

# A Relay-Assisted Distributed Source Coding Problem

Prakash Ishwar

Dept. Electrical and Computer Engineering  
Boston University  
Boston, MA 02215, USA  
Email: pi@bu.edu

S. Sandeep Pradhan

Dept. Electrical Engineering and Computer Science  
University of Michigan  
Ann Arbor, MI 48109, USA  
Email: pradhanv@eecs.umich.edu

**Abstract**—A relay-assisted distributed source coding problem with three statistically correlated sources is formulated and studied. Natural lower bounds for the rates are presented. The lower bounds are shown to be achievable for three special classes of joint source distributions. The achievable coding strategies for the three special classes are observed to have markedly different characteristics. A special class for which the lower bounds are not achievable is presented. A single unified coding strategy which subsumes all four special classes is derived.

## I. INTRODUCTION

Consider the following distributed source coding problem which is illustrated in Fig. 1.  $n$  samples  $X^n := (X_1, \dots, X_n)$  of an information source available at location- $a$  need to be reproduced at location- $c$  with high probability. The message from location- $a$ , of bit rate  $R_a$  bits per source sample, is relayed through an intermediate location- $b$  which has access to  $n$  samples  $Y^n$  of a second information source which are statistically correlated to  $X^n$ . The message from location- $b$  to location- $c$  is of bit rate  $R_b$  bits per source sample. Location- $c$  has access to  $n$  samples  $Z^n$  of a third information source which are correlated to  $X^n$  and  $Y^n$ . The reproduction at location- $c$  is  $\hat{X}^n$ . The goal is to characterize the set of minimal rate pairs  $(R_a, R_b)$  for which  $\mathbb{P}(X^n \neq \hat{X}^n) \downarrow 0$  as  $n \uparrow \infty$  (the rate region).

The problem just described belongs to the general class of source network problems [1, Chapter 3] (also called distributed source coding problems) which have been studied in the Information Theory literature in the 70's and 80's. Beginning the late 90's, the construction of practical distributed source codes for emerging wireless sensor network applications [2] received much attention and renewed interest in the study of general source network problems. For instance, the three sources in Fig. 1 can represent the correlated observation samples of three physically separated, radio-equipped, sensing devices which are measuring some physical quantities of interest in the vicinity of their respective locations. In some situations, the device at location- $c$  may need to learn about the observations at location- $a$  but is uninterested in the observations of an intermediate location- $b$  whose connectivity with both location- $a$  and location- $c$  is orders of magnitude stronger than that between location- $a$  and location- $c$ . Depending on the relative strengths of the connections between location- $a$  and location- $b$  and between location- $b$  and location- $c$ , it may be necessary to

use codes of different rates on these links to achieve the desired objective of reproducing  $X^n$  at location- $c$ . This motivates the characterization of the entire rate region of minimal rate pairs  $(R_a, R_b)$  for which  $\mathbb{P}(X^n \neq \hat{X}^n)$  can be driven down to zero as  $n \uparrow \infty$ . More recently, the study of general source networks has also benefited from insights which are emerging from the related area of network coding [3].

In the terminology of [1, p. 246], the source network of Fig. 1 has three input vertices corresponding to the three sources, one output vertex corresponding to the decoder at location- $c$ , and two intermediate vertices corresponding to the encoders at locations  $a$  and  $b$ . The depth of this network which is the length of the longest path from an input to an output is 3. The structure of this network is significantly different from that of several related source network problems of depth 2 studied in the 70's and 80's whose rate regions are fairly well understood (see [4], [1, Chapter 3], and references therein). In more recent terminology, the specific source network of Fig. 1 would be the simplest example of the so-called single-demand line source networks (see for example [5], [6]).

Despite their apparent simplicity, the general rate regions of line source networks are not well understood except in very special cases when the joint probability distribution of the sources and/or the fidelity criteria for source reproductions have some special properties. For instance, inner and (non-matching) outer bounds for the rate-distortion region for a slightly more general class of tree source networks are derived in [5]. However, it is assumed that there exists an underlying hidden source conditioned on which all the source observations become independent. Moreover, all locations need to reproduce the hidden source to within some fidelity as measured by a single-letter distortion function. A recent closely related work [7] studies the relay-assisted distributed source coding problem of Fig. 1 but when  $X^n-Z^n-Y^n$  forms a Markov chain (degraded side information) and when  $X^n$  needs to be reproduced at location- $b$  and location- $c$  only to within expected distortion levels as measured by associated single-letter fidelity criteria. Inner and (non matching) outer bounds for the rate-distortion region are derived and some special cases are discussed. In the context of asymptotically lossless reconstruction of  $X^n$  at (only) location- $c$ , the condition  $X^n-Z^n-Y^n$  corresponds to the special case 2 in Section III-B.

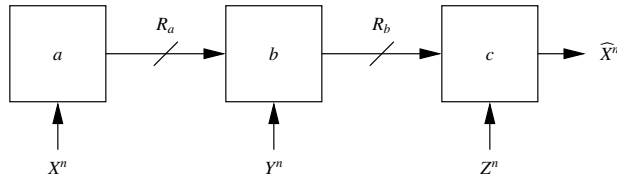


Fig. 1. Relay-assisted distributed source coding.

Other related works include [8], [9].

