A Study of Semantic Data Compression

Basak Guler Aylin Yener
Electrical Engineering Department
The Pennsylvania State University, University Park, PA 16802
bxg215@psu.edu *yener@ee.psu.edu*

Prithwish Basu
Raytheon BBN Technologies
Cambridge, MA
pbasu@bbn.com

***Abstract*—A two-way semantic model is considered with two sources sharing their ideas chosen from different sets of facts. These facts may be expressed in the form of RDF (Resource Description Framework) triples. A set of conclusions can be derived by using the logical relations between these facts. This set of conclusions depends on the current interest of the network, thus not all combinations of facts lead to a useful conclusion. Users are interested in sharing only the facts that lead to these conclusions. Additionally, users do not want to use extra resources for sharing the facts that lead to the same conclusions. We consider the worst-case semantic communication performance of this network. We provide upper and lower bounds for each user to learn useful facts from one another, and show that increasing the number of rounds of interaction can improve the worst-case performance over the existing schemes by reducing the total number of bits transmitted.**

***Index Terms*—Semantic compression, interactive communication, semantic networks.**

I. Introduction

Sources in modern communication networks, humans, computers, or smart devices, aim at sharing meaningful and useful information, instead of merely delivering any information and maximizing the throughput. The impact of the meaning of messages delivered by the sources have been explored in [1]–[3] by extending the classical information theory framework. The communication problem we envision in this paper is the worst-case data compression performance in a two-way channel to convey information that leads to useful conclusions, instead of merely conveying any information. A related work is [5] which studies the distributed source coding problem to show that interaction can improve the performance of function computation by reducing the total rate requirement, as opposed to reconstructing sources.

We introduce our scheme with the following example. Consider a network with two persons. The first person has 3 facts represented by $\mathcal{X} = \{x_1, x_2, x_3\}$, whereas the second person has 2 facts $\mathcal{Y} = \{y_1, y_2\}$. A single fact is conveyed by each person. Users are interested in 2 conclusions $\{c_1, c_2\} = \mathcal{C}$ drawn from these facts, $x_1 \wedge y_2 \to c_1$ and $x_3 \wedge y_1 \to c_2$. Therefore there are only 2 desired conclusions instead of $|\mathcal{X}| \cdot |\mathcal{Y}| = 6$. We assume that persons are not allowed to share the conclusions directly, but they share the facts that lead to the desired conclusions. We assume both parties have a shared knowledge of the mapping between the facts and conclusions before communication takes place. The idea is that communicating any combination of facts is not always useful. Specifically, a person may choose to ignore a message if it is not relevant, or when the fact does not help infer any necessary conclusion in the current situation. We refer to this network problem as a *semantic communication* problem, as the logical relations define the semantic aspects of the messages and the communicating parties are concerned in obtaining the facts that lead to meaningful conclusions, instead of encoding and decoding any fact as in conventional communication systems. We show that this modeling can greatly reduce the number of bits to be transmitted from both parties to learn each other's message.

II. System Model

We consider a network model in which two persons are communicating using a two-way channel. The first person knows the facts $x_i \in \mathcal{X}$ where $\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}$ and the second person knows the facts $y_j \in \mathcal{Y}$ with $\mathcal{Y} = \{y_1, \ldots, y_{|\mathcal{Y}|}\}$. These facts refer to the logical symbols generated by each source. The first person wants to learn the second person's fact whereas the second person wants to learn the fact of the first person as long as these facts lead to a conclusion. We assume the relations between the facts and conclusions are restricted to conjunctive expressions in propositional logic [4]. We define a subset $\mathcal{S}$ of $\mathcal{X} \times \mathcal{Y}$ to refer to the set of tuples of facts from the two users that result in a desired conclusion:

$$\mathcal{S} = \{(x_i, y_j) : x_i \wedge y_j \to c_k,\ c_k \in \mathcal{C},\ x_i \in \mathcal{X},\ y_j \in \mathcal{Y}\} \tag{1}$$

where $\mathcal{C}$ denotes the set of all desired conclusions. In case the facts of two parties do not belong to $\mathcal{S}$, i.e., they do not result in a desired conclusion, users are not interested in learning each other's fact. We assume a noiseless channel exists between the persons and focus on the source coding problem. We consider a zero-error, two-way transmission scheme with a deterministic encoding-decoding protocol with multiple rounds. The mapping ϕ is used to encode the facts from $\mathcal{S}$ to codewords represented by bit streams. Let $\phi(x_i, y_j) = [\phi_k(x_i, y_j)]_{k=1}^r$ be the sequence of codewords exchanged during an r-round communication. $\phi_k(x_i, y_j)$ represents the codewords transmitted from both parties at round k:

