Computer Aided Proofs for Rate Regions of Independent Distributed Source Coding Problems

Congduan Li, Steven Weber, John MacLaren Walsh
ECE Department, Drexel University
Philadelphia, PA 19104

***Abstract*—The rate regions of independent distributed source coding (IDSC) problems, a sub-class of the broader family of multi-source multi-sink networks, are investigated. An IDSC problem consists of multiple sources, multiple encoders, and multiple decoders, where each encoder has access to all sources, and each decoder has access to a certain subset of the encoders and demands a certain subset of the sources. Instead of manually deriving the rate region for a particular problem, computer tools are used to obtain the rate regions for hundreds of non-isomorphic (symmetry-removed) IDSC instances. A method for enumerating all non-isomorphic IDSC instances of a particular size is given. For each non-isomorphic IDSC instance, the Shannon outer bound, superposition coding inner bound, and several achievable inner bounds based on linear codes, are considered and calculated. For all of the hundreds of IDSC instances considered, vector binary inner bounds match the Shannon outer bound, and hence, exact rate regions are proven together with code constructions that achieve them.**

I. Introduction

Many important practical problems, such as efficient information transfer over a wireless or wired network, efficient data storage on disks in a distributed storage system, and efficient transmission in a video streaming system, have been shown to involve determining the rate region of an abstracted network under network coding. In many cases, the abstracted network coding problem is a multi-source multi-sink multicast problem, which is open in general.

This paper addresses one type of the multi-source multi-sink multicast networks, the independent distributed source coding (IDSC) problems. As will be introduced in §II, the IDSC model consists of multiple sources, multiple encoders, and multiple decoders, where each encoder has access to all sources, and each decoder has access to a certain subset of the encoders and demands a certain subset of the sources. Since most of these problems are open, it is not easy to derive analytical expressions for the rate regions of them. This motivates us to use computers to calculate the rate regions and give proofs automatically.

Yeung *et al.*'s early work [1], which was motivated from satellite communication systems, and Yan *et al.*'s celebrated paper [2] for general networks, provided a method, in principle, to calculate the capacity region of networks under network coding. Since the calculation involves Γ_N^*, the region of entropic vectors, which is not fully characterized yet for $N \geq 4$, we have no direct way to obtain exact capacity region when there are more than four network variables, both source and edge variables included. The problems presented in this paper are in this case since they have four to six variables. However, one can use bounds on the region of entropic vectors to calculate inner and outer bounds on the capacity region and then determine the capacity region by comparing the bounds. Other interesting characteristics of the network can be determined through bound comparison, including the sufficiency of classes of linear codes to obtain the entire rate region, and the ability of Shannon-type inequalities to determine the rate region. Multiple methods can be utilized to directly calculate these bounds through polyhedral projection in this manner [3]–[6].

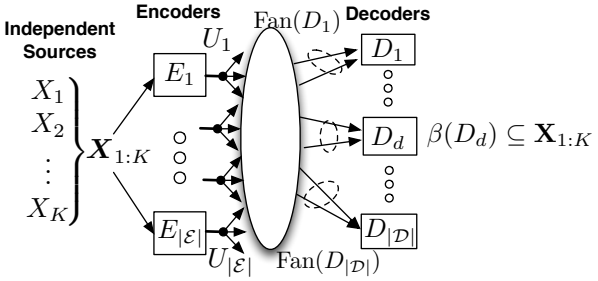

Figure 1: A general IDSC model A*: K independent sources are available to encoders indexed by $\mathcal{E}$; a decoder $D_d, d \in \mathcal{D}$ has access to $\mathrm{Fan}(D_d) \subseteq \mathcal{E}$ and demands $\beta(D_d) \subseteq \mathbf{X}_{1:K}$. The number of variables in the network is $N = K + |\mathcal{E}|$.*

A special class of IDSC problems with prioritized sources, multilevel diversity coding systems (MDCS), has been studied in [3], [4], [6]–[10]. This paper investigates general IDSC problems with the following agenda: I) enumerate all non-isomorphic instances for a given size; II) for each instance, obtain the exact rate region by comparing the calculated outer and inner bounds on the rate region using computer programs; III) generate human readable proofs. The inner bounds considered include the (binary) representable matroid inner bounds and superposition coding rate region. It is shown that for the 219 IDSC instances considered, superposition is not always optimal, but binary linear codes are always optimal, and the Shannon outer bound is tight for every instance.

II. System Model

In an IDSC instance, as shown in Fig. 1 and denoted as A, there are K *independent* sources $\boldsymbol{X}_{1:K} \equiv (X_1, \ldots, X_K)$ where source k has support $\mathcal{X}_k$. As is standard in source coding, each source X_k is in fact an i.i.d. sequence of random variables $\{X_k^t, t = 1, 2, \ldots\}$ in t, and X_k is a representative random variable with this distribution.