As discussed in what follows, the study of even the simplest line source network with three sources (Fig. 1) offers new insights for efficient information processing and coding in source networks. The presence of an intermediate relay with correlated side information  $Y^n$  makes the problem both interesting and challenging in a number of ways. Perhaps the two most striking features of this problem are the following. (i) It is not necessary for the relay to reproduce the source  $X^n$ .<sup>1</sup> Hence, the relay need not “fully decode” the message from location- $a$ . (ii) The presence of correlated  $Y^n$  at the relay should be helpful but it is not necessary that  $Y^n$  be reproduced at location- $c$  for  $X^n$  to be reproduced at location- $c$ . Thus we see that there is a very interesting set of interactions and tradeoffs which come into play in this problem. These considerations also give rise to a number of distinct coding strategies for the relay. For instance, the relay can simply forward the message from location- $a$  to location- $c$  without any processing, it can encode some of its own observations and convey it to location- $c$ , or it can process the message from location- $a$  using  $Y^n$  to decode an intermediate description of  $X^n$  and then jointly re-encode the intermediate description together with a part of  $Y^n$  and send it to location- $c$ , etc.

The problem is mathematically formulated in Section II. Section III derives a natural outer bound for the rate region and discusses three special source correlation structures for which the outer bound is tight. The study of these three special cases is rewarding because it exposes three coding strategies of significantly different flavors. An interesting special case for which the outer bound is not tight is also presented. The distinctiveness of the three coding strategies motivates the search for a single unified coding strategy which subsumes all the previously discussed strategies as special cases. This is accomplished by Theorem 1 in Section IV. Concluding remarks are made Section V.

**Notation:** For  $n \in \mathbb{N}$ ,  $a^n := (a_1, \dots, a_n)$ . Kronecker delta  $\delta_{uv}$ :  $\delta_{uv} = 1$  if  $u = v$  and is zero otherwise.

## II. STATEMENT OF THE PROBLEM

Let  $((X_i, Y_i, Z_i))_{i=1}^n$ ,  $n \in \mathbb{N}$ , be a three-component discrete-time, memoryless, stationary source, that is, for  $i = 1, 2, 3, \dots$ ,  $(X_i, Y_i, Z_i) \sim$  i.i.d.  $p_{X,Y,Z}(x, y, z)$  with  $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ . It is assumed that  $\max\{|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|\} < \infty$ .  $X^n$  is observed at

<sup>1</sup>If  $X^n$  is also required to be reproduced at location- $b$  then the problem is much easier and the rate region is immediate (cf. remark at the end of Section II). Thus the interesting problem is when  $X^n$  must be reproduced at only location- $c$ .

location- $a$ ,  $Y^n$  at location- $b$ , and  $Z^n$  at location- $c$ . The goal is to reproduce  $X^n$  at location- $c$ . The constraint is that the message from location- $a$  has to pass through the intermediate location- $b$ . The problem is illustrated and summarized in Fig. 1

**Definition 1:** A relay-assisted distributed source code with parameters  $(n, m_a, m_b)$  is the triple  $(e_a, e_b, g_c)$ :

$$e_a : \mathcal{X}^n \rightarrow \{1, \dots, m_a\}, \quad e_b : \{1, \dots, m_a\} \times \mathcal{Y}^n \rightarrow \{1, \dots, m_b\}, \\ g_c : \{1, \dots, m_b\} \times \mathcal{Z}^n \rightarrow \mathcal{X}^n.$$

Here,  $n$  is the blocklength,  $e_a$  will be called the source encoder,  $e_b$  will be called the relay encoder, and  $g_c$  will be called the source decoder. The quantities  $(1/n) \log_2(m_a)$  and  $(1/n) \log_2(m_b)$  will be called the source and relay block-coding rates (in bits per sample) respectively.

**Definition 2:** A rate-pair  $(R_a, R_b)$  is said to be operationally admissible for relay-assisted distributed source coding if, for every  $\epsilon > 0$ , there exists an  $N_\epsilon \in \mathbb{N}$ , such that, for all  $n \geq N_\epsilon$ , there exists a relay-assisted distributed source code with parameters  $(n, m_a, m_b)$  satisfying

$$\frac{1}{n} \log_2(m_a) \leq R_a + \epsilon, \quad (2.1)$$

$$\frac{1}{n} \log_2(m_b) \leq R_b + \epsilon, \quad (2.2)$$

$$\mathbb{P}(g_c(e_b(e_a(X^n), Y^n), Z^n) \neq X^n) \leq \epsilon. \quad (2.3)$$

Let  $W_a := e_a(X^n)$ ,  $W_b := e_b(W_a, Y^n)$ ,  $\widehat{X}^n := g_c(W_b, Z^n)$ , and  $P_e := \mathbb{P}(\widehat{X}^n \neq X^n)$ .

**Definition 3:** The operationally admissible rate region for relay-assisted distributed source coding,  $\mathcal{R}_{op}$ , is the set of all operationally admissible rate-pairs for relay-assisted distributed source coding.

It can be verified from the above definitions that  $\mathcal{R}_{op}$  is a closed convex subset of  $[0, \infty)^2$  which contains  $[\log_2 |\mathcal{X}|, \infty)^2$ .

**Problem statement:** Characterize  $\mathcal{R}_{op}$  in terms of computable single-letter [1] information quantities.