$$\phi_k(x_i, y_j) = [\phi_k^X(x_i, y_j), \phi_k^Y(x_i, y_j)] \tag{2}$$

where $\phi_k^X(x_i, y_j)$ is the codeword transmitted from the first person and $\phi_k^Y(x_i, y_j)$ is the codeword from the second person at round k. We also define $\phi^X(x_i, y_j) = [\phi_k^X(x_i, y_j)]_{k=1}^r$ and $\phi^Y(x_i, y_j) = [\phi_k^Y(x_i, y_j)]_{k=1}^r$ as the sequences of codewords transmitted from the first and second users in r rounds. Since we assume a two-way transmission scheme, both persons can transmit arbitrary length codewords simultaneously. We allow null transmissions. The maximal length codeword for the mapping ϕ is defined as:

$$l(\phi) = \max_{\mathcal{S}} \quad |\phi(x_i, y_j)|, \quad (x_i, y_j) \in \mathcal{S} \tag{3}$$

Our aim is to find the best encoding scheme in terms of the maximal codeword length, which is given as follows:

$$\bar{l} = \min_{\phi} \quad l(\phi) \tag{4}$$

Since useful fact pairs are only a subset of the fact space, each fact from one person is partially connected to the facts from the other person. We define the ambiguity set $\mathcal{A}_i^X$ as the set of all possible facts from the second person that lead to a conclusion with x_i:

$$\mathcal{A}_i^X = \{y_j : x_i \wedge y_j \to c_k,\ y_j \in \mathcal{Y},\ c_k \in \mathcal{C}\} \tag{5}$$

Similarly, define the ambiguity set $\mathcal{A}_j^Y$ for every fact y_j as:

$$\mathcal{A}_j^Y = \{x_i : x_i \wedge y_j \to c_k,\ x_i \in \mathcal{X},\ c_k \in \mathcal{C}\} \tag{6}$$

978-1-4799-0248-4/13/$31.00 ©2013 IEEE

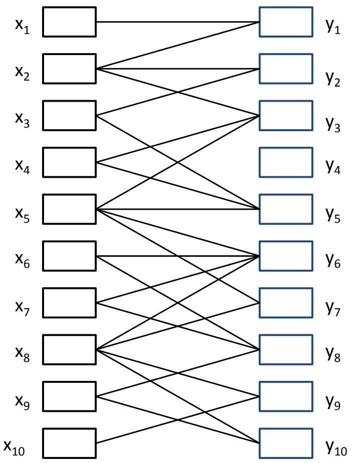

Fig. 1. Network model with 2 users and 10 facts for each user.

where $i = 1, \dots, |\mathcal{X}|$, $j = 1, \dots, |\mathcal{Y}|$. The number of elements in each set is given by $|\mathcal{A}_i^X| = d_i$ and $|\mathcal{A}_j^Y| = m_j$, respectively, bounded by a maximum $d_i \leq d$ and $m_j \leq m$ for some $d, m \in \mathbb{N}$ for every i, j. We note that these relations are purely semantical and that any pair of facts can appear with nonzero probability whether they are meaningful together or not. However, persons are interested in recovering messages only if they lead to a conclusion, irrespective of their statistical frequencies.

Lemma 1. *[6] Let $(x_i, y_j), (x_{\hat{i}}, y_j), (x_{\hat{i}}, y_{\hat{j}}) \in \mathcal{S}$ be fact pairs such that $i, \hat{i} \in \mathcal{X}, i \neq \hat{i}$ and $j, \hat{j} \in \mathcal{Y},\ j \neq \hat{j}$. If $\phi(x_i, y_j)$ or $\phi(x_{\hat{i}}, y_{\hat{j}})$ is a prefix of the other, then the following holds:*

$$\phi(x_i, y_j) = \phi(x_{\hat{i}}, y_j) = \phi(x_{\hat{i}}, y_{\hat{j}}) \tag{7}$$

Proposition 1. *The set of codewords for the facts in the ambiguity set $\mathcal{A}_i^X$ for each $x_i \in \mathcal{X}$ is prefix-free. Similarly, the set of codewords in $\mathcal{A}_j^Y$ for each $y_j \in \mathcal{Y}$ is prefix-free.*

This proposition follows from Lemma 1 and the property that the codewords corresponding to the facts in the ambiguity set have to be prefix-free for the receiver to interpret the codeword correctly for error-free decoding.