It is assumed that all sources are available to each of the encoders indexed by the finite set $\mathcal{E}$. The output $\mathrm{Out}(E_e)$ of

an encoder E_e is a description/message variable $U_e, e \in \mathcal{E}$. The message variables, or the encoders, are mapped to a collection of decoders indexed by the set $\mathcal{D}$. The collection of mappings is denoted as a bipartite set $\mathcal{G}$, $\mathcal{G} \subseteq \mathcal{E} \times \mathcal{D}$ of edges where $(E_e, D_d) \in \mathcal{G}$ if E_e is accessible by D_d. The set of encoders mapped to a particular decoder D_d is called the *fan* of D_d, and is denoted as $\mathrm{Fan}(D_d) = \{E_e | (E_e, D_d) \in \mathcal{G}\}$. Similarly, the set of decoders connected to a particular encoder E_e is called the *fan* of E_e, and is denoted by $\mathrm{Fan}(E_e) = \{D_d | (E_e, D_d) \in \mathcal{G}\}$. The demands of a decoder D_d are $\beta(D_d) \subseteq \mathbf{X}_{1:k}$. Furthermore, with a little abuse of notation, we denote $\mathrm{Fan}(\mathbf{X}_\mathcal{I}) = \{i \in \mathcal{D} | \beta(D_i) = \mathbf{X}_\mathcal{I}\}, \mathcal{I} \subseteq 1:K$. Note that the total number of variables in the network is $N = K + |\mathcal{E}|$, including all source and coded message variables.

We define an IDSC instance as *minimal* if it obeys:

(**C1**.) $\forall i, j \in \mathcal{D}$, $\mathrm{Fan}(D_i) \neq \mathrm{Fan}(D_j)$;
(**C2**.) If $\mathrm{Fan}(D_i) \subseteq \mathrm{Fan}(D_j)$, $i \neq j$, then $\beta(D_i) \cap \beta(D_j) = \emptyset$;
(**C3**.) $\bigcup_{i \in \mathcal{D}} \mathrm{Fan}(D_i) = \mathcal{E}$;
(**C4**.) $\nexists\, k, l \in \mathcal{E}, k \neq l$ such that $\mathrm{Fan}(E_k) = \mathrm{Fan}(E_l)$.
(**C5**.) $\forall k \in \{1, 2, \ldots, K\}$, $\exists\, d \in \mathcal{D}$ such that $X_k \in \beta(D_d)$.
(**C6**.) $\forall k, l \in \{1, 2, \ldots, K\}, k \neq l$, we have $\{d \in \mathcal{D} | X_k \in \beta(D_d)\} \neq \{d \in \mathcal{D} | X_l \in \beta(D_d)\}$.

The first condition (**C1**) indicates the trivial requirement that no two decoders should have the same fan, for otherwise these two decoders can be combined. (**C2**) is necessary for pursuing minimality of representation of a network because the demand at decoder D_i implies that decoder D_j can also decode $\beta(D_i)$. This condition also makes sure that there does not exist a contradiction in the decoding capabilities, for instance, a decoder may have access to more encoders than another but demands less. A special case of this condition is when the demands of some decoders are identical. This requires that the fan of such decoders cannot be a subset of one another. (**C3**) requires that every encoder must be in use in the reconstruction. (**C4**) requires that no two encoders have exactly the same fan, for otherwise the two encoders can be combined. (**C5**) ensures that no source is redundant. (**C6**) ensures that no two sources can be combined as a single source.

A. Representation of IDSC instances

One representation of an IDSC instance is to list the fan and demands of each decoder. Since $\mathrm{Fan}(D_d) \subseteq \mathcal{E}, \forall d \in \mathcal{D}$, one can represent $\mathrm{Fan}(D_d)$ using a $|\mathcal{E}|$-bit indicator vector or a corresponding integer value, where the entries of the vector from left to right are mapped to $E_{|\mathcal{E}|}, \ldots, E_1$ and a one in a position indicates the inclusion of the associated encoder in $\mathrm{Fam}(D_d)$. Similarly, the demands $\beta(D_d)$ can be coded with a K-bit vector or integer with D_1 corresponding to the last bit. We can define a bijection $\theta : n \leftrightarrow find(bin(n) \neq 0)$ between positions of indicators (non-zero values) in binary sequence and corresponding integer value. For example, $\theta(6) = \{2, 3\}$ because the binary sequence of 6 is (110) and the non-zero positions are $2, 3$ from the right side. Furthermore, we denote that $\mathbf{X}_{\theta(n)} = \{X_i, i \in \theta(n)\}$. With this encoding, an IDSC instance can be easily represented by a matrix with $2^K - 1$ rows, where the row indices represent the demands of decoders and entries are integers representing the fan of each decoder. An all-zero vector at i-th row means there is no decoder that only requires $\mathbf{X}_{\theta(i)}$. Each row may include some zeros to make them the same length. For example, the configuration matrix for the fourth $(2, 2)$ IDSC instance, shown as $iv)$ in Fig. 2, is $\left[\begin{smallmatrix} 0 & 3 \\ 1 & 2 \\ 0 & 0 \end{smallmatrix}\right]$, where the three row vectors indicate that there is one decoder that has access to $\{E_1, E_2\}$ ($(11)_2 = 3$) and demands X ($\theta(1) = \{1\}$). Two decoders that have access to E_1 ($(01)_2 = 1$), E_2 ($(10)_2 = 2$) respectively both demand Y ($\theta(2) = \{2\}$). Note that this implies the first decoder can also decode Y, but we only list X as its demand due to minimality condition (**C2**).