*Remark:* The problem of characterizing the rate region becomes considerably easier if  $X^n$  has to be reproduced at both location- $b$  and location- $c$  with an error probability which tends to zero as  $n$  tends to infinity. The rate region in this case is given by  $R_a \geq H(X|Y)$ ,  $R_b \geq H(X|Z)$ : For any arbitrarily

small positive real number  $\epsilon$  and all  $n$  sufficiently large,  $X^n$  can be encoded at location- $a$  at the rate  $H(X|Y)$  bits per sample, and decoded at location- $b$  with a probability which can be made smaller than  $\epsilon$ . This follows from the coding theorem due to Slepian and Wolf [10] with  $X^n$  as the source of information to be encoded at location- $a$  and  $Y^n$  as the side-information available to a decoder at location- $b$ . The Slepian-Wolf coding theorem applied a second time with  $X^n$  as the source of information available at location- $b$  and  $Z^n$  as the side-information available to a decoder at location- $c$  shows that, for any arbitrarily small positive real number  $\epsilon$  and all  $n$  sufficiently large,  $X^n$  can be *re-encoded* at location- $b$  at the rate  $H(X|Z)$  bits per sample, and decoded at location- $c$  with a probability which can be made smaller  $\epsilon$ . The converse to the Slepian-Wolf coding theorem proves that the lower bound  $R_a \geq H(X|Y)$  cannot be improved. The cutset bound (3.5) in Section III-A proves that the lower bound  $R_b \geq H(X|Z)$  cannot be improved. Thus the rate-region is fully characterized.

### III. SPECIAL CASES

#### A. The cutset bounds

**Proposition 1:** *If a rate-pair  $(R_a, R_b)$  is operationally admissible for relay-assisted distributed source coding then*

$$R_a \geq H(X|Y, Z), \quad (3.4)$$

$$R_b \geq H(X|Z) \quad (3.5)$$

where  $(X, Y, Z) \sim p_{X,Y,Z}(x, y, z)$  and  $H(X|Z)$ ,  $H(X|Y, Z)$ , are conditional entropies [10] in bits.

*Proof:*  $(R_a, R_b) \in \mathcal{R}_{op} \Rightarrow \forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} : \forall n \geq N_\epsilon$ , there exists a relay-assisted distributed source code with parameters  $(n, m_a, m_b)$  satisfying conditions (2.1), (2.2), and (2.3). By Fano's inequality [10],

$$H(X^n|\widehat{X}^n) \leq 1 + nP_e \log_2 |\mathcal{X}| \leq 1 + n\epsilon \log_2 |\mathcal{X}|. \quad (3.6)$$

For all  $\epsilon > 0$  and all  $n \geq N_\epsilon$ , we have the following information inequalities

$$\begin{aligned} R_a &\stackrel{(a)}{\geq} \frac{1}{n} \log_2 m_a - \epsilon \\ &\stackrel{(b)}{\geq} \frac{1}{n} H(W_a) - \epsilon \\ &\stackrel{(c)}{\geq} \frac{1}{n} H(W_a|Y^n, Z^n) - \epsilon \\ &\stackrel{(d)}{=} \frac{1}{n} I(W_a; X^n|Y^n, Z^n) - \epsilon \\ &= \frac{1}{n} (H(X^n|Y^n, Z^n) - H(X^n|Y^n, Z^n, W_a)) - \epsilon \\ &\stackrel{(e)}{=} H(X|Y, Z) - \frac{1}{n} H(X^n|Y^n, Z^n, W_a, \widehat{X}^n) - \epsilon \\ &\stackrel{(f)}{\geq} H(X|Y, Z) - \frac{1}{n} H(X^n|\widehat{X}^n) - \epsilon \\ &\stackrel{(g)}{\geq} H(X|Y, Z) - \frac{1}{n} - \epsilon (\log_2 |\mathcal{X}| + 1) \end{aligned}$$

where (a) is due to (2.1), (b) is because  $W_a$  takes no more than  $m_a$  distinct values, (c) because conditioning does not increase entropy, (d) because  $W_a$  is a deterministic function of  $X^n$ , (e) is because  $(X_i, Y_i, Z_i) \sim$  i.i.d.  $p_{X,Y,Z}(x, y, z)$  and because  $W_b$  and  $\widehat{X}^n$  are respectively functions of  $(W_a, Y^n)$  and  $(W_b, Z^n)$ , (f) is because unconditioning does not decrease entropy, and (g) is due to Fano's inequality (3.6). Since the last inequality holds for all  $\epsilon > 0$  and all  $n \geq N_\epsilon$ ,

$$R_a \geq H(X|Y, Z).$$

In a similar manner, for all  $\epsilon > 0$  and all  $n \geq N_\epsilon$ , we have the following information inequalities

$$\begin{aligned} R_b &\stackrel{(a)}{\geq} \frac{1}{n} \log_2 m_b - \epsilon \\ &\stackrel{(b)}{\geq} \frac{1}{n} H(W_b) - \epsilon \\ &\stackrel{(c)}{\geq} \frac{1}{n} H(W_b|Z^n) - \epsilon \\ &\geq \frac{1}{n} I(W_b; X^n|Z^n) - \epsilon \\ &= \frac{1}{n} (H(X^n|Z^n) - H(X^n|Z^n, W_b)) - \epsilon \\ &\stackrel{(d)}{=} H(X|Z) - \frac{1}{n} H(X^n|Z^n, W_b, \widehat{X}^n) - \epsilon \\ &\stackrel{(e)}{\geq} H(X|Z) - \frac{1}{n} H(X^n|\widehat{X}^n) - \epsilon \\ &\stackrel{(f)}{\geq} H(X|Z) - \frac{1}{n} - \epsilon (\log_2 |\mathcal{X}| + 1) \end{aligned}$$

