X|Y)$ is also achieved by optimizing a distribution over W. Theorem 13 extends Körner's Theorem to conditional chromatic and conditional graph entropies.

Theorem 13.

$$\lim_{n \to \infty} \frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) = H_G(X|Y)$$

This theorem extends the previous result to the conditional case. In other words, in order to encode a memoryless source, we first color a graph G^n for sufficiently large n. Then, we encode each source symbol with its corresponding vertex's color. Finally, we use a Slepian-Wolf code on the sequence of colors achieving a rate arbitrarily close to $H(c(\mathbf{X})|\mathbf{Y})$. This allows for computation of the function at the decoder. The resulting code has rate arbitrarily close to the Orlitsky-Roche bound $H_G(X|Y)$ by Theorem 13. Figure 4 illustrates our coding scheme.

The Orlitsky-Roche achievability proof uses random coding arguments and proves the existence of an optimal coding scheme, but does not specify it precisely. Our coding scheme allows the use of heuristics available for finding good colorings as well as the use of optimal source codes that achieve the conditional entropy. Finding the minimum entropy colorings required to achieve the bound is NP-hard [22], [19], but even simple colorings (weakly)[2] improve over the bound $H(X|Y)$ that arises when trying to recover X completely at the receiver by the Data Processing Inequality. This solution gives the corner points of the achievable rate region for the distributed functional compression problem, considered next.

B. Distributed Functional Compression ($D = 0$)

In this section, we prove rate bounds for the distributed functional compression problem. The derived rate region is always achievable and sometimes tight. The region directly evolves from the coloring arguments discussed in the above section.

Recall the lossless distributed functional compression problem shown in Figure 2. Our goal is to provide an achievability scheme that extends the modular scheme given in Figure 4 for the side information case. Again, the code first precodes the data (coloring) and then describes the colors using existing Slepian-Wolf source codes.

The Slepian-Wolf Theorem [3] states that in order to recover a joint source (X, Y) at a receiver, it is both necessary and sufficient to encode separately sources X and Y at rates (R_x, R_y) where

$$R_x \geq H(X|Y)$$
$$R_y \geq H(Y|X)$$
$$R_x + R_y \geq H(X, Y).$$

Denote this region as $\mathcal{R}(X, Y)$.

For any n and functions g_x and g_y defined on $\mathcal{X}^n$ and $\mathcal{Y}^n$, respectively, denote by $\mathcal{R}^n(g_x, g_y)$ the Slepian-Wolf region for $g_x(\mathbf{X})$ and $g_y(\mathbf{Y})$ normalized by the block length. Precisely, $\mathcal{R}^n(g_x, g_y)$ is the set of all (R_x, R_y) where

$$R_x \geq \frac{1}{n} H(g_x(\mathbf{X})|g_y(\mathbf{Y})),$$
$$R_y \geq \frac{1}{n} H(g_y(\mathbf{Y})|g_x(\mathbf{X})),$$
$$R_x + R_y \geq \frac{1}{n} H(g_x(\mathbf{X}), g_y(\mathbf{Y})).$$

If Y is sent at rate $H(Y)$, it can be faithfully recovered at the receiver. Thus, the rate for X is $H_{G_x}(X|Y)$ as given by Orlitsky and Roche. Similarly, when $R_x \geq H(X)$, $R_y \geq H_{G_y}(Y|X)$. Therefore, we know the corner points for the rate region for the distributed functional compression problem.

Our goal is to determine the region and give a scheme analogous to the one given in Figure 4 that achieves all rates in the given region. We proceed with the following philosophy: color X and Y using the characteristic graphs

[2]For all functions $f(\mathbf{X}, \mathbf{Y})$, $H(c(\mathbf{X})|\mathbf{Y}) \leq H(\mathbf{X}|\mathbf{Y})$. For any non-injective function $f(\mathbf{X}, \mathbf{Y})$ and large enough block length n, $H(c(\mathbf{X})|\mathbf{Y}) < H(\mathbf{X}|\mathbf{Y})$.
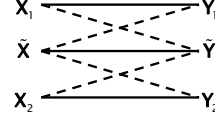

Fig. 5. An illustration of the Zigzag Condition.

G_x and G_y, and encode the colored sequences using codes achieving the Slepian-Wolf bounds. We want to characterize when this approach is valid. In other words, we want to find the conditions under which colorings of the characteristic graphs are sufficient to determine $f(x, y)$ for the zero distortion problem.

1) Zigzag Condition: A condition which is necessary and sufficient for the proposed coloring scheme to give a legitimate code follows.

Condition 14 (Legitimate Coloring). *For any n, consider ε-colorings c_x and c_y of G_x^n and G_y^n with associated probability distribution $\hat{p}$. The colorings c_x and c_y and the source distribution $\hat{p}(x, y)$ are said to satisfy the Legitimate Coloring condition if for all colors $(\gamma, \sigma) \in c_x(\mathcal{X}^n) \times c_y(\mathcal{Y}^n)$, and all $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in c_x^{-1}(\gamma) \times c_y^{-1}(\sigma)$ such that $\hat{p}(\mathbf{x}_1, \mathbf{y}_1)\hat{p}(\mathbf{x}_2, \mathbf{y}_2) > 0$, $f(\mathbf{x}_1, \mathbf{y}_1) = f(\mathbf{x}_2, \mathbf{y}_2)$.*

Condition 14 determines whether the colorings are sufficient for the decoder to reconstruct the desired function, but gives little insight into how to design good colorings. The Zigzag Condition, described next, gives conditions under which colorings (c_x, c_y) are legitimate only if c_x is a coloring of G_x and c_y is a coloring of G_y.

Condition 15 (Zigzag Condition). *A discrete memoryless source $\{(X_i, Y_i)\}_{i \in \mathbb{N}}$ with distribution $p(x, y)$ satisfies the Zigzag Condition if for any ε and some n, $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in T_\varepsilon^n$, there exists some $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in T_\varepsilon^n$ such that $(\tilde{\mathbf{x}}, \mathbf{y}_i), (\mathbf{x}_i, \tilde{\mathbf{y}}) \in T_{\frac{\varepsilon}{2}}^n$ for each $i \in \{1, 2\}$, and $(\tilde{x}_j, \tilde{y}_j) = (x_{ij}, y_{(3-i)j})$ for some $i \in \{1, 2\}$ for each j.*

Figure 5 illustrates the Zigzag Condition. If a solid line connects two values, then the pair is in T_ε^n. If a dashed line connects two values, then the pair is in $T_{\frac{\varepsilon}{2}}^n$. Thus, the Zigzag Condition is quite strict in the sense that many pairs must be typical. For any source that does not satisfy the Zigzag Condition, coloring G_x^n and G_y^n independently still allows the decoder to uniquely determine the function but may use more rate than required because it treats as jointly typical (x^n, y^n) pairs that are unlikely to occur together.

2) Rate Region: Let $\mathcal{S}^\varepsilon = \bigcup_{n=1}^\infty \bigcup_{(c_x^n, c_y^n)} \mathcal{R}^n(c_x^n, c_y^n)$ where for all n, c_x^n and c_y^n are ε-colorings of G_x^n and G_y^n. Let $\mathcal{S}$ be largest set that is a subset of $\mathcal{S}^\varepsilon$ for all $\varepsilon > 0$. Let $\bar{\mathcal{S}}$ be the set closure of $\mathcal{S}$. Finally, let $\mathcal{R}$ be the rate region for the distributed functional compression problem. We can now state the rate region in the notation just given.

Theorem 16. *For any $\varepsilon > 0$, $\mathcal{S}^\varepsilon$ is an inner bound to the rate region, and thus, $\mathcal{S}$ is an inner bound to the rate region. In other words, $\mathcal{R} \subseteq \mathcal{S}$. Moreover, under the Zigzag Condition, the rate region for the distributed functional source coding*

problem is $\mathcal{R} = \bar{\mathcal{S}}$.

Theorem 16 extends Theorem 13 to the distributed case by showing that optimal (independent) coloring of characteristic graphs G_x^n and G_y^n always yields a legitimate code, describing a family of problems for which these codes guarantee an optimal solution. While the above result is not a single letter characterization, *any* non-trivial (non-injective) coloring does better than the Slepian-Wolf rates, by the Data Processing Inequality (cf. [13]).

The Orlitsky-Roche bound is consistent with our region at $R_y = H(Y)$ by the following argument. If $R_y = H(Y)$, then $c_y(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y}$ typical with some $\mathbf{x}$. Thus, the rate R_x must be $R_x \geq \frac{1}{n} H(c_x(\mathbf{X})|\mathbf{Y})$ which is minimized at $H_{G_x}(X|Y)$ by Theorem 13.

Next, we derive a characterization of the minimum joint rate, $R_x + R_y$ in terms of graph entropies.