Another notation for an IDSC instance that we will use extensively in this paper is the tuple $(\mathbf{X}_{1:K}, \mathcal{E}, \mathcal{D}, \mathcal{G}, \beta(D_d), d \in \mathcal{D})$ where $\mathbf{X}_{1:K} = (X_1, \ldots, X_K)$ represents the K sources, $\mathcal{E}$ is the encoder set, $\mathcal{D}$ is the decoder set with corresponding demands $\beta(D_d), d \in \mathcal{D}$, and $\mathcal{G}$ is the set of edges between encoders and decoders which indicates the accesses of each decoder: if D_d has access to E_e, the edge $(E_e, D_d) \in \mathcal{G}$.

B. Enumeration of non-isomorphic IDSC Problems

We begin by defining a notion of equivalent or isomorphic IDSC problem instances.

Definition 1 (Isomorphic IDSC instances)**:** Two $(K, |\mathcal{E}|)$ IDSC instances, $\mathsf{A} = (\mathbf{X}_{1:K}, \mathcal{E}, \mathcal{D}, \mathcal{G}, \beta(D_d), d \in \mathcal{D})$ and $\mathsf{A}' = (\mathbf{X}'_{1:K}, \mathcal{E}', \mathcal{D}', \mathcal{G}', \beta(D_{d'}), d' \in \mathcal{D}')$, are *isomorphic*, denoted as $\mathsf{A} \cong \mathsf{A}'$, if there exist a permutation of sources $\sigma : \mathbf{X}_{1:K} \to \mathbf{X}'_{1:K}$, a permutation of encoders $\mathcal{E}$, $\pi : \mathcal{E} \to \mathcal{E}'$ such that $\mathbf{X}'_{1:K} = \sigma(\mathbf{X}_{1:K})$, $\mathcal{E}' = \pi(\mathcal{E})$, $\mathcal{D}' = \mathcal{D}$, $\mathcal{G}' = \{(\pi(E_e), D_d) | (E_e, D_d) \in \mathcal{G}\}$, and $\beta(D_{d'}) = \sigma(\beta(D_d))$.

Since the isomorphism merely permutes the sources and/or encoders, to study all possible IDSC instances, it suffices to consider one representative in each isomorphism class, i.e., only consider non-isomorphic IDSC instances. One method to do this is to remove isomorphisms from the list of all IDSC instances. Utilizing such a method has the additional benefit of indicating the total list of IDSC instances (with symmetries included) for comparison with the list and number of non-isomorphic instances. In order to obtain all IDSC instances, similar as [6], we use the observation that the fan of decoders with the same decoding ability must be a *Sperner family* of the encoders set $\mathcal{E}$, as required by condition (**C2**). A Sperner family of $\mathcal{E}$, sometimes also called an *independent system* or a *clutter*, is a collection of subsets of $\mathcal{E}$ such that no element is contained in another. Since the fan of a certain subset of the sources can be empty, the empty set, or the all-zero row vector, is used to represent the mapping between encoders and such "decoders".

An algorithm to enumerate isomorphic and non-isomorphic IDSC instances is given in Algorithm 1. We consider the subsets of sources as demands of decoders in a binary-count order. The enumeration process works as follows.

1) List all Sperner families, $\mathrm{Sper}(\mathcal{E})$ of the encoders set $\mathcal{E}$ (required by (**C2**)); this can be done by considering all combinations of subsets of $\mathcal{E}$ that satisfy the clutter property;

2) All Sperner families, including the empty set, are possible configurations for decoders demanding the first source, except when $K = 1$, from (**C5**);
3) Suppose we have configurations for up to the fan of $\mathbf{X}_{\theta(i-1)}$. Now we will augment the existing partially-configured IDSC instances with configurations for the fan of $\mathbf{X}_{\theta(i)}$, i.e., the decoders merely demanding $\mathbf{X}_{\theta(i)}$. We consider all the existing instances one by one. Take one of them as an example. Since the fan of decoders in $\mathrm{Fan}(\mathbf{X}_{\theta(i)})$ cannot have same fan as the existing decoders (required by (**C1**)), we should remove the Sperner families containing at least one element that is already selected in the existing decoders. Furthermore, for every $j < i$ such that $\theta(j) \subset \theta(i)$, all Sperner families containing at least one element that is a subset of the fan of a decoder in $\mathrm{Fan}(\mathbf{X}_{\theta(i)})$ shall be removed as well (required by (**C2**)). The remaining Sperner families are possible configurations for $\mathrm{Fan}(\mathbf{X}_{\theta(i)})$. Repeat until $i = 2^K - 1$.
4) At the last stage, when $i = 2^K - 1$, the fan of $\mathbf{X}_{1:K}$ are considered. If all encoders have been assigned access to at least one decoder (required by (**C3**)), no two encoders have the same fan (required by (**C4**)), and each source is required by at least one decoder (required by (**C5**)), an IDSC instance with simplest structure has been obtained;
5) After step 4), all IDSC instances are obtained with isomorphism. Then we remove isomorphism by keeping one instance in every isomorphism class, where the IDSC instances can be obtained by permuting sources and/or encoders from one another, and remove the others in the isomorphism class from the list of all IDSC instances.