where (a) is due to (2.2), (b) is because  $W_b$  takes no more than  $m_b$  distinct values, (c) because conditioning does not increase entropy, (d) is because  $(X_i, Z_i) \sim$  i.i.d.  $p_{X,Z}(x, z)$  and because  $\widehat{X}^n$  is a deterministic function of  $(W_b, Z^n)$ , (e) is because unconditioning does not decrease entropy, and (f) is due to Fano's inequality (3.6). Since the last inequality holds for all  $\epsilon > 0$  and all  $n \geq N_\epsilon$ ,

$$R_b \geq H(X|Z).$$

#### B. Special cases in which the cutset bounds are tight

*Case-1:  $X$ - $Y$ - $Z$  is a Markov chain.* In this case,  $H(X|Y, Z) = H(X|Y)$ . Thus for any arbitrarily small positive real number  $\epsilon$  and all  $n$  sufficiently large,  $X^n$  can be encoded at location- $a$  at the rate  $H(X|Y, Z)$  bits per sample, and decoded at location- $b$  with a probability which can be made smaller than  $\epsilon$ . This follows from the Slepian-Wolf coding theorem with  $X^n$  as the source of information to be encoded at location- $a$  and  $Y^n$  as the side-information available to a decoder at location- $b$ . The Slepian-Wolf coding theorem applied a second time with  $X^n$  as the source of information available at location- $b$  and  $Z^n$  as the side-information available to a decoder at location- $c$  shows that, for any arbitrarily small positive real number  $\epsilon$  and all  $n$  sufficiently

large,  $X^n$  can be *re-encoded* at location- $b$  at the rate  $H(X|Z)$  bits per sample, and decoded at location- $c$  with a probability which can be made smaller  $\epsilon$ . We shall call this relay-assisted distributed source coding strategy as *relay decode and re-encode*.

*Case-2:  $X$ - $Z$ - $Y$  is a Markov chain.* In this case,  $H(X|Y, Z) = H(X|Z)$ . The Slepian-Wolf coding theorem with  $X^n$  as the source of information available at location- $a$  and  $Z^n$  as the side-information available to a decoder at location- $c$  shows that, for any arbitrarily small positive real number  $\epsilon$  and all  $n$  sufficiently large,  $X^n$  can be encoded at location- $a$  at the rate  $H(X|Y, Z)$  bits per sample, and decoded at location- $c$  with a probability which can be made smaller than  $\epsilon$ . The operation at the intermediate location- $b$  is to simply *forward* the message from location- $a$  to location- $c$  without any processing. We shall call this relay-assisted distributed source coding strategy as *relay forward*.

*Case-3:  $H(Y|X, Z) = 0$ , that is, for all tuples  $(x, z) \in \mathcal{X} \times \mathcal{Z}$  for which  $p_{X,Z}(x, z) > 0$ ,  $y$  is a deterministic function of  $(x, z)$ .* In this case,  $H(X|Z) = H(Y|Z) + H(X|Y, Z)$  which is seen by expanding  $H(X, Y|Z)$  in two different ways by the chain rule for conditional entropy [10]. The Slepian-Wolf coding theorem with  $Y^n$  as the source of information available at location- $b$  and  $Z^n$  as the side-information available to a decoder at location- $c$  shows that, for any arbitrarily small positive real number  $\epsilon$  and all  $n$  sufficiently large,  $Y^n$  can be encoded at location- $b$  at the rate  $H(Y|Z)$  bits per sample, and decoded at location- $c$  with a probability which can be made smaller than  $\epsilon$ . The Slepian-Wolf coding theorem applied a second time with  $X^n$  as the source of information available at location- $a$  and  $(Y^n, Z^n)$  as the side-information available to a decoder at location- $c$  shows that, for any arbitrarily small positive real number  $\epsilon$  and all  $n$  sufficiently large,  $X^n$  can be encoded at location- $a$  at the rate  $H(X|Y, Z)$  bits per sample, and decoded at location- $c$  with a probability which can be made smaller than  $\epsilon$ . The operation at the intermediate location- $b$  is to (i) *encode* its observations to be reproduced at location- $c$  and (ii) *forward* the message from location- $a$  to location- $c$  without any processing. We shall call this relay-assisted distributed source coding strategy as *relay encode and forward*.

### C. A case in which the cutset bounds are not tight

*Case-4:  $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$ ,  $p_{Y,Z}(y, z) = 0.5(1 - p)\delta_{yz} + 0.5p(1 - \delta_{yz})$ ,  $p \in (0, 1)$ , and  $X = Y \wedge Z$  where  $Y \wedge Z$  is the Boolean AND function.* According to the cutset bounds,  $R_a \geq H(X|Y, Z) = 0$  and  $R_b \geq H(X|Z) = 0.5H(Y \wedge Z|Z = 1) = 0.5H(Y|Z = 1) = 0.5h_2(p)$  where  $h_2(\alpha) := -\alpha \log_2(\alpha) - (1 - \alpha) \log_2(1 - \alpha)$ ,  $\alpha \in [0, 1]$  is the binary entropy function [10]. If  $R_a = 0$  then it can be argued that the problem is equivalent to a depth-2 network source coding problem with sources  $Y^n$  and  $Z^n$

available at location- $b$  and location- $c$  respectively and the goal being to reproduce the sample-wise Boolean AND function of  $Y^n$  and  $Z^n$  at location- $c$  with a probability which tends to one as  $n$  tends to infinity. This latter problem has been studied by Yamamoto [11] and by Han and Kobayashi [12]. Lemma 1 in [12] applied to the current problem proves that  $R_b \geq H(Y|Z) = h_2(p) > 0.5h_2(p)$ . Thus the cutset bounds are not tight for general  $p_{X,Y,Z}$ .