Corollary 17. *Under the Zigzag Condition, if there is a unique point that achieves the minimum joint rate, it must be $R_x + R_y = H_{G_x}(X) + H_{G_y}(Y)$.*

In this case, each encoder uses only its corresponding marginal ($p(x)$ or $p(y)$) when encoding. The resulting rates are $H(c_x(\mathbf{x}))$ and $H(c_y(\mathbf{y}))$, respectively.

When the jointly optimal rate is not unique, Theorem 18 bounds the difference between the minimal sum-rate and $H_{G_x}(X) + H_{G_y}(Y)$.

Theorem 18. *Let $I_{G_x}(X;Y) = H_{G_x}(X) - H_{G_x}(X|Y)$ be the graph information of X and Y for the graph G_x. Let $I_{G_y}(Y;X) = H_{G_y}(Y) - H_{G_y}(Y|X)$ be the graph information of Y and X for the graph G_y. Let R_{xy} equal the minimal sum-rate, $R_x + R_y$. Then, under the Zigzag Condition:*

$$\left[H_{G_x}(X) + H_{G_y}(Y)\right] - R_{xy} \leq \min\left\{I_{G_x}(X;Y), I_{G_y}(Y;X)\right\}.$$

Thus, for the case in Corollary 17, the mutual information of the minimum entropy colorings of G_x^n and G_y^n goes to zero as $n \to \infty$:

$$\lim_{n\to\infty} \frac{1}{n} I(c(\mathbf{X}); c(\mathbf{Y})) = 0.$$

If the independent sets of G_x are large, then $H_{G_x}(X)$ and $H_{G_x}(X|Y)$ are close, and $I_{G_x}(X;Y)$ is close to zero. Therefore, coloring followed by fixed block length compression (using $p(x)$, not $p(x,y)$) is not too far from optimal by Theorem 18. (Similarly, for G_y.) Another case when the right hand side of Theorem 18 is small is when X and Y have small mutual information. In fact, if X and Y are independent, the right hand side is zero and Corollary 17 applies.

The region given in Theorem 16 has several interesting properties. First, it is convex by time-sharing arguments for any two points in the region. Second, when there is a unique point (R_x, R_y) achieving the minimal sum-rate, we can give a single-letter characterization for that point (Corollary 17). When it is not unique, we have given a simple bound on performance.

Figure 6 presents a possible rate region for the case where the minimal sum rate is not uniquely achieved. (For ease of reading, we drop the subscripts for G_x and G_y and write G for both.)
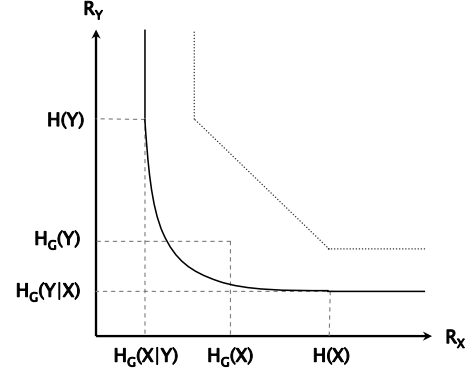

Fig. 6. An example rate region for the zero distortion distributed functional compression problem.

The "corners" of this rate region are $(H_{G_x}(X|Y), H(Y))$ and $(H(X), H_{G_y}(Y|X))$, the Orlitsky-Roche points, which can be achieved with graph coloring, in the limit sense, as described earlier. For any rate $R_x \in (H_{G_x}(X|Y), H(X))$, the joint rate required is less than or equal to the joint rate required by a time-sharing of the Orlitsky-Roche scheme. The inner region denoted by the dotted line is the Slepian-Wolf rate region.

The other point we characterize is the minimum joint rate point (when unique) given as $(H_{G_x}(X), H_{G_y}(Y))$. Thus, we have given a single-letter characterization for three points in the region.

C. Functional Compression with Side Information ($D > 0$)

We now consider the functional rate distortion problem; we give a new characterization of the rate distortion function given by Yamamoto. We also give an upper bound on that rate distortion function which leads to an achievability scheme that mirrors those given in the functional side information problem.

Recall the Yamamoto rate distortion function (Theorem 8). According to the Orlitsky-Roche result (Theorem 7), when $D = 0$, any distribution over independent sets of the characteristic graph (with the Markov chain $W - X - Y$ imposed) is in $\mathcal{P}(0)$. Any distribution in $\mathcal{P}(0)$ can be thought of as a distribution over independent sets of the characteristic graph.

We claim that finding a suitable reconstruction function, $\hat{f}$, is equivalent to finding the decoding function g on $\mathcal{W} \times \mathcal{Y}$ from Theorem 8.

For any m, let $\mathcal{F}_m(D)$ denote the set of all functions $\hat{f}_m : \mathcal{X}^m \times \mathcal{Y}^m \to \mathcal{Z}^m$ such that

$$\lim_{m\to\infty} E[d(f(\mathbf{X}, \mathbf{Y}), \hat{f}_m(\mathbf{X}, \mathbf{Y}))] \leq D.$$

To prove the claim, we consider blocks of length mn. The functions in the expectation above will be on $\mathcal{X}^{mn} \times \mathcal{Y}^{mn}$.

Let $\mathcal{F}(D) = \bigcup_{m\in\mathbb{N}} \mathcal{F}_m(D)$. Let $G(\hat{f})$ denote the characteristic graph of $\mathbf{X}$ with respect to $\mathbf{Y}$, $p(\mathbf{x}, \mathbf{y})$, and $\hat{f}$ for any $\hat{f} \in \mathcal{F}(D)$. For each m and all functions $\hat{f} \in \mathcal{F}(D)$, denote for brevity the normalized graph entropy $\frac{1}{m} H_{G(\hat{f})}(\mathbf{X}|\mathbf{Y})$ as $H_{G(\hat{f})}(X|Y)$.

Theorem 19.

$$R(D) = \inf_{\hat{f} \in \mathcal{F}(D)} H_{G(\hat{f})}(X|Y)$$

Note that $G(\hat{f})$ must be a subgraph of the characteristic graph G^m (for appropriate m) with respect to f. Because G^m is finite, there are only finitely many subgraphs. Thus, for any fixed error ε and associated block length m, this is a finite optimization. This theorem implies that once the suitable reconstruction function $\hat{f}$ is found, the functional side information bound (and achievability scheme) using the graph $G(\hat{f})$ is optimal in the limit.

Unfortunately, $\mathcal{F}(D)$ is an (uncountably) infinite set, but the set of graphs associated with these functions is countably infinite. Moreover, any allowable graph dictates an ordinal function, but it has no meaning in terms of distortion. Given the ordinal function $\hat{f}$, choosing the cardinal values that minimize expected distortion is a tractable optimization problem. This shows that if one could approximate f with $\hat{f}$, the compression rate may improve (even when $\hat{f}$ is not optimal).

The problem of finding an appropriate function $\hat{f}$ is equivalent to finding a new graph whose edges are a subset of the edges of the characteristic graph. This motivates Corollary 20 where we use the a graph parameterized by D to look at a subset of $\mathcal{F}(D)$. The resulting bound is not tight, but it provides a practical tool for tackling a very difficult problem.

Define the D-characteristic graph of X with respect to Y, $p(x, y)$, and $f(x, y)$, as having verticies $V = \mathcal{X}$ and the pair (x_1, x_2) is an edge if there exists some $y \in \mathcal{Y}$ such that $p(x_1, y)p(x_2, y) > 0$ and $d(f(x_1, y), f(x_2, y)) > D$. Denote this graph as G^D. Because $d(z_1, z_2) = 0$ if and only if $z_1 = z_2$, the 0-characteristic graph is the characteristic graph, i.e. $G^0 = G$.

Corollary 20. *The rate $H_{G^D}(X|Y)$ is achievable.*

Constructing this graph is not computationally difficult when the number of vertices is small. Given the graph, we have a set of equivalence classes for $\hat{f}$. One can then optimize $\hat{f}$ by choosing those values for the equivalence classes that minimize distortion. However, any legal values (values that lead to the graph G^D) will necessarily still have distortion within D. Indeed, this construction guarantees not only that expected distortion is less than or equal to D, but also that maximal distortion is always less than or equal to D. There are many possible improvements to be made here.

Theorem 13 and the corresponding achievability scheme, Corollary 20, gives a simple coding scheme that may potentially lead to large compression gains.

D. Possible Extensions

In all of the above problems, our achievability schemes are modular, providing a separation between the computation of the function and the lossless compression of the function descriptors.