III. Lower Bound on Semantic Codeword Length

In this section, a lower bound is derived for maximal code length of the two-way semantic network.

Theorem 1. *Let i_o be the index of the first fact x_{i_o} such that $|\mathcal{A}_{i_o}| = d$. Similarly, define j_o as the index of the fact y_{j_o} such that $|\mathcal{A}_{j_o}| = m$. A lower bound on the maximal codeword length of the two-way semantic network is given as:*

$$\bar{l} \geq \max_{\substack{(x_{i_o}, y_j) \in \mathcal{S} \\ (x_i, y_{j_o}) \in \mathcal{S}}} \{\lceil \log(d) \rceil + \lceil \log(m_j) \rceil, \lceil \log(d_i) \rceil + \lceil \log(m) \rceil\} \tag{8}$$

Proof: The following result follows from Proposition 1:

$$\phi^X(x_i, y_j) \geq \lceil \log(|\mathcal{A}_i^X|) \rceil = \lceil \log(d_i) \rceil \tag{9}$$

$$\phi^Y(x_i, y_j) \geq \lceil \log(|\mathcal{A}_j^Y|) \rceil = \lceil \log(m_j) \rceil \tag{10}$$

for all $x_i \in \mathcal{X}$ and $y_j \in \mathcal{Y}$. Then the worst-case codeword length for the mapping ϕ satisfies:

$$\bar{l} = \min_{\phi} \max_{(x_i, y_j) \in \mathcal{S}} \quad |\phi(x_i, y_j)| \tag{11}$$

$$= \min_{\phi} \max_{(x_i, y_j) \in \mathcal{S}} \quad (|\phi^X(x_i, y_j)| + |\phi^Y(x_i, y_j)|) \tag{12}$$

$$\geq \max_{(x_i, y_j) \in \mathcal{S}} \quad (\lceil \log(|\mathcal{A}_i^X|) \rceil + \lceil \log(|\mathcal{A}_j^Y|) \rceil) \tag{13}$$

$$= \max_{(x_i, y_j) \in \mathcal{S}} \quad (\lceil \log(d_i) \rceil + \lceil \log(m_j) \rceil) \tag{14}$$

$$\geq \max_{\substack{(x_{i_o}, y_j) \in \mathcal{S} \\ (x_i, y_{j_o}) \in \mathcal{S}}} \{\lceil \log(d) \rceil + \lceil \log(m_j) \rceil, \lceil \log(d_i) \rceil + \lceil \log(m) \rceil\} \tag{15}$$

■

IV. Semantic Relations Between Facts and the Upper Bound

In this section we consider the upper bound on the encoding schemes for sharing the facts between the two parties. We start with the following naive upper bound.

Theorem 2. *The maximum code length for the two-way semantic network with one round of interaction is:*

$$\bar{l} \leq \lceil \log(\chi(\mathcal{G}_X)) \rceil + \lceil \log(\chi(\mathcal{G}_Y)) \rceil \tag{16}$$

where $\chi(\mathcal{G}_X)$ and $\chi(\mathcal{G}_Y)$ are the chromatic numbers of the characteristic graphs $\mathcal{G}_X$ and $\mathcal{G}_Y$, respectively.

Proof: Let

$$\mathcal{R}_X = \{x_i : (x_i, y_j) \in \mathcal{S},\ x_i \in \mathcal{X},\ y_j \in \mathcal{Y}\} \tag{17}$$

$$\mathcal{R}_Y = \{y_j : (x_i, y_j) \in \mathcal{S},\ x_i \in \mathcal{X},\ y_j \in \mathcal{Y}\} \tag{18}$$