The numbers of IDSC instances for some $(K, |\mathcal{E}|)$ pairs are listed in the first three columns of Table I. Due to the limit of space, we only list the 4 non-isomorphic minimal $(2, 2)$ IDSC instances in Fig. 2.

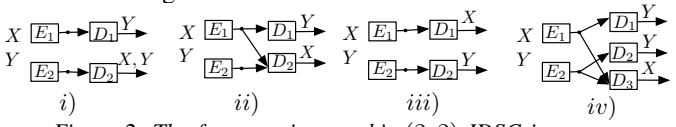

Figure 2: The four non-isomorphic $(2, 2)$ IDSC instances.

The fast growth of number of Sperner families with increasing $|\mathcal{E}|$ renders Algorithm 1 inefficient for large numbers of encoders. As a special case of the general network model in [11], we can use the algorithm in [11], based on Leiterspiel [12], to list the non-isomorphic IDSC instances directly.

Input: Encoder index set $\mathcal{E}$, Number of sources K
Output: All non-isomorphic IDSC instances $\mathcal{M}$, all IDSC instances with isomorphism $\mathcal{M}'$

Initialization: List all Sperner families $\mathrm{Sper}(\mathcal{E})$ of $\mathcal{E}$, $pool = \mathrm{Sper}(\mathcal{E})$, $\mathsf{A} = (\mathbf{X}_{1:K}, \mathcal{E}, \emptyset, \emptyset, \emptyset)$, $\mathcal{M}' = \mathsf{A}$;
for $i = 1 : (2^K - 1)$ **do**
  $\mathcal{M}'' = \mathcal{M}', \mathcal{M}' = \emptyset$;
  for *every* $\mathsf{A} \in \mathcal{M}''$ **do**
    if $i \neq 2^K - 1$ **then**
      $pool = \mathrm{Sper}(\mathcal{E}) \setminus \{\mathcal{I} \in \mathrm{Sper}(\mathcal{E}) | Aug(\mathsf{A}, \mathcal{I}, i)$ violates (**C1**)-(**C2**)$\}$;
    end
    else
      $pool = \mathrm{Sper}(\mathcal{E}) \setminus \{\mathcal{I} \in \mathrm{Sper}(\mathcal{E}) | Aug(\mathsf{A}, \mathcal{I}, i)$ violates (**C1**)-(**C5**)$\}$;
    end
    $\mathcal{A} = Aug(\mathsf{A}, pool, i), \mathcal{M}' = \mathcal{M}' \cup \mathcal{A}$;
  end
end
Remove isomorphisms: $\mathcal{M} = isoRemoval(\mathcal{M}')$;
where the Aug function is defined as follows
Function: $\mathsf{A} = Aug(\mathsf{A}, pool, i)$
$\mathcal{A} = \emptyset$;
for *every* $\mathcal{I} \in pool$ **do**
  $\mathcal{D}' = \mathcal{D} \cup \{|\mathcal{D}| + 1, \ldots, |\mathcal{D}| + |\mathcal{I}|\}$;
  $\beta(|\mathcal{D}| + 1) = \cdots = \beta(|\mathcal{D}| + |\mathcal{I}|) = \mathbf{X}_{\theta(i)}$;
  for $j = 1 : |\mathcal{I}|$ **do**
    $\mathcal{G}' = \mathcal{G} \cup \{(E_e, D_{|\mathcal{D}|+j}) | E_e \in \mathcal{I}(j)\}$;
  end
  $\mathcal{A} = \mathcal{A} \cup (\mathbf{X}_{1:K}, \mathcal{E}, \mathcal{D}', \mathcal{G}', \beta(D_{d'}), d' \in \mathcal{D}')$;
end

Algorithm 1: Enumerate isomorphic and non-isomorphic $(K, |\mathcal{E}|)$ IDSC instances