The computation module is a graph coloring module. The specific problem of interest for our scheme is NP-hard [22], [19], but there is ample literature providing near-optimal graph
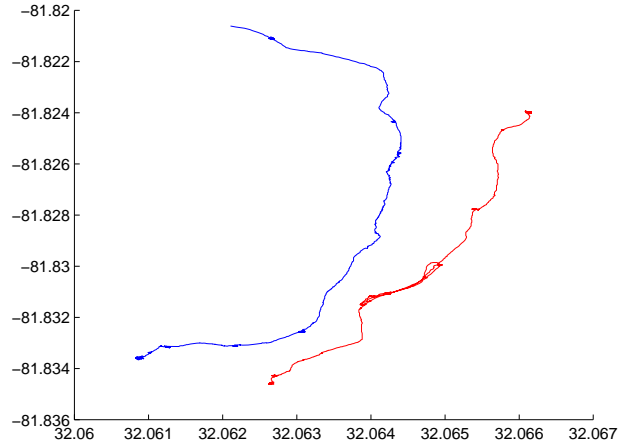

Fig. 7. Blue Force Tracking Data.

coloring heuristics for special graphs or heuristics that work well in certain cases [20], [21].

The lossless correlation module is a standard entropy coding scheme such as a Slepian-Wolf code. There are many practical algorithms with near-optimal performance for these codes. For example DISCUS codes [6] and source-splitting techniques [7].

Given the separation, the problem of functional compression becomes more tractable. While the overall problem may still be NP-hard, one can combine the results from each module to provide heuristics that are good for the specific task at hand.

We note that our results treat only two sources X and Y. Lossless distributed source codes for the more general scenario of N sources exists in the literature [13, p. 415]. Thus, it seems likely that given a suitable extension of the graph coloring technique and the Zigzag Condition, our results would generalize to N sources. We focus on the two-source scenario because, as with Slepian-Wolf, it gives many insights into the problem. We leave the extension to future work.

IV. SIMULATION RESULTS

In this section, we present an application of the work presented in the previous sections. We consider a sensor network scenario in which there are several sources communicating with some central receiver. This receiver wishes to learn some function of the sources.

Specifically, we consider Blue Force Tracking, which is a GPS system used by the U.S. Armed Forces to track friendly and enemy movements. Sometimes the nodes in the system communicate with each other, and sometimes they communicate with some central receiver, such as a UAV, which is the case considered here.

We present preliminary experimental results for the algorithm given for the distributed functional compression. We obtained tracking data from SRI International.[3] This data
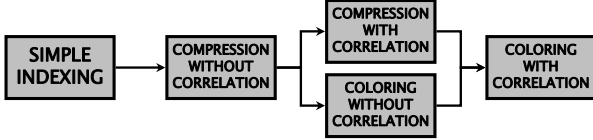

Fig. 8. Levels of compression.

Fig. 9. Empirical probability versus rate.

represents GPS location data. It includes information on various mobiles, including latitude and longitude coordinates. We ignored the other information (e.g., altitude) for the purpose of this simulation. See Figure 7. The blue represents the trajectory through time of one vehicle, and the red curve represents the trajectory of the other.

We focused on these two mobiles, our sources. We assume that our sources are the positional differences (i.e. Δ-encoding), $\mathbf{X}_1$ and $\mathbf{X}_2$, where each is actually a pair, $(\Delta X_{1,\text{LAT}}, \Delta X_{1,\text{LON}})$ and $(\Delta X_{2,\text{LAT}}, \Delta X_{2,\text{LON}})$, of the latitude and longitude data. The use of Δ-encoding assumes that the positional differences form a Markov chain, a common assumption. Our goal is to test the hypothesis that there can be significant encoding gains with even very simple coloring schemes when a function and not the full sources need to be recovered. We consider three relative proximity functions[4] for our analysis:

$$f_{\text{LAT}}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{1}_{|\Delta X_{1,\text{LAT}} - \Delta X_{2,\text{LAT}}| < Z},$$
$$f_{\text{LON}}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{1}_{|\Delta X_{1,\text{LON}} - \Delta X_{2,\text{LON}}| < Z},$$
$$f(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{1}_{\sqrt{(\Delta X_{1,\text{LAT}} - \Delta X_{2,\text{LAT}})^2 + (\Delta X_{1,\text{LON}} - \Delta X_{2,\text{LON}})^2} < Z}.$$

Thus, the functions are 1 when the sources change their relative position by less than Z (along some axis or both), and 0 otherwise. To compare the results of our analysis with current methods, we consider the joint rate $R_1 + R_2$ where $\mathbf{X}_1$ is communicated at rate R_1 and $\mathbf{X}_2$ is communicated at rate R_2. There are several means of rate reduction summarized in Figure 8.

First, the most common (in practice) means of communication is to actually communicate the full index of the value. This means that if $\mathbf{X}_1$ takes M_1 possible values and $\mathbf{X}_2$ takes M_2 possible values, each source will communicate those values using $\log M_1$ bits and $\log M_2$ bits, respectively. Thus, the joint rate is $\log M_1 M_2$. This is clearly inefficient.

Second, we can immediately reduce the rate by compressing each source before communication. Therefore the rate for $\mathbf{X}_1$ would be $H_1 = H(\mathbf{X}_1)$, and the rate for $\mathbf{X}_2$ would be $H_2 = H(\mathbf{X}_2)$. The joint rate would be $H_1 + H_2$. This is strictly better than the first method unless the sources are uniformly distributed.

Third, we can further reduce the rate using correlation, or Slepian-Wolf, encoding. We could use any of the techniques

already developed to achieve near optimal rates, such as DISCUS codes [6] and source-splitting [7]. The joint rate would be $H_{12} = H(\mathbf{X}_1, \mathbf{X}_2)$. This is strictly better than the second method unless the sources are independent.

Fourth, we could use our coloring techniques from Section III-B. If we considered each source communicating its color to the central receiver, the joint rate will be $H_1^\chi + H_2^\chi = H(c_1(\mathbf{X}_1)) + H(c_2(\mathbf{X}_2))$. This may not be better than the above third method, though it certainly will be for independent sources. It will always be better than the second method.

Finally, we could use Slepian-Wolf coding over the colors to achieve a joint rate of $H_{12}^\chi = H(c_1(\mathbf{X}_2), c_2(\mathbf{X}_2))$, which will be strictly better than the third method unless c_1 and c_2 are both injective and strictly better than the fourth method unless $c_1(\mathbf{X}_1)$ and $c_2(\mathbf{X}_2)$ are independent. Thus, the rate relations are as follows:

$$\log M_1 M_2 \geq H_1 + H_2 \geq \begin{matrix} H_{12} \\ H_1^\chi + H_2^\chi \end{matrix} \geq H_{12}^\chi.$$

In our simulations, we test various values of Z to see how the likelihood $p = P[f_{\text{LON}}(\mathbf{X}_1, \mathbf{X}_2) = 1]$, which changes with Z affects the rate reduction.[5] Intuitively, we expect that as p becomes more extreme and approaches either 0 or 1, the rate reduction will become more extreme and approach 100%. (Because if $f_{\text{LON}} = 1$ or $f_{\text{LON}} = 0$ with probability 1, there is nothing to communicate and the rate required is 0. This is shown in Figure 9 where we plot the empirical probability $p = p(Z)$ versus the rate H_{12}^χ.

We expect it would be more symmetric about $1/2$ if we used optimal encoding schemes. However, we are only considering $G = G^1$ (no power graphs) when coloring, as well as a quite simple coloring algorithm. Had we used power graphs, our rate gains would be higher, though the computational complexity would increase exponentially with n. Our coloring algorithm was a simple greedy algorithm that did not use any of the probability information nor was it an ε-coloring. We expect better gains with more advanced graph coloring schemes.

---

[4]We would have liked to use a true proximity function, but then we could not form a valid comparison because our uncolored rate would be in terms of Δ-encoding, but our coloring would necessarily have to be in terms of an encoding of the true position. Therefore, we examine functions that measure how far two mobiles moved towards or away from each other relative to their previous distance, a kind of distance of positional differences.
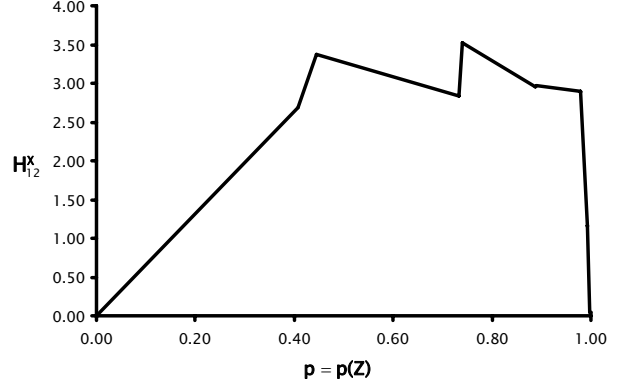
[5]We only show our results for f_{LON} for brevity. The intuition remains for f and f_{LAT}.

TABLE III
EMPIRICAL RESULTS FOR OUR ENCODING SCHEME.