Define a characteristic graph $\mathcal{G}_X = (V_X, E_X)$ for the first person with the vertex set $V_X = \mathcal{R}_X$ and an edge $(x_i, x_{\hat{i}})$ if there exists y_j such that $x_i \wedge y_j \rightarrow c_k$, $x_{\hat{i}} \wedge y_j \rightarrow c_{\hat{k}}$ and $c_k \neq c_{\hat{k}}$. Denote the chromatic number of this graph by $\chi(\mathcal{G}_X)$. The first person sends the index of the color of her fact, which requires no more than $\lceil \log(\chi(\mathcal{G}_X)) \rceil$ bits. A similar graph $\mathcal{G}_Y = (V_Y, E_Y)$ is defined for the second user, who sends the index of the color of his fact by using at most $\lceil \log(\chi(\mathcal{G}_Y)) \rceil$ bits. Now, both parties can use their own facts to determine the conclusion. This communication takes only one round and requires $\lceil \log(\chi(\mathcal{G}_X)) \rceil + \lceil \log(\chi(\mathcal{G}_Y)) \rceil$ bits in the worst-case. ■

Observe that the above approach may become inefficient when the number of facts in the support set increases. We want to know if semantics can help to reduce the transmitted number of bits and whether source coding can benefit from logical relations. We first utilize a coding scheme from interactive communication to show that allowing another round of interaction between the two parties can greatly reduce the semantic compression rate. Later, we propose a scheme to improve the upper bound by allowing users to take turns in multiple rounds, even when the number of non-empty transmissions from each user stays the same. In order to show this, we utilize the following hypergraph partitioning results:

Lemma 2. *[7] If for all $x_i \in \mathcal{X}$, $|\mathcal{A}_i^X| \leq d$ and for all $y_j \in \mathcal{Y}$, $|\mathcal{A}_j^Y| \leq m$, then the worst-case codeword length is bounded by:*

$$\bar{l} \leq \lceil \log(dm) \rceil + \lceil \log(\min(d, m)) \rceil \tag{19}$$

Lemma 3. *[8] Let $H = (V, E)$ be a hypergraph with a vertex set V of size $|V|$. Hyperedges are given as $E_i \subseteq V$ for $i = 1, \dots, |E|$ and each hyperedge consists of at most d elements, i.e., $|E_i| \leq d$. Given $\epsilon > 0$, there exists a constant $c(\epsilon)$ such that $\forall p \geq (\ln \sqrt{|V||E|})^{1+\epsilon}$ and $p > 1$, a partition $V_1, V_2, \dots V_{\lceil \frac{d}{p} c(\epsilon) \rceil}$ of V can be found with the property $|V_k \cap E_i| < p$, for $i = 1, \dots, |E|$ and $k = 1, \dots, \lceil \frac{d}{p} c(\epsilon) \rceil$.*

We introduce the following definitions to be used in the remaining part of the paper. Define the set of colors used for the characteristic graphs $\mathcal{G}_X$ and $\mathcal{G}_Y$ in Theorem 2 as $\mathcal{Q}_X = \{q_1^X, \dots, q_{|\mathcal{Q}_X|}^X\}$ and $\mathcal{Q}_Y = \{q_1^Y, \dots, q_{|\mathcal{Q}_Y|}^Y\}$, respectively, with chromatic numbers

$|\mathcal{Q}_X| = \chi(\mathcal{G}_X)$ and $|\mathcal{Q}_Y| = \chi(\mathcal{G}_Y)$. Let $q(x_i)$ and $q(y_j)$ be the colors assigned to the facts x_i and y_j. Then, we define the ambiguity set $\mathcal{T}_n^X$ of color q_n^X as follows:

$$\mathcal{T}_n^X = \{q_z^Y : x_i \wedge y_j \rightarrow c_k,\ x_i \in \mathcal{X},\ q(x_i) = q_n^X,\ y_j \in \mathcal{Y},\\ q(y_j) = q_z^Y,\ q_z^Y \in \mathcal{Q}_Y, c_k \in \mathcal{C}\}, \quad n = 1, \dots, |\mathcal{Q}_X| \quad (20)$$

Similarly, define the ambiguity set $\mathcal{T}_j^Y$ for each q_j^Y as:

$$\mathcal{T}_z^Y = \{q_n^X : x_i \wedge y_j \rightarrow c_k,\ x_i \in \mathcal{X},\ q(x_i) = q_n^X,\ q_n^X \in \mathcal{Q}_X,\\ y_j \in \mathcal{Y},\ q(y_j) = q_z^Y, c_k \in \mathcal{C}\}, \quad z = 1, \dots, |\mathcal{Q}_Y| \quad (21)$$

The following result for the semantic two-way source coding problem shows that logical relations can improve the semantic source coding performance:

Theorem 3. *Given $|\mathcal{T}_n^X| \leq d$ for $n = 1, \dots, |\mathcal{Q}_X|$ and $|\mathcal{T}_z^Y| \leq m$ for $z = 1, \dots, |\mathcal{Q}_Y|$. Then*

$$\bar{l} \leq \lceil \log(d) \rceil + \lceil \log(m) \rceil\\ + (1+\epsilon) \log\log(\sqrt{\chi(\mathcal{G}_X)\chi(\mathcal{G}_Y)}) + 2\log\, c(\epsilon) + 5 \quad (22)$$

Proof: The proof follows from Lemma 3 and Theorem 4 in [8]. The idea is to use hypergraph partitioning to obtain a partition of $\mathcal{Q}_X$ and $\mathcal{Q}_Y$ such that in each partition, the set of colors that lead a conclusion when combined with the color of the fact from the other user has a number of elements no greater than p.

Specifically, there exists a partition of $\mathcal{Q}_X$ as $\mathcal{Q}_{X1}, \mathcal{Q}_{X2}, \dots \mathcal{Q}_{X\lceil \frac{m}{p} c(\epsilon) \rceil}$, such that for any q_z^Y:

$$|\mathcal{Q}_{Xu} \cap q_n^X : x_i \wedge y_j \rightarrow c_k| \leq p, \quad (23)$$

for $x_i \in \mathcal{X},\ q(x_i) = q_n^X,\ q_n^X \in \mathcal{Q}_X,\ y_j \in \mathcal{Y},\ q(y_j) = q_z^Y,\ c_k \in \mathcal{C},\ u = 1, \dots, \lceil \frac{m}{p} c(\epsilon) \rceil$. Similarly, a partition exists for the second user on $\mathcal{Q}_Y$ as $\mathcal{Q}_{Y1}, \mathcal{Q}_{Y2}, \dots, \mathcal{Q}_{Y\lceil \frac{d}{p} c(\epsilon) \rceil}$, such that for any q_n^x:

$$|\mathcal{Q}_{Yu} \cap q_z^Y : x_i \wedge y_j \rightarrow c_k| \leq p, \quad (24)$$

with $x_i \in \mathcal{X},\ q(x_i) = q_n^X,\ y_j \in \mathcal{Y},\ q(y_j) = q_z^Y,\ q_z^Y \in \mathcal{Q}_Y,\ c_k \in \mathcal{C},\ u = 1, \dots, \lceil \frac{d}{p} c(\epsilon) \rceil$. Then the two-way communication takes place as follows. Let $p = (\ln \sqrt{|\mathcal{Q}_X||\mathcal{Q}_Y|})^{1+\epsilon}$. The first person sends the index of the partition of $\mathcal{Q}_X$ that the color of her fact is in. Note that this requires no more than $\lceil \log(\frac{m}{p} c(\epsilon)) \rceil$ bits. On the other side, the second person sends the index of the partition of $\mathcal{Q}_Y$ that the color of his fact lies in by using at most $\lceil \log(\frac{d}{p} c(\epsilon)) \rceil$ bits. The first person can use her color and the received partition index to restrict the possible second user colors in a p-dimensional subspace. The second person can use a similar elimination method, leaving at most p possible colors from the first user. Thus, communication is now limited to an at most $p \times p$ subset of $\mathcal{Q}_X \times \mathcal{Q}_Y$. The number of bits required for both parties to learn both facts is no more than $3\lceil \log(p) \rceil$, which is obtained by using $m = d = p$ in Lemma 2.

Thus the total number of bits is given as:

$$\bar{l} \leq \lceil \log(\frac{m}{p} c(\epsilon)) \rceil + \lceil \log(\frac{d}{p} c(\epsilon)) \rceil + 3\lceil \log(p) \rceil \quad (25)$$

$$\leq (\log(\frac{m}{p} c(\epsilon)) + 1) + (\log(\frac{d}{p} c(\epsilon)) + 1) + 3(\log(p) + 1) \quad (26)$$

$$\leq \log(m) + \log(d) + (1+\epsilon) \log\ln(\sqrt{|\mathcal{Q}_X||\mathcal{Q}_Y|})\\ + 2\log\, c(\epsilon) + 5 \quad (27)$$

$$\leq \lceil \log(d) \rceil + \lceil \log(m) \rceil + (1+\epsilon) \log\log(\sqrt{|\mathcal{Q}_X||\mathcal{Q}_Y|})\\ + 2\log\, c(\epsilon) + 5 \quad (28)$$