## IV. A GENERAL INNER RATE REGION

**Theorem 1:** (*Single coding strategy covering all cases*) Let  $(X, Y, Z, U, V)$  be random variables taking values in  $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{V}$ , where  $\mathcal{U}$  and  $\mathcal{V}$  are finite alphabets. For all  $(x, y, z, u, v) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{V}$ , let

$$p_{X,Y,Z,U,V}(x, y, z, u, v) = p_{X,Y,Z}(x, y, z) \cdot p_{U|X}(u|x) \cdot p_{V|U,Y}(v|u, y),$$

where  $p_{U|X}(\cdot|\cdot)$  and  $p_{V|U,Y}(\cdot|\cdot)$  are conditional pmfs, that is, the auxiliary random variables  $U, V$  satisfy the following two Markov chains:  $U$ - $X$ - $(Y, Z)$  and  $V$ - $(U, Y)$ - $(X, Z)$ . Then all rate-pairs  $(R_a, R_b)$  satisfying

$$R_a \geq I(X; U|Y) + H(X|V, Z) \quad (4.7)$$

$$R_b \geq I(Y, U; V|Z) + H(X|V, Z) \quad (4.8)$$

are operationally admissible for relay-assisted distributed source coding and cover all the four previously discussed special cases.

*Proof-sketch:* We first show that all the four previously discussed special cases are covered by Theorem 1. For Case-1, set  $\mathcal{U} = \mathcal{V} = \mathcal{X}$  and  $U = V = X$ , for Case-2, set  $|\mathcal{U}| = |\mathcal{V}| = 1$ , for Cases 3 and 4, set  $|\mathcal{U}| = 1$ ,  $\mathcal{V} = \mathcal{Y}$ , and  $V = Y$ .

We now sketch the proof that all rate-pairs  $(R_a, R_b)$  satisfying (4.7) and (4.8) are operationally admissible for relay-assisted distributed source coding. The proof sketch follows well known random coding and random binning arguments in [1], [10], [13].

Let  $(X, Y, Z, U, V)$  be random variables taking values in  $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{V}$  with the joint pmf  $p_{X,Y,Z,U,V}(x, y, z, u, v) = p_{X,Y,Z}(x, y, z) \cdot p_{U|X}(u|x) \cdot p_{V|U,Y}(v|u, y)$  as stated in Theorem 1. Let  $\mathbf{X} := X^n$ ,  $\mathbf{Y} := Y^n$ ,  $\mathbf{Z} := Z^n$ ,  $\epsilon > 0$ , and  $C_x := (x(1), \dots, x(|\mathcal{X}|^n))$  be an ordering of the set of all source words  $\mathcal{X}^n$ .

*Random codebook generation for location-a:* Generate  $m'_u$  i.i.d. codewords  $U(j) := U^n(j)$ ,  $j = 1, \dots, m'_u$ , of block-length  $n$ , according to the product pmf  $p_U(\mathbf{u}) := \prod_{i=1}^n p_U(u_i)$ , independent of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . The tuple of codewords  $C_u := (U(1), \dots, U(m'_u))$  is available at locations a and b. Generate  $m'_u$  i.i.d. bin-indices  $A_j$ ,  $j = 1, \dots, m'_u$ , according to the uniform pmf over the set  $\{1, \dots, m_u\}$ , independent of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  and  $\{U(j), j = 1, \dots, m'_u\}$ . The tuple of bin-indices  $\mathcal{A} := (A_1, \dots, A_{m'_u})$  is available at locations a and b.

The interpretation is that for  $j = 1, \dots, m'_u$ , codeword  $U(j)$  is assigned to a bin of codewords which has a bin-index  $A_j$ .

Also generate a second set of  $|\mathcal{X}|^n$  i.i.d. bin-indices  $\tilde{A}_j$ ,  $j = 1, \dots, |\mathcal{X}|^n$ , according to the uniform pmf over the

set  $\{1, \dots, m_x\}$ , independent of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  and  $\{\mathbf{U}(j), A_j, j = 1, \dots, m'_u\}$ . The tuple of bin-indices  $\tilde{\mathcal{A}} := (\tilde{A}_1, \dots, \tilde{A}_{|\mathcal{X}|^n})$  is available at locations a and c.

The interpretation is that for  $j = 1, \dots, |\mathcal{X}|^n$ , the sourceword  $\mathbf{x}(j) \in C_x$  is assigned to a bin of sourcewords which has a bin-index  $\tilde{A}_j$ .

*Encoding at location-a:* Let

$$J_u := \min\{j = 1, \dots, m'_u : (\mathbf{X}, \mathbf{U}(j)) \in T_\epsilon^{(n)}(p_{X,U})\}$$

with the convention that  $\min\{\emptyset\} = 1$ , where  $\{\emptyset\}$  denotes the empty set, and where  $T_\epsilon^{(n)}(p_{X,U})$  denotes the  $\epsilon$  strongly  $p_{X,U}$ -typical subset of  $\mathcal{X}^n \times \mathcal{U}^n$  [1]. Also let  $J_x$  be the index of  $\mathbf{X}$  in the tuple of sourcewords  $C_x$ , that is,  $\mathbf{X} = \mathbf{x}(J_x)$ . Set  $W_a := (A_{J_u}, \tilde{A}_{J_x})$ .