Rates	$\log M_1 M_2$	$H_1 + H_2$	H_{12}	$H_1^\chi + H_2^\chi$	H_{12}^χ
$f(\mathbf{X}_1, \mathbf{X}_2)$	9.32 (0%)	7.93 (15%)	7.78 (17%)	5.44 (42%)	5.29 (43%)
$f_{\text{LAT}}(\mathbf{X}_1, \mathbf{X}_2)$	9.32 (0%)	7.93 (15%)	7.78 (17%)	3.38 (64%)	3.37 (64%)
$f_{\text{LON}}(\mathbf{X}_1, \mathbf{X}_2)$	9.32 (0%)	7.93 (15%)	7.78 (17%)	3.55 (62%)	3.53 (62%)

In Table III, we present the rate results for the various stages of compression in Figure 8. All units are in bits. In the table, we use the values of Z that provide the lowest rate reductions; in other words, we use the worst case rates by testing various Z as in Figure 9. The percentage next to each number shows the percentage decrease in rate. Thus, for the first column, we see 0%, and in the second we see

$$1 - \frac{7.93}{9.32} \approx 0.149 \approx 15\%.$$

We can see that the sources are close to independent, as H_{12} is only slightly smaller than $H_1 + H_2$. Therefore, there is not much gain when considering the correlation between the sources. Nevertheless, the coloring provides a great deal of coding gain. For the simpler f_{LAT} and f_{LON}, the rate has been reduced almost threefold.

This provides evidence that our techniques can indeed lead to large rate gains. For the more simple functions, the rate has been reduced over 60%. Further, considering that the indices are often sent without compression, it is worth noting that even simple compression is 15% better.

V. PROOFS AND ANCILLARY RESULTS

In this section, we provide full proofs of all our previously-stated results.

A. Functional Compression with Side Information

We recall Theorem 13:

$$\lim_{n \to \infty} \frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) = H_G(X|Y). \tag{4}$$

To prove this, we borrow proof techniques from Körner [17], and Orlitsky and Roche [2]. We first state some more typicality results. We use the notion of ε-strong typicality.

Lemma 21. *Suppose $(\mathbf{X}, \mathbf{Y})$ is a sequence of n random variables drawn independently and according to the joint distribution $p(x, y)$, which is the marginal of $p(w, x, y)$. Let an n-sequence $\mathbf{W}$ be drawn independently according to its marginal, $p(w)$. Suppose the joint distribution $p(w, x, y)$ forms a Markov Chain, $W - X - Y$. Then, for all $\varepsilon > 0$, there is a $\varepsilon_1 = k \cdot \varepsilon$, where k depends only on the distribution $p(w, x, y)$, such that for sufficiently large n,*

1) $P[\mathbf{X} \notin T_\varepsilon^n] < \varepsilon_1$, $P[\mathbf{Y} \notin T_\varepsilon^n] < \varepsilon_1$, *and* $P[(\mathbf{X}, \mathbf{Y}) \notin T_\varepsilon^n] < \varepsilon_1$,
2) *For all* $\mathbf{x} \in T_\varepsilon^n$, $P[(\mathbf{x}, \mathbf{W}) \in T_\varepsilon^n] \geq 2^{-n(I(W;X)+\varepsilon_1)}$.
3) *For all* $\mathbf{y} \in T_\varepsilon^n$, $P[(\mathbf{y}, \mathbf{W}) \in T_\varepsilon^n] \leq 2^{-n(I(W;X)-\varepsilon_1)}$.
4) *For all* $(\mathbf{w}, \mathbf{x}) \in T_\varepsilon^n$,

$$P[(\mathbf{w}, \mathbf{Y}) \in T_\varepsilon^n | (\mathbf{x}, \mathbf{Y}) \in T_\varepsilon^n] \geq 1 - \varepsilon_1.$$

Part 1 follows from [13, Lemma 13.6.1], parts 2 and 3 follow from [13, Lemma 13.6.2], and part 4 follows from [13, Lemma 14.8.1].

1) Lower Bound: Consider any n and the corresponding OR-product graph G^n. Let c be an ε-coloring of G^n that achieves $H_{G^n}^\chi(\mathbf{X}|\mathbf{Y})$, i.e.

$$H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) = H(c(\mathbf{X})|Y).$$

As stated earlier, a minimum entropy coloring exists because the set of all colorings on a graph with a finite number of verticies (here, fewer than $|\mathcal{X}|^n$) is finite. With this coloring, we prove that there is a scheme that encodes at rate $\frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y})$ such that the decoder can compute $f(\mathbf{X}, \mathbf{Y})$ with small probability of error. Proving that would establish achievability of the rate $\frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y})$. Theorem 7 proves that no achievable rate can be below $H_G(X|Y)$. This will give us the following lemma.

Lemma 22.

$$\liminf_{n \to \infty} \frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \geq H_G(X|Y).$$

Proof: For any $n > 0$, let c, as above, denote a coloring on G^n that achieves $H_{G^n}^\chi(\mathbf{X}|\mathbf{Y})$, for G the characteristic graph of X with respect to Y, $p(x, y)$, and $f(x, y)$. Let $\Sigma = \{c(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^n\}$, the set of all colors. For every color $\sigma \in \Sigma$ and $\mathbf{y} \in \mathcal{Y}^n$, let $g(\sigma, \mathbf{y}) = f(\bar{\mathbf{x}}, \mathbf{y})$ where $\bar{\mathbf{x}} \in c^{-1}(\sigma) = \{\mathbf{x} : c(\mathbf{x}) = \sigma\}$ with $p(\bar{\mathbf{x}}, \mathbf{y}) > 0$. If no such $\bar{\mathbf{x}}$ exists, g is undefined.

Consider $(\mathbf{x}, \mathbf{y})$ as an instance of the source. Thus, $p(\mathbf{x}, \mathbf{y}) > 0$. Suppose $c(\mathbf{x}) = \sigma$ and $\mathbf{y}$ are available at the decoder where c is defined as above. Then, there is a decoding error when $g(\sigma, \mathbf{y}) \neq f(\mathbf{x}, \mathbf{y})$. This is true only if there exists some $\bar{\mathbf{x}} \in \mathcal{X}^n$ such that $c(\bar{\mathbf{x}}) = \sigma$, $p(\bar{\mathbf{x}}, \mathbf{y}) > 0$, and $f(\bar{\mathbf{x}}, \mathbf{y}) \neq f(\mathbf{x}, \mathbf{y})$. However, because $c(\bar{\mathbf{x}}) = c(\mathbf{x})$, it is true that $(\mathbf{x}, \bar{\mathbf{x}}) \notin E$, where E is the edge set of G^n. Therefore, for all $\bar{\mathbf{y}} \in \mathcal{Y}^n$ with $p(\mathbf{x}, \bar{\mathbf{y}})p(\bar{\mathbf{x}}, \bar{\mathbf{y}}) > 0$, it must be true that $f(\mathbf{x}, \bar{\mathbf{y}}) = f(\bar{\mathbf{x}}, \bar{\mathbf{y}})$. This means $f(\bar{\mathbf{x}}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y})$ for all $\bar{\mathbf{x}} \in c^{-1}(\sigma)$ with $p(\bar{\mathbf{x}}, \mathbf{y}) > 0$. Thus, $g(\sigma, \mathbf{y}) = f(\mathbf{x}, \mathbf{y})$. This shows that the side information along with just the color is sufficient to determine f.

It remains to be seen how to make $c(\mathbf{x})$ and $\mathbf{y}$ available at the decoder. Recall that if a source is encoded at a rate equal to its entropy, it can be recovered to arbitrarily small probability of error at the decoder. Thus, having $\mathbf{Y}$ available at the decoder as side information is the same as encoding $\mathbf{Y}$ at rate greater than $H(Y)$. Recall the Slepian-Wolf Theorem [3] on sources C and Y states that if $R_y > H(Y)$, an encoding with rate $R_c > H(C|Y)$ suffices to recover (C, Y) at the decoder.

We consider our source as $(c(\mathbf{X}), \mathbf{Y})$. Thus, an encoding of rate at least $H(c(\mathbf{X})|\mathbf{Y})$ suffices to recover the functions with arbitrary probability of error. Encoders (and corresponding decoders) exist by the Slepian-Wolf Theorem. Let $\tilde{e} : \Sigma^m \to \{1, \ldots, 2^{mR_c}\}$ be such an encoding with $\tilde{r} : \{1, \ldots, 2^{mR_c}\} \times \mathcal{Y}^{nm} \to \Sigma^m \times \mathcal{Y}^{nm}$ its corresponding decoder. Thus, the idea here is to first color n-blocks of the source. Then, one encodes m-blocks of the colors. The overall rate will be $\frac{1}{m}H(c(\mathbf{X})|\mathbf{Y}))$.