■

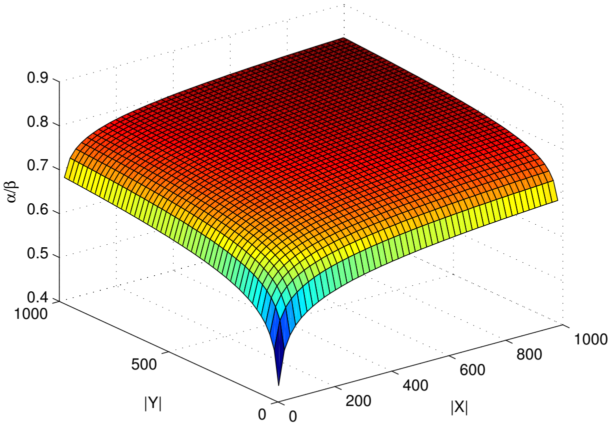

Fig. 2. Various set sizes for $\mathcal{X}$ and $\mathcal{Y}$ vs. fraction α/β.

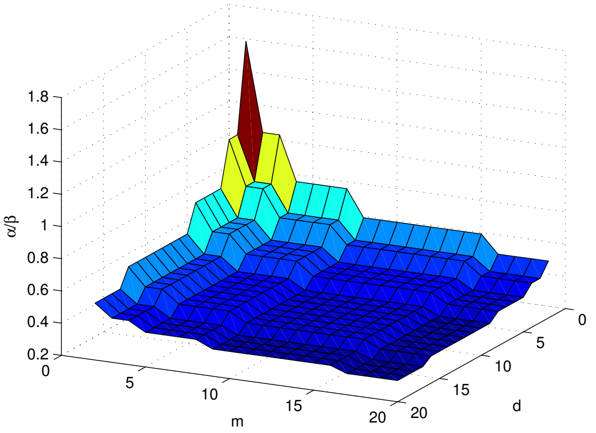

Fig. 3. Various d and m values vs. fraction α/β.

This upper bound is achieved by three rounds of communication when each person makes two non-empty transmissions. We assume the partitioning protocol is agreed upon by the two parties before communication takes place. If d and m are such that $d \simeq m$, then we observe that $\lceil \log(d) \rceil + \lceil \log(m) \rceil \gg \log\log(\sqrt{|\mathcal{Q}_X||\mathcal{Q}_Y|})$ for almost all values of d and m.

Next we show that the upper and lower bounds are tight for some graphs. Let us consider the following example. Define a network as in Fig. 1 with the following condition. Let n' be an index of a color such that $\mathcal{T}_{n'}^X = d$ and z' be the color such that $\mathcal{T}_{z'}^Y = m$. Assume that $q_{n'}^X \in \mathcal{T}_{z'}^Y$ or $q_{z'}^Y \in \mathcal{T}_{n'}^X$. Then the lower bound becomes $\bar{l} \geq \lceil \log(d) \rceil + \lceil \log(m) \rceil$. Thus, for the case when $d \simeq m$, the upper bound is tight. However, if this condition is not satisfied, the upper bound can grow significantly when the fact spaces become larger.

We present the relation between the worst case upper and lower bounds for various d, m values and set sizes in Fig. 2 and Fig.3. For the sake of this example, we assume every fact pair leads to a different conclusion, and thus the characteristic graphs for both users are complete graphs where $\chi(\mathcal{G}_X) = \mathcal{X}$ and $\chi(\mathcal{G}_Y) = \mathcal{Y}$. We define $\alpha = \log\log(\sqrt{|\mathcal{X}||\mathcal{Y}|})$ and $\beta = \lceil \log(d) \rceil + \lceil \log(m) \rceil$. Fig. 2 shows various set sizes for $\mathcal{X}$ and $\mathcal{Y}$ vs. the fraction α/β for fixed $d = 2$ and $m = 5$. We present in Fig. 3 the relation between various d and m values and the fraction α/β for $|\mathcal{X}| = 1000$ and $|\mathcal{Y}| = 1000$. The figures suggest that, when d and m values are small with respect to the set sizes, there is a significant gap between the upper and lower bounds, which reduces only when d and m start approaching the set sizes.