The interpretation is that the encoder at location-a first “quantizes” the sourceword  $\mathbf{X}$  to the codeword  $\mathbf{U}(J_u)$ . This “quantization” is accomplished by searching through the tuple of codewords  $C_u$  and finding the first codeword  $\mathbf{U}(J_u)$  which is  $\epsilon$  strongly  $p_{X,U}$  typical with  $\mathbf{X}$ . The encoder also randomly partitions the tuple of codewords  $C_u$  into nonoverlapping bins. The “quantized” codeword  $\mathbf{U}(J_u)$  is assigned to a bin with bin-index  $A_{J_u}$  in this random partitioning. In addition, the encoder also randomly partitions the tuple of all sourcewords  $C_x$  into nonoverlapping bins. The observed sourceword  $\mathbf{X}$ , which occurs at position  $J_x$  in the tuple  $C_x$ , is assigned to a bin with bin-index  $\tilde{A}_{J_x}$  in this random partitioning. The bin-index  $A_{J_u}$  of  $\mathbf{U}(J_u)$  together with the bin-index  $\tilde{A}_{J_x}$  of  $\mathbf{X} = \mathbf{x}(J_x)$  comprise the message  $W_a$  sent from location-a to location-b.

*Random codebook generation for location-b:* Generate  $m'_v$  i.i.d. relay-codewords  $\mathbf{V}(j) := V^n(j)$ ,  $j = 1, \dots, m'_v$ , of block-length  $n$ , according to the product pmf  $p_V(\mathbf{v}) := \prod_{i=1}^n p_V(v_i)$ , independent of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ ,  $\{\mathbf{U}(j), A_j, j = 1, \dots, m'_u\}$ , and  $\{\tilde{A}_j, j = 1, \dots, |\mathcal{X}|^n\}$ . The tuple of relay-codewords  $C_v := (\mathbf{V}(1), \dots, \mathbf{V}(m'_v))$  is available at locations b and c. Generate  $m'_v$  i.i.d. bin-indices  $B_j$ ,  $j = 1, \dots, m'_v$ , according to the uniform pmf over the set  $\{1, \dots, m_v\}$ , independent of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ ,  $\{\mathbf{U}(j), A_j, j = 1, \dots, m'_u\}$ ,  $\{\tilde{A}_j, j = 1, \dots, |\mathcal{X}|^n\}$ , and  $\{\mathbf{V}(j), j = 1, \dots, m'_v\}$ . The tuple of bin-indices  $\mathcal{B} := (B_1, \dots, B_{m'_v})$  is available at locations b and c.

The interpretation is that for  $j = 1, \dots, m'_v$ , codeword  $\mathbf{V}(j)$  is assigned to a bin of codewords which has a bin-index  $B_j$ .

*Encoding at location-b:* Noting that  $W_a = (A_{J_u}, \tilde{A}_{J_x})$  and  $\mathbf{Y}$  are available at location-b, let

$$\hat{J}_u := \min\{j = 1, \dots, m'_u : A_j = A_{J_u}, (\mathbf{U}(j), \mathbf{Y}) \in T_\epsilon^{(n)}(p_{U,Y})\}$$

and

$$J_v := \min\{j = 1, \dots, m'_v : (\mathbf{U}(\hat{J}_u), \mathbf{Y}, \mathbf{V}(j)) \in T_\epsilon^{(n)}(p_{U,Y,V})\}.$$

Set  $W_b := (B_{J_v}, \tilde{A}_{J_x})$ .

The interpretation is that the encoder at location-b first estimates the codeword  $\mathbf{U}(J_u)$  as  $\mathbf{U}(\hat{J}_u)$ . This is accomplished by searching through the bin of codewords in  $C_u$  whose bin-index is  $A_{J_u}$  and finding the first codeword  $\mathbf{U}(\hat{J}_u)$  which is  $\epsilon$  strongly  $p_{U,Y}$  typical with  $\mathbf{Y}$ . The encoder next “jointly-quantizes” the pair  $(\mathbf{U}(\hat{J}_u), \mathbf{Y})$  to the relay-codeword  $\mathbf{V}(J_v)$ . This “joint-quantization” is accomplished by searching through the tuple

of relay-codewords  $C_v$  and finding the first relay-codeword  $\mathbf{V}(J_v)$  which is  $\epsilon$  strongly  $p_{U,Y,V}$  typical with  $\mathbf{U}(\hat{J}_u)$  and  $\mathbf{Y}$ . The encoder also randomly partitions the tuple of relay-codewords  $C_v$  into nonoverlapping bins. The “quantized” relay-codeword  $\mathbf{V}(J_v)$  is assigned to a bin with bin-index  $B_{J_v}$  in this random partitioning. The bin-index  $B_{J_v}$  of  $\mathbf{V}(J_v)$  together with the bin-index  $\tilde{A}_{J_x}$  of  $\mathbf{X} = \mathbf{x}(J_x)$  comprise the message  $W_b$  sent from location-b to location-c.

*Decoding at location-c:* Noting that  $W_b = (B_{J_v}, \tilde{A}_{J_x})$  and  $\mathbf{Z}$  are available at location-c, let

$$\hat{J}_v := \min\{j = 1, \dots, m'_v : B_j = B_{J_v}, (\mathbf{V}(j), \mathbf{Z}) \in T_\epsilon^{(n)}(p_{V,Z})\}$$

and

$$\hat{J}_x := \min\{j = 1, \dots, |\mathcal{X}|^n : \tilde{A}_j = \tilde{A}_{J_x}, (\mathbf{X}(j), \mathbf{V}(\hat{J}_v), \mathbf{Z}) \in T_\epsilon^{(n)}(p_{X,V,Z})\}.$$

Set  $\hat{\mathbf{X}} := \mathbf{x}(\hat{J}_x)$ .