Formally, fix some $n \in \mathbb{N}$. Suppose $\varepsilon > 0$. With an encoder as above, let m be such that

$$\Pr\left[\tilde{r}(\sigma, \mathbf{Y}) \neq (\sigma, \mathbf{Y})\right] < \varepsilon. \tag{5}$$

To show achievability, we need to prove that there exists an encoder $e : \mathcal{X}^{nm} \to \{1, \ldots, 2^{mH(c(\mathbf{X})|Y)}\}$ and a decoder $r : \{1, \ldots, 2^{mH(c(\mathbf{X})|Y)}\} \times \mathcal{Y}^{nm} \to \mathcal{Z}^{nm}$ such that the probability of error is also small:

$$\Pr\left[r(e(\mathbf{X}), \mathbf{Y}) \neq f(\mathbf{X}, \mathbf{Y})\right] < \varepsilon. \tag{6}$$

To prove this, define our encoder as

$$e(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \tilde{e}(c(\mathbf{x}_1), \ldots, c(\mathbf{x}_m)).$$

Then define the decoder as

$$r(e(\mathbf{x}_1, \ldots, \mathbf{x}_m), (\mathbf{y}_1, \ldots, \mathbf{y}_m)) = (g(c(\mathbf{x}_1), \mathbf{y}_1), \ldots, g(c(\mathbf{x}_m), \mathbf{y}_m)),$$

when $\tilde{r}$ correctly recovers the pair $(c(\mathbf{x}_i), \mathbf{y}_i)$ and is undefined otherwise.

The probability that $\tilde{r}$ fails is less than ε by (5). If $\tilde{r}$ does not fail, then, as described earlier, the function will be correctly recovered. Thus, we have shown (6).

Therefore, for any n, the rate $\frac{1}{n}H_{G^n}^\chi(\mathbf{X}|\mathbf{Y})$ is achievable. Thus,

$$\liminf_{n\to\infty} \frac{1}{n}H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \geq H_G(X|Y),$$

completing our proof of the lower bound. $\blacksquare$

2) Upper Bound: Next, we prove that the encoding rate required to recover $c(\mathbf{X})$ given $\mathbf{Y}$ is at most $H_G(X|Y)$:

Lemma 23.

$$\limsup_{n\to\infty} \frac{1}{n}H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \leq H_G(X|Y).$$

Proof: Suppose $\varepsilon > 0$, $\delta > 0$. Suppose n is (sufficiently large) such that: (1) Lemma 21 applies with some $\varepsilon_1 < 1$, (2) $2^{-n\delta} < \varepsilon_1$, and (3) $n > 2 + \frac{3}{2\varepsilon_1}$.

Let $p(w, x, y)$ be the distribution that achieves the $H_G(X|Y)$ with the Markov property $W - X - Y$. (This is guaranteed to exist by Theorem 7.) Denote by $p(w)$, $p(x)$, and $p(y)$ the marginal distributions. For any integer M, define an M-*system* $(\mathbf{W}_1, \ldots, \mathbf{W}_M)$ where each $\mathbf{W}_i$ is drawn independently with distribution $p(\mathbf{w}) = \prod_{i=1}^n p(w_i)$.

Our encoding scheme will declare an error if $(\mathbf{x}, \mathbf{y}) \notin T_\varepsilon^n$. This means that the encoder will code over ε_1 colorings of the characteristic graphs. By construction, this error happens with probability less than ε_1. Henceforth assume that $(\mathbf{x}, \mathbf{y}) \in T_\varepsilon^n$.

Next, our encoder will declare an error when there is no i such that $(\mathbf{W}_i, \mathbf{x}) \in T_\varepsilon^n$. This occurs with probability

$$\Pr[(\mathbf{W}_i, \mathbf{x}) \notin T_\varepsilon^n \forall i] \overset{(a)}{\leq} \prod_{i=1}^M \Pr[(\mathbf{W}_i, \mathbf{x}) \notin T_\varepsilon^n]$$
$$\overset{(b)}{=} (1 - \Pr[(\mathbf{W}, \mathbf{x}) \in T_\varepsilon^n])^M$$
$$\overset{(c)}{\leq} \left(1 - 2^{-n(I(W;X)+\varepsilon_1)}\right)^M$$
$$\overset{(d)}{\leq} 2^{-M \cdot 2^{-n(I(W;X)+\varepsilon_1)}}$$

where (a) and (b) follow because the $\mathbf{W}_i$ are independent and identically distributed, (c) follows from Lemma 21 part 2, and (d) follows because for $\alpha \in [0, 1]$, $(1-\alpha)^n \leq 2^{-n\alpha}$. Assuming $M > 2^{n(I(W;X)+\varepsilon_1+\delta)}$,

$$\Pr[(\mathbf{W}_i, \mathbf{x}) \notin T_\varepsilon^n \forall i] \leq 2^{-\delta n} < \varepsilon_1,$$

because n is large enough such that the final inequality holds. Henceforth, fix an M-system $(\mathbf{W}_1, \ldots, \mathbf{W}_M)$ for some $M > 2^{n(I(W;X)+\varepsilon_1+\delta)}$. Further, assume there is some i such that $(\mathbf{W}_i, \mathbf{x}) \in T_\varepsilon^n$.

For each $\mathbf{x}$, let the smallest (or any) such i be denoted as $\hat{c}(\mathbf{x})$. Note that $\hat{c}$ is an ε_1-coloring of the graph G^n. For each $\mathbf{y}$, define:

$$S(\mathbf{y}) = \{\hat{c}(\mathbf{x}) : (\mathbf{x}, \mathbf{y}) \in T_\varepsilon^n\},$$
$$Z(\mathbf{y}) = \{\mathbf{W}_i : (\mathbf{W}_i, \mathbf{y}) \in T_\varepsilon^n\},$$
$$s(\mathbf{y}) = |S(\mathbf{y})|,$$
$$z(\mathbf{y}) = |Z(\mathbf{y})|.$$

Then, $s(\mathbf{y}) = \sum_{i=1}^M \mathbf{1}_{i \in S(\mathbf{y})}$, because our coloring scheme $\hat{c}$ is simply an assignment of the indices of the M-system. Thus, we know

$$E[s(\mathbf{Y})] = \sum_{i=1}^M \Pr[i \in S(\mathbf{Y})].$$

Similarly, we get $z(\mathbf{y}) = \sum_{i=1}^M \mathbf{1}_{\mathbf{W}_i \in Z(\mathbf{y})}$. Thus,

$$E[z(\mathbf{Y})] = \sum_{i=1}^M \Pr[\mathbf{W}_i \in Z(\mathbf{Y})]$$
$$\geq \sum_{i=1}^M \Pr[\mathbf{W}_i \in Z(\mathbf{Y}) \text{ and } i \in S(\mathbf{Y})]$$
$$= \sum_{i=1}^M \Pr[i \in S(\mathbf{Y})]P[\mathbf{W}_i \in Z(\mathbf{Y})|i \in S(\mathbf{Y})]$$

We know that if $i \in S(\mathbf{Y})$, there is some $\mathbf{x}$ such that $\hat{c}(\mathbf{x}) = i$ and $(\mathbf{x}, \mathbf{Y}) \in T_\varepsilon^n$. For each such $\mathbf{x}$, we must have (by definition of our coloring), $(\mathbf{W}_i, \mathbf{x}) \in T_\varepsilon^n$. For each such $\mathbf{x}$ where $(\mathbf{W}_i, \mathbf{x}) \in T_\varepsilon^n$,

$$\Pr[\mathbf{W}_i \in Z(\mathbf{Y})|i \in S(\mathbf{Y})] = \Pr[(\mathbf{W}_i, \mathbf{Y}) \in T_\varepsilon^n|(\mathbf{x}, \mathbf{Y}) \in T_\varepsilon^n]$$
$$\geq 1 - \varepsilon_1$$

by Lemma 21 part 4. Thus, we have $E[z(\mathbf{Y})] \geq (1 - \varepsilon_1)E[s(\mathbf{Y})]$. This, along with Jensen's inequality, imply

$$E[\log s(\mathbf{Y})] \leq \log E[s(\mathbf{Y})]$$
$$\leq \log E\left[\frac{z(\mathbf{Y})}{1 - \varepsilon_1}\right].$$

Finally, (using a Taylor series expansion) we know $\log \frac{1}{1-\varepsilon_1} \leq \varepsilon_2 = 2\varepsilon_1 + \frac{1}{2}$ when $0 < \varepsilon_1 < 1$. Thus,

$$E[\log s(\mathbf{Y})] \leq \log E[z(\mathbf{Y})] + \varepsilon_2, \quad (7)$$

We compute

$$E[z(\mathbf{Y})] = \sum_{i=1}^{M} P[(\mathbf{W}_i, \mathbf{Y}) \in T_\varepsilon^n]$$
$$= M \cdot P[(\mathbf{W}_i, \mathbf{Y}) \in T_\varepsilon^n]$$

because the $\mathbf{W}_i$ are i.i.d. Therefore,

$$E[z(\mathbf{Y})] \leq M \cdot 2^{-n(I(W;Y)-\varepsilon_1)} \quad (8)$$

by Lemma 21 part 3.