Next, we investigate whether we can improve this upper bound and reduce the gap by allowing multiple rounds of communication. This allows the proposed scheme to work in low-rate communication environments when sources have limited bandwidth but do not mind having extra rounds of interaction. The following scheme shows that the upper bound can be improved by increasing the total number of rounds, even in the case when the number of non-empty transmissions from each user stays the same.

Consider the characteristic graph used in Theorems 2 and 3. Denote the set of colors in the first round by $\mathcal{Q}_X^1 = \mathcal{Q}_X$ and $\mathcal{Q}_Y^1 = \mathcal{Q}_Y$ for the first and the second user, respectively. We assume $d > m$ without loss of generality. Let $p_1 = (\ln\sqrt{|\mathcal{Q}_X^1||\mathcal{Q}_Y^1|})^{1+\epsilon}$. From Lemma 3, $\mathcal{Q}_X^1$ can be partitioned into $\lceil \frac{m}{p_1} c(\epsilon)\rceil$ groups such that for each partition u:

$$|\mathcal{Q}_{Xu}^1 \cap q_n^X : x_i \wedge y_j \rightarrow c_k| \leq p_1, \tag{29}$$

where $x_i \in \mathcal{X},\ y_j \in \mathcal{Y},\ \ q(x_i) = q_n^X,\ q_n^X \in \mathcal{Q}_X^1,\ c_k \in \mathcal{C}$ and $u = 1, \ldots, \lceil \frac{m}{p_1} c(\epsilon)\rceil$. In this round, the first person sends the index of the partition that her fact resides in by using no more than $\lceil \log(\frac{m}{p_1} c(\epsilon))\rceil$ bits, and the second person makes an empty transmission. Let $\hat{u}$ be the index of the partition sent by the first person in the first round. In the second round, after receiving the index from the first user, the second person considers the following set:

$$\begin{aligned}\mathcal{Q}_Y^2 = \{q_z^Y : x_i \wedge y_j \rightarrow c_k,\ x_i \in \mathcal{X},\ q(x_i) \in \mathcal{Q}_{X\hat{u}}^1 \\ y_j \in \mathcal{Y},\ q(y_j) = q_z^Y, q_z^Y \in Q_Y^1\}\end{aligned} \tag{30}$$

Note that $|\mathcal{Q}_Y^2| \leq \min\{d|\mathcal{Q}_{X\hat{u}}^1|, |\mathcal{Q}_Y^1|\}$. Next, consider a hypergraph $\mathcal{H} = (V, E)$ with the vertex set $V = \mathcal{Q}_Y^2$ and edges $E = \{E_1, \ldots, E_{|E|}\}$ where

$$\begin{aligned}E_n = \{q_z^Y : x_i \wedge y_j \rightarrow c_k,\ x_i \in \mathcal{X},\ q(x_i) = q_n^X \\ y_j \in \mathcal{Y},\ q(y_j) = q_z^Y,\ q_z^Y \in \mathcal{Q}_Y^2\}.\end{aligned} \tag{31}$$

for each $n = 1, \ldots, |\mathcal{Q}_{X\hat{u}}^1|$. Thus the total number of edges is $|E| = |\mathcal{Q}_{X\hat{u}}^1|$, and the number of elements in each hyperedge satisfies:

$$|E_n| \leq m, \quad n = 1, \ldots, |\mathcal{Q}_{X\hat{u}}^1| \tag{32}$$

We note that the first person can also determine this set by using the partition index for her color and the logical relations between the two facts of both parties. We then define the following variable:

$$p_2 = (\ln\sqrt{|V||E|})^{1+\epsilon} = (\ln\sqrt{|\mathcal{Q}_Y^2||\mathcal{Q}_{X\hat{u}}^1|})^{1+\epsilon} < p_1 \tag{33}$$

Then the second user can partition $\mathcal{Q}_Y^2$ into $\lceil \frac{d}{p_2} c(\epsilon)\rceil$ groups and sends the index of his fact's color, which requires no more than $\lceil \log(\frac{m}{p_2} c(\epsilon))\rceil$ bits. After receiving the partition index, the first user can use her color to reduce the number of possible colors from the second user to p_2.