The interpretation is that the decoder at location-c first estimates the relay-codeword  $\mathbf{V}(J_v)$  as  $\mathbf{V}(\hat{J}_v)$ . This is accomplished by searching through the bin of relay-codewords in  $C_v$  whose bin-index is  $B_{J_v}$  and finding the first codeword  $\mathbf{V}(\hat{J}_v)$  which is  $\epsilon$  strongly  $p_{V,Z}$  typical with  $\mathbf{Z}$ . The decoder next estimates the sourceword  $\mathbf{X} = \mathbf{x}(J_x)$  as  $\hat{\mathbf{X}} = \mathbf{x}(\hat{J}_x)$ . This is accomplished by searching through the set of all sourcewords in  $C_x$  whose bin-index is  $\tilde{A}_{J_x}$  and finding the first sourceword  $\mathbf{X}(\hat{J}_x)$  which is  $\epsilon$  strongly  $p_{X,V,Z}$  typical with  $\mathbf{V}(\hat{J}_v)$  and  $\mathbf{Z}$ .

*Error events:*

$$\mathcal{E}_1: \{\mathbf{X} \notin T_\epsilon^{(n)}(p_X)\}.$$

As  $n \uparrow \infty$ , by the strong law of larger numbers,  $\mathbb{P}(\mathcal{E}_1) \downarrow 0$ .

$$\mathcal{E}_2: \{(\mathbf{X}, \mathbf{U}(J_u)) \notin T_\epsilon^{(n)}(p_{X,U})\}.$$

There exists an  $\epsilon_1(n, \epsilon) > 0$  such that (i)  $\epsilon_1(n, \epsilon) \downarrow 0$  as  $n \uparrow \infty$  and  $\epsilon \downarrow 0$  and (ii)  $\mathbb{P}(\mathcal{E}_2 | \mathcal{E}_1^c) \downarrow 0$  if

$$(1/n) \log_2 m'_u = I(X; U) + \epsilon_1(n, \epsilon). \quad (4.9)$$

$$\mathcal{E}_3: \{(\mathbf{X}, \mathbf{U}(J_u), \mathbf{Y}) \notin T_\epsilon^{(n)}(p_{X,U,Y})\}.$$

$\mathbb{P}(\mathcal{E}_3 | \mathcal{E}_2^c) \downarrow 0$  as  $n \uparrow \infty$  by the Markov lemma [10] because  $U-X-Y$  is a Markov chain.

$$\mathcal{E}_4: \{\exists j \neq J_u : A_j = A_{J_u}, (\mathbf{U}(j), \mathbf{Y}) \in T_\epsilon^{(n)}(p_{U,Y})\},$$

There exists an  $\epsilon_2(n, \epsilon) > 0$  such that (i)  $\epsilon_2(n, \epsilon) \downarrow 0$  as  $n \uparrow \infty$  and  $\epsilon \downarrow 0$  and (ii)  $\mathbb{P}(\mathcal{E}_4) \downarrow 0$  if

$$(1/n)(\log_2 m'_u - \log_2 m_u) = I(U; Y) - \epsilon_2(n, \epsilon). \quad (4.10)$$

Conditions (4.9) and (4.10) and the Markov chain  $U-X-(Y, Z)$  imply the following condition on  $(1/n) \log_2 m_u$ :

$$\begin{aligned} \frac{1}{n} \log_2 m_u - \epsilon_1(n, \epsilon) - \epsilon_2(n, \epsilon) &= I(X; U) - I(U; Y) \\ &= I(X; U|Y). \end{aligned} \quad (4.11)$$

Hence under condition (4.11),  $\mathbb{P}(\hat{J}_u \neq J_u) \downarrow 0$  as  $n \uparrow \infty$ .

$$\mathcal{E}_5: \{(\mathbf{U}(J_u), \mathbf{Y}, \mathbf{V}(J_v)) \notin T_\epsilon^{(n)}(p_{X,U,Y,V})\},$$

There exists an  $\epsilon_3(n, \epsilon) > 0$  such that (i)  $\epsilon_3(n, \epsilon) \downarrow 0$  as  $n \uparrow \infty$  and  $\epsilon \downarrow 0$  and (ii)  $\mathbb{P}(\mathcal{E}_5 | (\mathcal{E}_3 \cup \mathcal{E}_4)^c) \downarrow 0$  if

$$(1/n) \log_2 m'_v = I(U, Y; V) + \epsilon_3(n, \epsilon). \quad (4.12)$$

$\mathcal{E}_6: \{(U(J_u), Y, V(J_v), X, Z) \notin T_\epsilon^{(n)}(p_{U,Y,V,X,Z})\}.$

$\mathbb{P}(\mathcal{E}_6 | \mathcal{E}_5^c) \downarrow 0$  as  $n \uparrow \infty$  by the Markov lemma [10] because  $V-(U, Y)-(X, Z)$  is a Markov chain.

$\mathcal{E}_7: \{\exists j \neq J_v : B_j = B_{J_v}, (V(j), Z) \in T_\epsilon^{(n)}(p_{V,Z})\},$

There exists an  $\epsilon_4(n, \epsilon) > 0$  such that (i)  $\epsilon_4(n, \epsilon) \downarrow 0$  as  $n \uparrow \infty$  and  $\epsilon \downarrow 0$  and (ii)  $\mathbb{P}(\mathcal{E}_7) \downarrow 0$  if

$$(1/n)(\log_2 m'_v - \log_2 m_v) = I(V; Z) - \epsilon_4(n, \epsilon). \quad (4.13)$$

Conditions (4.12) and (4.13) and the Markov chain  $V-(U, Y)-(X, Z)$  imply the following condition on  $(1/n)\log_2 m_v$ :

$$\begin{aligned} \frac{1}{n} \log_2 m_v - \epsilon_3(n, \epsilon) - \epsilon_4(n, \epsilon) &= I(U, Y; V) - I(V; Z) \\ &= I(U, Y; V|Z). \end{aligned} \quad (4.14)$$

Hence under condition (4.14),  $\mathbb{P}(\hat{J}_v \neq J_v) \downarrow 0$  as  $n \uparrow \infty$ .