By the definition of $S(\mathbf{y})$, we know that determining $\hat{c}$ given $\mathbf{Y} = \mathbf{y}$ requires at most $\log s(\mathbf{y})$ bits. Therefore, we have

$$H(\hat{c}(\mathbf{X})|\mathbf{Y}) \leq E[\log s(\mathbf{Y})]. \quad (9)$$

Putting it all together, we have

$$H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \overset{(a)}{\leq} H(\hat{c}(\mathbf{X})|\mathbf{Y})$$
$$\overset{(b)}{\leq} E[\log s(\mathbf{Y})]$$
$$\overset{(c)}{\leq} \log E[z(\mathbf{Y})] + \varepsilon_2$$
$$\overset{(d)}{\leq} \log \left(M \cdot 2^{-n(I(W;Y)-\varepsilon_1)}\right) + \varepsilon_2$$
$$\overset{(e)}{=} \log \left(2^{n(I(W;X)-I(W;Y)+2\varepsilon_1+\delta)} + 1\right) + \varepsilon_2$$
$$\overset{(f)}{\leq} n(I(W;X) - I(W;Y) + 2\varepsilon_1 + \delta) + 1 + \varepsilon_2$$

where (a) follows by definition of the conditional chromatic entropy, (b) follows from inequality (9), (c) follows from inequality (7), (d) follows from inequality (8), (e) follows by setting $M = \lceil 2^{n(I(W;X)+\varepsilon_1+\delta)} \rceil$, and (f) follows because $\log(\alpha + 1) \leq \log(\alpha) + 1$ for $\alpha \geq 1$.

For Markov chains $W - X - Y$,

$$I(W;X) - I(W;Y) = I(W;X|Y).$$

Thus, for our optimal distribution $p(w, x, y)$, we have

$$H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \leq n(H_G(X|Y) + 2\varepsilon_1 + \delta) + 1 + \varepsilon_2$$

Because $n > 2 + \frac{3}{2\varepsilon_1}$, $\frac{1+\varepsilon_2}{n} < \varepsilon_1$. Thus, $\frac{1}{n}H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \leq H_G(X|Y) + 3\varepsilon_1 + \delta$. This completes the proof for the upper bound:

$$\limsup_{n\to\infty} \frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) \leq H_G(X|Y).$$

The lower and upper bounds, Lemmas 22 and 23, combine to give Theorem 13:

$$\lim_{n\to\infty} \frac{1}{n} H_{G^n}^\chi(\mathbf{X}|\mathbf{Y}) = H_G(X|Y).$$

$\square$

B. Distributed Functional Compression

Recall Theorem 16 states that the achievable rate region for the distributed functional compression problem, under the Zigzag Condition (Condition 15), is the set closure of the set of all rates that can be realized via graph coloring.

We prove this by first showing that if the colors are available at the decoder, the decoder can successfully compute the function. This proves achievability. Next, we show that all valid encodings are ε-colorings of the characteristic graphs (and their powers). This establishes the converse.

1) Achievability: We first prove the achievability of all rates in the region given in the theorem statement.

Lemma 24. *For sufficiently large n and ε-colorings c_x and c_y of G_x^n and G_y^n, respectively, there exists*

$$\hat{f} : c_x(\mathcal{X}^n) \times c_y(\mathcal{Y}^n) \to \mathcal{Z}^n$$

such that $\hat{f}(c_x(\mathbf{x}), c_y(\mathbf{y})) = f(\mathbf{x}, \mathbf{y})$ for all $(\mathbf{x}, \mathbf{y}) \in T_\varepsilon^n$.

Proof: Suppose $(\mathbf{x}, \mathbf{y}) \in T_\varepsilon^n$, and we have colorings c_x and c_y. We proceed by constructing $\hat{f}$. For any two colors $\gamma \in c_x(\mathcal{X}^n)$ and $\sigma \in c_y(\mathcal{Y}^n)$, let $\hat{\mathbf{x}} \in c_x^{-1}(\gamma)$ and $\hat{\mathbf{y}} \in c_y^{-1}(\sigma)$ be any (say the first) pair such that $p(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in T_\varepsilon^n$. Define $\hat{f}(\gamma, \sigma) = f(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. There must be such a pair because certainly $(\mathbf{x}, \mathbf{y})$ qualifies.

To show this function is well-defined on elements in the support, suppose $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are both in T_ε^n. Suppose further that $c_x(\mathbf{x}_1) = c_x(\mathbf{x}_2)$ and $c_y(\mathbf{y}_1) = c_y(\mathbf{y}_2)$. Then, we know that there is no edge $(\mathbf{x}_1, \mathbf{x}_2)$ in the high probability subgraph of G_x^n or $(\mathbf{y}_1, \mathbf{y}_2)$ in the edge set of the high probability subgraph of G_y^n by definition of graph coloring.

By the Zigzag Condition, there exists some $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ such that $(\tilde{\mathbf{x}}, \mathbf{y}_1), (\tilde{\mathbf{x}}, \mathbf{y}_2), (\mathbf{x}_1, \tilde{\mathbf{y}}), (\mathbf{x}_2, \tilde{\mathbf{y}}) \in T_{\frac{\varepsilon}{2}}^n$. We claim that there is no edge between $(\mathbf{x}_i, \tilde{\mathbf{x}})$ or $(\mathbf{y}_i, \tilde{\mathbf{y}})$ for either i. We prove this for $(\mathbf{x}_1, \tilde{\mathbf{x}})$ with the other cases following naturally. Suppose there were an edge. Thus, there is some $\hat{\mathbf{y}}$ such that $f(\mathbf{x}_1, \hat{\mathbf{y}}) \neq f(\tilde{\mathbf{x}}, \hat{\mathbf{y}})$. This implies that $f(x_{1j}, \hat{y}_j) \neq f(\tilde{x}_j, \hat{y}_j)$ for some j. Define $\tilde{\mathbf{y}}'$ as $\tilde{\mathbf{y}}$ in every component but the j-th, where it is $\hat{y}_j$.

We know that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$|\nu_{(\mathbf{x}_1, \tilde{\mathbf{y}})}(x, y) - p(x, y)| \leq \frac{\varepsilon}{2|\mathcal{X}||\mathcal{Y}|}$$

by definition of $\frac{\varepsilon}{2}$-typicality. Therefore,

$$|\nu_{(\mathbf{x}_1, \tilde{\mathbf{y}}')}(x, y) - p(x, y)| \leq \frac{\varepsilon}{2|\mathcal{X}||\mathcal{Y}|}$$

for all (x, y) such that $y \neq \hat{y}_j$ and $y \neq \tilde{y}_j$.

Next, we can choose n large enough such that $n > \frac{2|\mathcal{X}||\mathcal{Y}|}{\varepsilon}$. Then, for $y = \hat{y}_j$ or $y = \tilde{y}_j$, the empirical frequency changes

by at most $\frac{1}{n}$. Thus, for all (x,y) (including $y = \hat{y}_j$ and $y = \tilde{y}_j$), we have

$$|\nu_{(\mathbf{x}_1, \tilde{\mathbf{y}}')}(x,y) - p(x,y)| \leq \frac{\varepsilon}{2|\mathcal{X}||\mathcal{Y}|} + \frac{1}{n} \leq \frac{\varepsilon}{|\mathcal{X}||\mathcal{Y}|}$$

Thus, $\tilde{\mathbf{y}}'$ is ε-typical with both $\mathbf{x}_1$ and $\mathbf{x}_2$. By construction, $f(\mathbf{x}_1, \tilde{\mathbf{y}}') \neq f(\mathbf{x}_2, \tilde{\mathbf{y}}')$. Therefore, there must be an edge in the high probability subgraph between $(\mathbf{x}_1, \mathbf{x}_2)$, an impossibility. Thus, there is no edge $(\mathbf{x}_1, \tilde{\mathbf{x}})$. The others follow similarly.

Thus, by definition of the graph,

$$f(\mathbf{x}_1, \mathbf{y}_1) = f(\tilde{\mathbf{x}}, \mathbf{y}_1) = f(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = f(\mathbf{x}_2, \tilde{\mathbf{y}}) = f(\mathbf{x}_2, \mathbf{y}_2).$$

Therefore, our function $\hat{f}$ is well-defined and has the desired property. ∎

Then, Lemma 24 implies that we can successfully compute our function f given colors of the characteristic graphs. Thus, if the decoder is given colors, it can look up f based on its table of $\hat{f}$. The question is now of faithfully (with probability of error less than ε) transmitting these colors to the receiver. However, when we consider the colors as sources, we know the achievable rates.