III. Rate Region Formulation

We now define the network codes and the capacity region for an IDSC instance. Each source node $k \in \{1, \ldots, K\}$ is associated with an IID sequence distributed according to its independent random variable X_k taking values in $\mathcal{X}_k$. Let $\mathbf{R} = (H(X_1), \ldots, H(X_K), R_1, \ldots, R_{|\mathcal{E}|}) \in \mathbb{R}_+^{K+|\mathcal{E}|}$ be a vector of source and edge rates. A $(n, \mathbf{R})$ block code over $\mathbb{F}_q$ consists of a series of block encoders, one for each $e \in \mathcal{E}$, which are functions that map a block of n source observations from all sources, to one of $\lceil q^{nR_e} \rceil$ different descriptions in $\eta_e = \{0, 1, \ldots, \lceil q^{nR_e} \rceil - 1\}$, $f_e^{(n)} : \prod_{k \in \{1,\ldots,K\}} \mathcal{X}_k^n \to \eta_e,\ e \in \mathcal{E}$, and a series of decoders $d \in \mathcal{D}$, which are functions $g_d^{(n)} : \prod_{e \in \mathrm{Fan}(D_d)} \eta_e \to \prod_{k \in \beta(D_d)} \mathcal{X}_k^n,\ d \in \mathcal{D}$. Denote by $U_e^n \in \eta_e$ the message for encoder e, $e \in \mathcal{E}$, which is the result of the encoding function f_e^n.

The achievable rate region of a network A, denoted as $\mathcal{R}_*(\mathsf{A})$, consists of all rate vectors $\mathbf{R} = (H(X_1), \ldots, H(X_K), R_1, \ldots, R_{|\mathcal{E}|})$ such that there exist sequences of encoding functions $f^n = (f_e^n, e \in \mathcal{E})$ and decoding functions $g^n = (g_d^n, d \in \mathcal{D})$ for which the probability of error at each decoder can be made arbitrarily small as $n \to \infty$, and the closure of this achievable region is the capacity region. Specifically, define the probability of error for each decoder $d \in \mathcal{D}$ as $p_d^{n,\mathrm{err}}(\mathbf{R}) = \mathbb{P}(g_d^n(U_{\mathrm{Fan}(D_d)}^n) \neq \beta(D_d)^{1:n})$, and the maximum over these as $p^{n,\mathrm{err}}(\mathbf{R}) = \max_{d \in \mathcal{D}} p_d^{n,err}$. A rate vector is in the rate region, $\mathbf{R} \in \mathcal{R}_*(\mathsf{A})$, is achievable if there exists a sequence of encoders $\{f_e^n\}$ and decoders $\{g_d^n\}$ such that $p^{n,\mathrm{err}}(\mathbf{R}) \to 0$ as $n \to \infty$, and the closure of the set of achievable rate vectors is the capacity region.

Extending the formula in [2] and following similar procedure as in [6], [13], the expression of the rate region of an IDSC problem expressed in terms of region of entropic vectors Γ_N^* and some network constraints, where N is the number of

variables in the network including both the source and encoder variables. Note that we assume that the output of a source will be the associated source variable itself and therefore we do not have the source coding constraints (usually represented as $\mathcal{L}_2$) for IDSC problems. Therefore, the rate region of an IDSC instance A is

$$\mathcal{R}_*(\mathsf{A}) = \mathrm{Proj}_{\boldsymbol{r},\boldsymbol{\omega}}(\overline{\mathrm{con}(\Gamma^*_{\mathrm{N}} \cap \mathcal{L}_{1,3})} \cap \mathcal{L}_{4',5}), \quad (1)$$

where $\mathrm{con}(\mathcal{B})$ is the conic hull of $\mathcal{B}$, and $\mathrm{Proj}_{\boldsymbol{r},\boldsymbol{\omega}}(\mathcal{B})$ is the projection of the set $\mathcal{B}$ on the coordinates $\left[\boldsymbol{r}^T, \boldsymbol{\omega}^T\right]^T$ where $\boldsymbol{r} = [R_e | e \in \mathcal{E}]$ and $\boldsymbol{\omega} = [H(X_k), k \in 1:K]$. Further, Γ^*_N and $\mathcal{L}_i, i = 1, 3, 4', 5$ are viewed as subsets of $\mathbb{R}^M$, $M = 2^N - 1 + |\mathcal{E}|$, $N = K + |\mathcal{E}|$, with coordinates $[\boldsymbol{h}^T, \boldsymbol{r}^T]^T$, with $\boldsymbol{h} \in \mathbb{R}^{2^N-1}$ indexed by subsets of $\mathcal{N}$ as is usual in entropic vectors, $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{E}|}$ playing the role of the capacities of edges, and any unreferenced dimensions (e.g. $\boldsymbol{r}$ in Γ^*_N) are left unconstrained (e.g. $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{E}|}$ in Γ^*_N). The $\mathcal{L}_i, i = 1, 3, 4', 5$ are network constraints representing source independency, codings at encoders, edge capacity constraints, sink nodes decoding constraints, respectively:

$$\mathcal{L}_1 = \left\{ \mathbf{h} \in \mathbb{R}^M : h_{\mathbf{X}_{1:K}} = \sum_{k \in 1:K} h_{X_k} \right\} \quad (2)$$

$$\mathcal{L}_3 = \{\mathbf{h} \in \mathbb{R}^M : h_{\mathbf{U}_e|\mathbf{X}_{1:K}} = 0, \forall e \in \mathcal{E}\} \quad (3)$$

$$\mathcal{L}_{4'} = \{(\mathbf{h}^T, \boldsymbol{r}^T)^T \in \mathbb{R}_+^{2^N-1+|\mathcal{E}|} : R_e \geq h_{U_e}, e \in \mathcal{E}\} \quad (4)$$

$$\mathcal{L}_5 = \{\mathbf{h} \in \mathbb{R}^M : h_{\beta(t)|\mathbf{U}_{\mathrm{In}(D_d)}} = 0, \forall d \in \mathcal{D}\}. \quad (5)$$

and we will denote $\mathcal{L}(\mathsf{A}) = \mathcal{L}_1 \cap \mathcal{L}_2 \cap \mathcal{L}_{4'} \cap \mathcal{L}_5$.

Similar to the solutions in [3]–[6], [13], we will replace Γ^*_N with polyhedral inner and outer bounds, typically from $\mathbb{F}_q$ representable matroids and the Shannon outer bound Γ^o_N, respectively, to check if the bounds match, because Γ^*_N is not fully characterized and even not polyhedral for $N \geq 4$ [14]. With polyhedral outer and inner bounds on Γ^*_N, (1) becomes a polyhedral computation problem which involves applying some constraints onto a polyhedra and then projecting down onto some coordinates. If the outer and inner bounds on rate region match, we obtain the exact rate region.

Regarding the inner bounds, as shown in [3]–[6], [13], there are two types of inner bounds obtained from $\mathbb{F}_q$-representable matroids. One is $\Gamma^{s,q}_N$, which is obtained directly from the conic hull of rank functions of $\mathbb{F}_q$-representable matroids on N elements, and their representations are associated with *scalar codes* in the sense described in [4], [6]. The other inner bound, associated with *vector codes*, is $\Gamma^{v,q,N'}_N$, which is obtained by considering the conic hull of rank functions of polymatroids which are N-partition of the ground sets of $\mathbb{F}_q$-representable matroids on $N' > N$ elements as described in [4], [6]. The tightness of the inner bound $\Gamma^{v,q,N'}_N$ is increasingly tight as $N' \to \infty$. With these bounds, we have associated rate regions

$$\mathcal{R}_k(\mathsf{A}) = \mathrm{Proj}_{\boldsymbol{r},\boldsymbol{\omega}}(\Gamma^k_N \cap \mathcal{L}(\mathsf{A})), \quad k \in \{o, (s,q), (v,q,N')\}$$

We also consider another important inner bound for the rate region, the superposition coding rate region, under which sources are coded independently from one another in each encoder, and the output of each encoder is the concatenation of the separately coded messages across the different sources. The associated region in this case is determined by the max-flow min-cut bound [14]. In particular, let $\mathbf{v} = (R_e, H(X_k), R_e(X_k) | e \in \mathcal{E}, k \in 1:K)$, where $R_e(X_k)$ is the sub-rate of source X_k at encoder E_e. After considering all decoders in $\mathcal{D}$, we can get the superposition coding rate region for an IDSC instance A as

$$\mathcal{R}_{sp}(\mathsf{A}) := \mathrm{Proj}_{\boldsymbol{r},\boldsymbol{\omega}} \left\{ \mathbf{v} \;\middle|\; \begin{array}{c} R_e = \sum_{i=1}^K R_e(X_i),\ e \in \mathcal{E}, \\ \sum_{e \in \mathrm{Fan}(D_d)} R_e(X_i) \geq H(X_i), \\ X_i \in \beta(D_d), \forall d \in \mathcal{D}. \end{array} \right\}.$$

We will consider all these bounds above for the thousands of non-isomorphic IDSC problems and prove their rate regions in the next section. If an inner bound matches with the Shannon outer bound, we say the associated codes are sufficient, i.e., every extreme ray in the outer bound on the rate region can be achieved by such codes.

Table I: Sufficiency of codes for IDSC instances: Columns 4–6 show the number of instances that the rate region inner bounds match with the Shannon outer bound.