In the third round, colors are now restricted to a $p_1 \times p_2$ dimensional subspace of $\mathcal{Q}_X \times \mathcal{Q}_Y$. Worst-case codeword length can then be bounded using Lemma 2 by substituting $d = p_1$ and $m = p_2$. Thus the number of bits required to recover the colors is no more than $\lceil \log(p_1)\rceil + 2\lceil \log(p_2)\rceil$. We also note that the same constant $c(\epsilon)$ can be used for both partitions. This follows from the construction of the constant $c(\epsilon)$ in [8], from which it follows that a constant that holds for p_1 also holds for $p_2 < p_1$. The following theorem provides the new upper bound.

Theorem 4. *The new upper bound for the worst-case code length for the two-way interactive semantic network is given by:*

$$\begin{aligned}\bar{l}_{new} \leq &\log(m) + \log(d) + (1+\epsilon)\log\log(\sqrt{|\mathcal{Q}_{X\hat{u}}^1||\mathcal{Q}_Y^2|}) \\ &+ 2\log\ c(\epsilon) + 5\end{aligned} \tag{34}$$

Proof: The total number of bits required satisfies:

$$\begin{aligned}\bar{l}_{new} \leq\ &\lceil \log(\frac{m}{p_1} c(\epsilon))\rceil + \lceil \log(\frac{d}{p_2} c(\epsilon))\rceil + \lceil \log(p_1)\rceil \\ &+ 2\lceil \log(p_2)\rceil \end{aligned} \tag{35}$$

$$\leq \log(m) + \log(d) + \log(p_2) + 2\log\ c(\epsilon) + 5 \tag{36}$$

$$\begin{aligned}\leq\ &\lceil \log(d)\rceil + \lceil \log(m)\rceil + (1+\epsilon)\log\log(\sqrt{|\mathcal{Q}_{X\hat{u}}^1||\mathcal{Q}_Y^2|}) \\ &+ 2\log\ c(\epsilon) + 5\end{aligned} \tag{37}$$

■

This scheme requires four rounds of interaction in total. However, each person again makes two non-empty transmissions as in Theorem 3. The new upper bound satisfies the following:

$$\bar{l}_{new} < \bar{l} \tag{38}$$

The number of rounds can be further increased to reduce the effect of set sizes of the facts on the upper bound. Due to space concerns, we only present that this follows from applying hypergraph partitioning within a partition described by our method sequentially. As the set sizes decrease in each round, the third term in (34) contributed by p also decreases, whereas the first two terms stay the same, with a different constant term. Thus the term that depends on the set sizes decreases, which is a desirable property for large networks.

V. Conclusion

In this paper, we have considered a semantic network with two sources interacting to share their facts. Depending on the network structure, some desired conclusions may be drawn from these facts. The two parties are interested in sharing only the facts that lead to desired conclusions. We have investigated lower and upper bounds for worst-case performance. We have proposed a method for utilizing the logical relationships between these facts, and show that performance can be improved by increasing the number of rounds of interaction. Future work includes semantic source coding with multiple sources and different world interpretations among the network entities.

References

[1] R. Carnap, Y. Bar-Hillel, "An outline of a theory of semantic information," *RLE Technical Reports*, vol. 247, MIT, Cambridge, MA, Oct. 1952.

[2] L. Floridi, "Outline of a theory of strongly semantic information," *Minds Mach.*, vol. 14, no. 2, pp. 197-221, 2004.

[3] F. M. Willems, T. Kalker, "Semantic compaction, transmission, and compression codes," *Proc. of Int. Symp. on Inf. Theory (ISIT)*, 2005, pp. 214218.

[4] P. Basu, J. Bao, M. Dean, J. A. Hendler, "Preserving quality of information by using semantic relationships," *PerCom Workshops,* 58-63, 2012.

[5] N. Ma, P. Ishwar, "Some results on distributed source coding for interactive function computation," *IEEE Trans. on Inf. Theory*, vol. 57, pp. 6180-6195, Sep. 2011.

[6] A. Orlitsky, "Worst-Case Interactive Communication I: Two Messages are Almost Optimal," *IEEE Trans. on Inf. Theory,* vol. 36, no. 5, Sept. 1990.

[7] R. Ahlswede, "Coloring Hypergraphs: A New Approach to Multi-user Coding," *J. of Combinatorics, Information & System Sciences,* vol. 4, no. 1, pp. 76-115.

[8] A. El Gamal, A. Orlitsky, "Interactive Data Compression," *25th Annual Symposium on Foundations of Computer Science,* pp. 100-108, 24-26 Oct. 1984.