Lemma 25. *For any n, ε-colorings c_x and c_y of G_x^n and G_y^n, respectively, the achievable rate region for joint source $(c_x(\mathbf{X}), c_y(\mathbf{Y}))$ is the set of all rates, (R_x^c, R_y^c), satisfying:*

$$R_x^c \geq H(c_x(\mathbf{X})|c_y(\mathbf{Y})),$$
$$R_y^c \geq H(c_y(\mathbf{Y})|c_x(\mathbf{X})),$$
$$R_x^c + R_y^c \geq H(c_x(\mathbf{X}), c_y(\mathbf{Y})).$$

Proof: This follows directly from the Slepian-Wolf Theorem [3] for the separate encoding of correlated sources. ∎

Suppose the probability of decoder error for the decoder guaranteed in Lemma 25 is less than $\frac{\varepsilon}{2}$. Then the total error in the coding scheme of first coloring G_x^n and G_y^n, and then encoding those colors to be faithfully decoded at the decoder is upper bounded by the sum of the errors in each stage. Thus, Lemmas 24 and 25 together to show that the probability that the decoder errs is less than ε for any ε provided large enough n (and block size m on the colors).

Finally, in light of the fact that n source symbols are encoded for each color, the achievable rate region for the problem under the Zigzag Condition is the set of all rates (R_x, R_y) such that

$$R_x \geq \frac{1}{n}H(c_x^n(\mathbf{X})|c_y^n(\mathbf{Y})),$$
$$R_y \geq \frac{1}{n}H(c_y^n(\mathbf{Y})|c_x^n(\mathbf{X})),$$
$$R_x + R_y \geq \frac{1}{n}H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y})).$$

where c_x^n and c_y^n are achievable ε-colorings (for any $\varepsilon > 0$). Thus every $(R_x, R_y) \in \mathcal{S}^\varepsilon$ is achievable for all $\varepsilon > 0$. Therefore, every $(R_x, R_y) \in \mathcal{S}$ is achievable.

2) Converse: Next, we prove that any distributed functional source code with small probability of error induces a coloring.

Suppose $\varepsilon > 0$. Define for all (n, ε),

$$\mathcal{F}_\varepsilon^n = \{\hat{f} : \Pr[\hat{f}(\mathbf{X}, \mathbf{Y}) \neq f(\mathbf{X}, \mathbf{Y})] < \varepsilon\}.$$

This is the set of all functions that equal f to within ε probability of error. (Note that all achievable distributed functional source codes are in $\mathcal{F}_\varepsilon^n$ for large enough n.)

Lemma 26. *Consider some function $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Any distributed functional code that reconstructs g with zero-error (with respect to a distribution $p(x,y)$) induces colorings on the characteristic graphs of X and Y with respect to g, $p(x,y)$, and Y and X, respectively.*

Proof: Suppose we have encoders e_x and e_y, decoder d, and characteristic graphs G_x^n and G_y^n. Then by definitions, a zero-error reconstruction implies that for any $(\mathbf{x}_1, \mathbf{y}_2)$, $(\mathbf{x}_2, \mathbf{y}_2)$ such that if $p(\mathbf{x}_1, \mathbf{y}_1) > 0$, $p(\mathbf{x}_2, \mathbf{y}_2) > 0$, $e_x(\mathbf{x}_1) = e_x(\mathbf{x}_2)$, and $e_y(\mathbf{y}_1) = e_y(\mathbf{y}_2)$, then

$$f(\mathbf{x}_1, \mathbf{y}_1) = f(\mathbf{x}_2, \mathbf{y}_2) = r(e_x(\mathbf{x}_1), e_y(\mathbf{y}_1)). \quad (10)$$

We now show that e_x and e_y are valid colorings of G_x^n and G_y^n. We demonstrate the argument for X. The argument for Y is analogous. We proceed by contradiction. If it were not true, then there must be some edge with both vertices with the same color. In other words, there must exist $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ such that $p(\mathbf{x}_1, \mathbf{y})p(\mathbf{x}_2, \mathbf{y}) > 0$, $e_x(\mathbf{x}_1) = e_x(\mathbf{x}_2)$, and $f(\mathbf{x}_1, \mathbf{y}) \neq f(\mathbf{x}_2, \mathbf{y})$. This is impossible (by taking $\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y}$ in equation (10)). Hence, we have induced colorings of the characteristic graphs. ∎

We now show that any achievable distributed functional code also induces an ε-coloring of the characteristic graphs.

Lemma 27. *All achievable distributed functional codes induce ε-colorings of the characteristic graphs.*

Proof: Let $g(\mathbf{x}, \mathbf{y}) = r(e_x(\mathbf{x}), e_y(\mathbf{y})) \in \mathcal{F}_\varepsilon^n$ be such a code. Then, we know that a zero-error reconstruction (with respect to p) of g induces colorings, e_x and e_y, of the characteristic graphs with respect to g and p by Lemma 26. Let the set of all $(\mathbf{x}, \mathbf{y})$ such that $g(\mathbf{x}, \mathbf{y}) \neq f(\mathbf{x}, \mathbf{y})$ be denoted as $\mathcal{C}$. Then because $g \in \mathcal{F}_\varepsilon^n$, we know that $\Pr[\mathcal{C}] < \varepsilon$. Therefore, the functions e_x and e_y restricted to $\mathcal{C}$ are ε-colorings of G_x and G_y (by definition). ∎

Thus, the Lemma 26 and Lemma 27 establish Theorem 16 in full.

3) Minimal Joint Rate: Recall Corollary 17 states that under the zigzag condition, when there is a unique point achieving the minimum joint rate, it must be $R_x + R_y = H_{G_x}(X) + H_{G_y}(Y)$.

Proof: First, we recall the rate pair $(H_{G_x}(X), H_{G_y}(Y))$ can be achieved via graph colorings. This is true by the achievability result of Theorem 16 along with Theorem 12, which states that graph colorings can achieve each of $H_{G_x}(X)$ and $H_{G_y}(Y)$. In the achievability proof above, we showed that, under the zigzag condition, any coloring scheme will lead to achievable rates. Therefore, $(H_{G_x}(X), H_{G_y}(Y))$ is in the rate region. (Note, that we have not yet used the uniqueness of the minimum.)

Suppose (R_x, R_y) achieves the minimum joint rate. By Theorem 16, this must be in some Slepian-Wolf region for the colors. Because it is a minimum, we must have $R_x + R_y = \frac{1}{n}H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y}))$. This can be achieved with

$R_x = \frac{1}{n}H(c_x^n(\mathbf{X}))$ and $R_y = \frac{1}{n}H(c_y^n(\mathbf{Y})|c_x^n(\mathbf{X}))$ or with $R_x = \frac{1}{n}H(c_x^n(\mathbf{X})|c_y^n(\mathbf{Y}))$ and $R_y = \frac{1}{n}H(c_y^n(\mathbf{Y}))$.

By assumption, there is only one such point, thus we must have $\frac{1}{n}I(c_x^n(\mathbf{X}); c_y^n(\mathbf{Y})) \to 0$ as $n \to \infty$. Thus, the minimal rate is $\frac{1}{n}H(c_x^n(\mathbf{X})) + \frac{1}{n}H(c_y^n(\mathbf{Y})) \to R_x + R_y$ as $n \to \infty$. We know for all n, $H_{G_x}(X) + H_{G_y}(Y) \leq \frac{1}{n}H(c_x^n(\mathbf{X})) + \frac{1}{n}H(c_y^n(\mathbf{Y}))$ by Theorem 12.

Therefore, we must have that the minimum achievable joint rate is $H_{G_x}(X) + H_{G_y}(Y)$. ∎

This corollary implies that minimum entropy colorings have decreasing mutual information as n increases. Thus, the closer we are to the optimum via graph coloring, the less complicated our Slepian-Wolf codes must be. In the limit, because mutual information is zero, each source only needs to code to entropy. Thus, the Slepian-Wolf codes are unnecessary when achieving the minimal joint rate. (Nevertheless, finding the minimum entropy colorings is, again, NP-hard.)

Next in Theorem 18, we consider the case when the minimum is not uniquely achievable.