$(K, \lvert\mathcal{E}\rvert)$	$\lvert\mathcal{M}'\rvert$	$\lvert\mathcal{M}\rvert$	$\mathcal{R}_{s,2}(\mathsf{A})$	$\mathcal{R}_{v,2,N+1}(\mathsf{A})$	$\mathcal{R}_{sp}(\mathsf{A})$
$(2,2)$	12	4	4	4	4
$(2,3)$	234	33	26	33	30
$(3,2)$	24	3	3	3	3
$(3,3)$	4752	179	143	179	148

IV. Computer Aided Proofs for Rate Regions

In this section, experimental results on thousands of IDSC instances are presented. We extended the ability of our computation package [15] to handle IDSC problems. We investigated rate regions for 219 non-isomorphic minimal IDSC instances representing 5130 isomorphic ones. These include the cases when $(K, |\mathcal{E}|) = (2,2), (2,3), (3,2), (3,3)$. For each non-isomorphic IDSC instance, we calculated several bounds on its rate region: the Shannon outer bound $\mathcal{R}_o$, the superposition coding inner bound $\mathcal{R}_{sp}$, the scalar binary representable matroid inner bound $\mathcal{R}_{s,2}$, the vector binary representable matroid inner bounds $\mathcal{R}_{v,2,N+1}$, where $N = K + |\mathcal{E}|$. If the outer bound on the rate region matches with an inner bound, we not only obtain the exact rate region but also know the codes that suffice to achieve any point in it.

Though it is infeasible to list all the 219 rate regions in this paper, a summary of results on the matches of various bounds is shown in Table I. The exact rate regions, their converses, and the codes that achieve them for all 219 non-isomorphic cases can be obtained at [16] and can be re-derived using [15]. For the non-isomorphic IDSC instances we considered, the Shannon outer bound is always tight on the rate regions, and the exact rate regions are obtained. Superposition coding does not always suffice for the IDSC instances. It suffices only for IDSC instances with $|\mathcal{E}| = 2$. Similarly, scalar binary codes also only suffice for the instances with $|\mathcal{E}| = 2$ but not for all instances with $|\mathcal{E}| = 3$. However, vector binary codes from binary matroids on $N + 1$ variables suffice for all the 219 instances. Thus, for the IDSC problems up to $K \leq 3, |\mathcal{E}| \leq 3$, vector binary codes suffice.

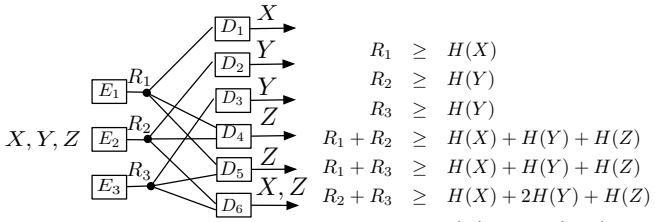

Figure 3: Block diagram and rate region $\mathcal{R}_*(\mathsf{A})$ for a $(3,3)$ IDSC instance A.

V. Example

In this section, an example is presented to show the tightness of various inner bounds on the rate region of an IDSC problem. Equivalently, we have computer aided proofs for both achievability part and converse part. In particular, consider the 3-source 3-encoder IDSC instance A with block diagram and rate region $\mathcal{R}_*(\mathsf{A})$ shown in Fig. 3.

Superposition coding is not optimal for this network. The superposition coding rate region is $\mathcal{R}_{sp}(\mathsf{A}) = \mathcal{R}_*(\mathsf{A}) \cap \{2R_1 + R_2 + R_3 \geq 3H(X) + 2H(Y) + 2H(Z)\}$. One of the extreme rays in the Shannon outer bound on rate region is $(R_1, R_2, R_3, H(X), H(Y), H(Z)) = (1, 1, 1, 1, 0, 1)$. This extreme ray cannot be achieved by superposition because when source X, Z are coded separately, the required coding rates will be 2 instead of 1 for at least one of the encoders.

Secondly, scalar binary codes do not suffice for this network, either. The scalar binary coding rate region is $\mathcal{R}_{s,2} = \mathcal{R}_*(\mathsf{A}) \cap \{R_1 + R_2 + R_3 \geq H(X) + 2H(Y) + 2H(Z)\}$. One of the extreme rays in the Shannon outer bound on rate region is $(R_1, R_2, R_3, H(X), H(Y), H(Z)) = (1, 1, 1, 0, 0, 2)$. This extreme ray cannot be achieved by scalar binary codes because no scalar code can encode a source with entropy of 2 into variables with entropy of 1.

However, vector binary codes suffice for this network. One can construct vector binary codes to achieve all extreme rays in the Shannon outer bound on the rate region. For instance, the extreme ray $(R_1, R_2, R_3, H(X), H(Y), H(Z)) = (1, 1, 1, 0, 0, 2)$ can be achieved by the vector binary code as follows. $U_1 = Z^1, U_2 = Z^2, U_3 = Z^1 + Z^2$, where Z^1, Z^2 are the two bits in source Z.