$\mathcal{E}_8: \{\exists j \neq J_x : \tilde{A}_j = \tilde{A}_{J_x}, (\mathbf{x}(j), V(J_v), Z) \in T_\epsilon^{(n)}(p_{V,Z})\}.$

There exists an  $\epsilon_5(n, \epsilon) > 0$  such that (i)  $\epsilon_5(n, \epsilon) \downarrow 0$  as  $n \uparrow \infty$  and  $\epsilon \downarrow 0$  and (ii)  $\mathbb{P}(\mathcal{E}_8 | (\mathcal{E}_6 \cup \mathcal{E}_7)^c) \downarrow 0$  if

$$\begin{aligned} H(X) - (1/n)\log_2 m_x &= I(X; V, Z) - \epsilon_5(n, \epsilon), \text{ i.e.,} \\ (1/n)\log_2 m_x - \epsilon_5(n, \epsilon) &= H(X|V, Z). \end{aligned} \quad (4.15)$$

Hence under conditions (4.11), (4.14), and (4.15),  $\mathbb{P}(\hat{J}_x \neq J_x) \downarrow 0$  as  $n \uparrow \infty$ . Since this probability is a weighted average over random codes, there exists at least one deterministic sequence of codes for which  $\mathbb{P}(\hat{J}_x \neq J_x) \downarrow 0$  as  $n \uparrow \infty$ . ■

## V. CONCLUDING REMARKS

A single-letter characterization of the entire rate region for the relay-assisted distributed source coding problem is still open at this time. Theorem 1 subsumes all the special cases for which the solution is known. Theorem 1 offers a single unified coding strategy based on familiar random coding and random binning arguments while incorporating the unique features of the noticeably different coding strategies of the special cases. Even though the rate regions of many simple source networks are not fully characterized, their study can be rewarding for they offer valuable insights for constructing interesting new distributed source coding strategies.

## ACKNOWLEDGMENT

The authors would like to thank Prof. S. Tatikonda, EE Yale University, and Prof. K. Ramchandran, EECS UC Berkeley, for introducing them to the problem and early discussions of some special cases. The authors would also like to acknowledge Dr. D. Petrovic, Atheros Communications, for special case 2 and thank Dr. C. Tian, AT&T Research Labs, for bringing to their attention references [7]–[9]. This material is based upon work supported by the US National Science Foundation (NSF) under awards (CAREER) CCF-0546598 and (CAREER) CCF-0448115. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

## REFERENCES

- [1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Budapest, Hungary: Akadémiai Kiadó, 1986.
- [2] S. S. Pradhan, J. Kusuma and K. Ramchandran, "Distributed compression in a dense microsensor network," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 51–60, Mar. 2002.
- [3] C. Fragouli and E. Soljanin, "Network Coding Ten Years after the Butterfly Emerged," in *IEEE Information Theory Society Newsletter*, vol. 57, no. 4, pp. 11–15, Dec. 2007.
- [4] T-S. Han and K. Kobayashi, "A Unified Achievable Rate Region for a General Class of Multiterminal Source Coding Systems," *IEEE Trans. Info. Theory*, vol. IT-26, pp. 277–288, May 1980.
- [5] S. C. Draper and G. W. Wornell, "Side Information Aware Coding Strategies for Sensor Networks," *J. Select. Areas Commun. (JSAC)*, vol. 22, no. 6, pp. 966–976, Aug. 2004.
- [6] M. Bakshi, M. Effros, W. Gu, and R. Koetter, "On network coding of independent and dependent sources in line networks," *Proc. IEEE International Symposium on Information Theory*, Nice, France, Jun. 2007.
- [7] D. Vasudevan, C. Tian, and S. Diggavi, "Lossy Source Coding for a Cascade Communication System with Side-Information," *Proc. Allerton Conference on Communication, Control and Computing*, IL, 27–29 Sep. 2006.
- [8] W. Gu and M. Effros, "Simple bounds for lossless source coding in a two-hop network," in *UCSD ITA Center Inaugural Workshop*, Feb. 2006.
- [9] H. Yamamoto, "Source coding theory for cascade and branching communication systems," *IEEE Trans. Info. Theory*, vol. IT-27, pp. 299–308, May 1981.
- [10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.
- [11] H. Yamamoto, "Wyner-Ziv theory for a general function of the correlated sources," *IEEE Trans. Info. Theory*, vol. IT-28, pp. 803–807, Sep. 1982.
- [12] T-S. Han and K. Kobayashi, "A dichotomy of functions  $F(X, Y)$  of correlated sources  $(X, Y)$ ," *IEEE Trans. Info. Theory*, vol. IT-33, pp. 69–76, Jan. 1987.
- [13] T. Berger, "Multiterminal Source Coding", in *The Information Theory Approach to Communications, (CISM Courses and Lectures No. 229)*, G. Longo, Ed., Vienna/New York: Springer-Verlag, 1978, pp. 171–231.