Proof: The joint rate must always satisfy:

$$
\begin{aligned}
R_x + R_y &= \frac{1}{n}H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y})) \\
&= \frac{1}{n}H(c_x^n(\mathbf{X})) + \frac{1}{n}H(c_y^n(\mathbf{Y})|c_x^n(\mathbf{X})) \\
&\geq H_{G_x}(X) + \frac{1}{n}H(c_y^n(\mathbf{Y})|\mathbf{X}) \\
&\geq H_{G_x}(X) + H_{G_y}(Y|X)
\end{aligned}
$$

The first inequality follows from the Data Processing Inequality on the Markov chain $c_y^n(\mathbf{Y}) - \mathbf{X} - c_x^n(\mathbf{X})$, and the second follows by definition of the conditional graph entropy. Similarly, we get:

$$
\begin{aligned}
R_x + R_y &= \frac{1}{n}H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y})) \\
&= \frac{1}{n}H(c_x^n(\mathbf{X})|c_y^n(\mathbf{Y})) + \frac{1}{n}H(c_y^n(\mathbf{Y})) \\
&\geq H_{G_x}(X|Y) + H_{G_y}(Y)
\end{aligned}
$$

Thus, the difference between the optimal rate $(R_x + R_y)$, and the rate given in Corollary 17 is bounded by the following two inequalities:

$$
\left[H_{G_x}(X) + H_{G_y}(Y)\right] - \left[R_x + R_y\right] \leq H_{G_x}(X) - H_{G_x}(X|Y)
$$
$$
\left[H_{G_x}(X) + H_{G_y}(Y)\right] - \left[R_x + R_y\right] \leq H_{G_y}(Y) - H_{G_y}(Y|X)
$$

∎

C. Functional Rate Distortion

In this section, we prove Theorem 19 and Corollary 20 for the functional rate distortion problem.

We restate Theorem 19 for completeness:

$$
R(D) = \min_{\hat{f} \in \mathcal{F}(D)} H_{G(\hat{f})}(X|Y)
$$

Proof: We prove that the given characterization is valid by first showing the rate $H_{G(\hat{f})}(X|Y)$ is achievable for any $\hat{f} \in \mathcal{F}(D)$, and next showing that every achievability scheme must be in $\mathcal{F}(D)$.

By Orlitsky and Roche, we know that the rate $H_{G(\hat{f})}(X|Y)$ is sufficient to determine the function $\hat{f}(\mathbf{X}, \mathbf{Y})$ at the receiver. By definition,

$$
\lim_{n \to \infty} E[d(f(\mathbf{X}, \mathbf{Y}), \hat{f}(\mathbf{X}, \mathbf{Y}))] \leq D.
$$

Thus, the rate $H_{G(\hat{f})}(X|Y)$ is achievable.

Next, suppose we have any achievable rate R, with corresponding *sequence* of encoding and decoding functions e_1^n and e_2^n respectively. Then the function $\hat{f}(\cdot, \cdot) = e_2^n(e_1^n(\cdot), \cdot)$ is a function $\hat{f} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}^n$ with the property (by achievability) that $\lim_{n \to \infty} E[d(f(\mathbf{X}, \mathbf{Y}), \hat{f}(\mathbf{X}, \mathbf{Y}))] \leq D$ (again because as $n \to \infty$, ε is driven to 0). Thus, $\hat{f} \in \mathcal{F}(D)$, completing the proof of Theorem 19. ∎

Next we prove Corollary 20, which states that $H_{G^D}(X|Y)$ is an achievable rate. We show this by demonstrating that any distribution on (W, X, Y) satisfying $W - X - Y$ and $X \in W \in \Gamma(G^D)$ also satisfies the Yamamoto requirement (i.e. is also in $\mathcal{P}(D)$).

Proof: Suppose $p(w, x, y)$ is such that $p(w, x, y) = p(w|x)p(x, y)$, or $W - X - Y$ is a Markov chain. Further suppose that $X \in W \in \Gamma(G^D)$. Then define $g(w, y) = f(x^*, y)$ where x^* is any (say, the first) $x \in w$ with $p(x^*, y) > 0$. This is well-defined because the nonexistence of x such that $p(x, y) > 0$ is a zero probability event, and $x \in w$ occurs with probability one by assumption.

Further, because w is an independent set, for any $x_1, x_2 \in w$, one must have $(x_1, x_2) \notin E^D$, the edge set of G^D. By definition of G^D, this means that for all $y \in \mathcal{Y}$ such that $p(x_1, y)p(x_2, y) > 0$, it must be the case that $d(f(x_1, y), f(x_2, y)) \leq D$. Therefore,

$$
E[d(f(X, Y), g(W, Y))] = E[d(f(X, Y), f(X^*, Y))] \leq D
$$

because both $X \in W$ and $X^* \in W$ are probability 1 events.

We have shown that for a given distribution achieving the conditional graph entropy, there is a function g on $\mathcal{W} \times \mathcal{Y}$ that has expected distortion less than D. In other words, any distribution satisfying $W - X - Y$ and $X \in W \in \Gamma(G^D)$ is also in $\mathcal{P}(D)$. Further, any such distribution can be associated with a coding scheme, by Orlitsky and Roche's work [2], that achieves the rate $I(W; X|Y)$. When the distribution is chosen such that $I(W; X|Y)$ is minimized, this is by definition equal to $H_{G^D}(X|Y)$. Thus, the rate $H_{G^D}(X|Y)$ is achievable, proving Corollary 20 and providing a single-letter upper bound for $R(D)$. ∎

VI. Conclusion

This article considered the problem of coding for computing in new contexts. We considered the functional compression problem with side information and gave novel solutions for both the zero and nonzero distortion cases. These algorithms gave an explicit decoupling of the computing from the correlation between the sources as a graph coloring problem. We proved that this decoupling is rate optimal. We extended this encoding scheme to the distributed functional compression with zero distortion. We gave an inner bound to the rate region, and gave the conditions under which the decoupling is optimal in the distributed case.

We never considered the nonzero distortion distributed functional compression problem mainly because even the case of $f(x,y) = (x,y)$ is unsolved. Nevertheless, it is our hope that the methods discussed in this article will yield new results for the more general problem.

All of our results concern two sources. An extension of these results to M sources seems plausible. However, the graph constructs used rely heavily on the two source structure and would need to be modified to deal with M sources. We leave that to future work.

Finally, we examined the applicability of our results. For Blue Force Tracking, we saw that even simple coloring schemes yielded large compression gains (64%).

In summary, this article is about modeling the distillation of relevant information from disparate sources. We hope the work presented herein serves as a step towards more research in this area.

REFERENCES

[1] C. E. Shannon and W. Weaver, *A Mathematical Theory of Communication*. Champaign, IL, USA: University of Illinois Press, 1963.

[2] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar. 2001.

[3] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[4] R. F. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, Nov. 1975.

[5] J. Körner and K. Marton, "How to encode the modulo-two sum of binary sources," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 219–221, Mar. 1979.

[6] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[7] T. P. Coleman, A. H. Lee, M. Médard, and M. Effros, "Low-complexity approaches to Slepian-Wolf near-lossless distributed data compression," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3546–3561, Aug. 2006.

[8] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[9] T. Berger and R. W. Yeung, "Multiterminal source encoding with one distortion criterion," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 228–236, Mar. 1989.

[10] J. Barros and S. Servetto, "On the rate-distortion region for separate encoding of correlated sources," IEEE Symposium on Information Theory (ISIT), Yokohama, Japan, 2003, p. 171.

[11] A. B. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic gaussian two-terminal source-coding problem," *IEEE Trans. Inf. Theory*, submitted February 2006.

[12] H. Feng, M. Effros, and S. Savari, "Functional source coding for networks with receiver side information," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Sep. 2004, pp. 1419–1427.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[14] H. Yamamoto, "Wyner-Ziv theory for a general function of the correlated sources," *IEEE Trans. Inf. Theory*, vol. 28, no. 5, pp. 803–807, Sep. 1982.

[15] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[16] H. S. Witsenhausen, "The zero-error side information problem and chromatic numbers," *IEEE Trans. Inf. Theory*, vol. 22, no. 5, pp. 592–593, Sep. 1976.

[17] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," 6th Prague Conference on Information Theory, 1973, pp. 411–425.

[18] N. Alon and A. Orlitsky, "Source coding and graph entropies," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1329–1339, Sep. 1996.

[19] J. Cardinal, S. Fiorini, and G. V. Assche, "On minimum entropy graph colorings," ISIT 2004, Jun.–Jul. 2004, p. 43.

[20] G. Campers, O. Henkes, and J. P. Leclerq, "Graph coloring heuristics: A survey, some new propositions and computational experiences on random and "Leighton's" graphs," in *Operations Research '87*, G. K. Rand, Ed., 1988, pp. 917–932.

[21] C. McDiarmid, "Colourings of random graphs," in *Graph Colourings*, ser. Pitman Research Notes in Mathematics Series, R. Nelson and R. J. Wilson, Eds. Longman Scientific & Technical, 1990, pp. 79–86.

[22] Q. Zhao and M. Effros, "Lossless and near-lossless source coding for multiple access networks." *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 112–128, Aug. 2003.