Finally, we consider the converse proof. As demonstrated in [6], automatic human readable proofs for MDCS instances can be generated by our software. The same procedure will give the proofs for IDSC instances as well. Take the last inequality in the rate region $\mathcal{R}_*(\mathsf{A})$ for instance, the inequalities and coefficients in Table II can be used to prove it as follows.

$$\begin{aligned} R_2 + R_3 &\overset{(1,2)}{\geq} H(U_2) + H(U_3) \overset{(3,4)}{=} H(Y, U_2) + H(Y, U_3) \\ &\overset{(5)}{\geq} H(Y) + H(Y, U_2, U_3) \\ &\overset{(6)}{\geq} H(Y) + H(X, Y, Z, U_2, U_3) - H(X, Z|U_2, U_3) \\ &\overset{(7,8)}{\geq} H(Y) + H(X, Y, Z) \overset{(9)}{=} H(X) + 2H(Y) + H(Z). \end{aligned}$$

These steps follow the capacities for $E_2 \& E_3$, the decoding constraints on $D_2 \& D_3$, the non-negativity of mutual information (twice), the decoding constraints on D_6, the source encoding constraints, and the source independence, resp.

Table II: Ordered inequalities with coefficients given by computer for proving $R_2 + R_3 \geq H(X) + 2H(Y) + H(Z)$ in Fig. 3.

Order	Coefficients	(In)equalities
1	1	$R_2 \geq H(U_2)$
2	1	$R_3 \geq H(U_3)$
3	1	$H(Y\|U_3) = 0$
4	1	$H(Y\|U_2) = 0$
5	1	$I(U_2; U_3\|Y) \geq 0$
6	1	$I(Y; X, Z\|U_2, U_3) \geq 0$
7	1	$H(XZ\|U_2, U_3) = 0$
8	1	$H(U_2, U_3\|X, Y, Z) \geq 0$
9	1	$H(X, Y, Z) = H(X) + H(Y) + H(Z)$

VI. Conclusion

This paper enumerated all 219 non-isomorphic minimal IDSC problems up to 3 sources and 3 encoders. For these IDSC instances, computation tools were used to calculate the various bounds on their rate regions and then their exact rate regions were obtained. Vector binary codes sufficed and the Shannon outer bound was tight for all the instances considered.

Acknowledgment

Support from NSF under CCF 1016588 & 1421828 is gratefully acknowledged.

References

[1] R. W. Yeung and Z. Zhang, "Distributed source coding for satellite communications," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 1111–1120, 1999.

[2] X. Yan, R. Yeung, and Z. Zhang, "An implicit characterization of the achievable rate region for acyclic multisource multisink network coding," *IEEE Trans. Inform. Theory*, vol. 58, no. 9, pp. 5625–5639, Sept 2012.

[3] C. Li, J. M. Walsh, and S. Weber, "Computational approaches for determining rate regions and codes using entropic vector bounds," in *50th Annual Allerton Conference on Communication, Control and Computing*, Oct 2012, pp. 1 –9.

[4] C. Li, J. Apte, J. M. Walsh, and S. Weber, "A new computational approach for determining rate regions and optimal codes for coded networks," in *IEEE Int. Symp. Network Coding*, Jun 2013, pp. 1 –6.

[5] J. Apte, C. Li, and J. Walsh, "Algorithms for computing network coding rate regions via single element extensions of matroids," in *2014 IEEE International Symposium on Information Theory (ISIT)*, June 2014.

[6] C. Li, S. Weber, and J. M. Walsh, "Multilevel diversity coding systems: Rate regions, codes, computation, & forbidden minors," *CoRR*, vol. abs/1407.5659, 2014.

[7] R. W. Yeung, "Multilevel diversity coding with distortion," *IEEE Trans. Information Theory*, vol. 41, pp. 412–422, 1995.

[8] K. P. Hau, "Multilevel diversity coding with independent data streams," Master's thesis, The Chinese University of Hong Kong, June 1995.

[9] J. R. Roche, R. W. Yeung, and K. P. Hau, "Symmetrical multilevel diversity coding," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 1059–1064, 1997.

[10] S. Mohajer, C. Tian, and S. Diggavi, "Asymmetric multilevel diversity coding and asymmetric gaussian multiple descriptions," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4367–4387, 2010.

[11] C. Li, "On multi-source multi-sink hyperedge networks: enumeration, rate region computation, and hierarchy," Ph.D. dissertation, Drexel University, 2015.

[12] A. Betten, M. Braun, H. Fripertinger, A. Kerber, A. Kohnert, and A. Wassermann, *Error-Correcting Linear Codes: Classification by Isometry and Applications*, ser. Algorithms and Computation in Mathematics. Springer Berlin Heidelberg, 2006.

[13] C. Li, S. Weber, and J. Walsh, "Network embedding operations preserving the insufficiency of linear network codes," in *52nd Annual Allerton Conf. on Communication, Control, and Computing*, Oct 2014.

[14] R. W. Yeung, *Information Theory and Network Coding*. Springer, 2008.

[15] C. Li, J. M. Walsh, and S. Weber, "Software for computing bounds on entropic vectors region and network rate region," available at http://www.ece.drexel.edu/walsh/aspitrg/software.html.

[16] ——, "Exact Rate Regions and Codes for all (K, E) IDSCs with $K, E \in \{2, 3\}$." available at http://goo.gl/WP80CK.