Data Exchange Problem with Helpers

Nebojsa Milosavljevic, Sameer Pawar, Salim El Rouayheb, Michael Gastpar and Kannan Ramchandran

Abstract—**In this paper we construct a deterministic polynomial time algorithm for the problem where a set of users is interested in gaining access to a common file, but where each has only partial knowledge of the file. We further assume the existence of another set of terminals in the system, called helpers, who are not interested in the common file, but who are willing to help the users. Given that the collective information of all the terminals is sufficient to allow recovery of the entire file, the goal is to minimize the (weighted) sum of bits that these terminals need to exchange over a noiseless public channel in order achieve this goal. Based on established connections to the multi-terminal secrecy problem, our algorithm also implies a polynomial-time method for constructing the largest shared secret key in the presence of an eavesdropper. We consider the following side-information settings: (i) side-information in the form of uncoded packets of the file, where the terminals' side-information consists of subsets of the file packets; (ii) side-information in the form of linearly correlated packets, where the terminals have access to linear combinations of the file packets; and (iii) the general setting where the the terminals' side-information has an arbitrary (i.i.d.) correlation structure. We provide a polynomial-time algorithm (in the number of terminals) that finds the optimal rate allocations for these terminals, and then determines an explicit optimal transmission scheme for cases (i) and (ii).**

I. Introduction

In recent years cellular systems have witnessed significant improvements in terms of data rates, and are nearly approaching the theoretical limits in terms of the physical layer spectral efficiency. At the same time, the rapid growth in the popularity of data-enabled mobile devices, such as smart phones and tablets, and the resulting explosion in demand for more throughput are challenging our abilities to deliver data, even with the current highly efficient cellular systems. One of the major bottlenecks in scaling the throughput with the increasing number of mobile devices is the "last mile" wireless link between the base station and the mobile devices – a resource that is shared among many terminals served within the cell. This motivates the study of paradigms where cell phone devices can cooperate among themselves to get the desired data in a peer-to-peer fashion without solely relying on the base station.

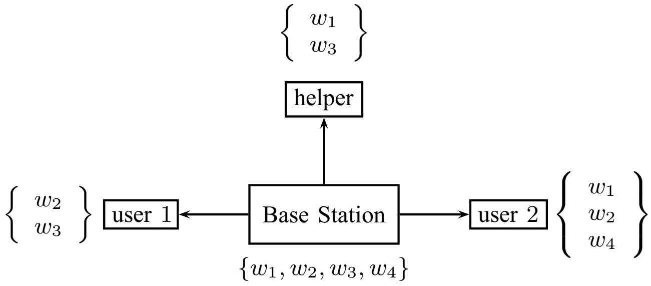

Fig. 1. An example of the data exchange problem with helpers. A base station has a file formed of four packets $w_1, \dots, w_4 \in \mathbb{F}_{q^n}$ and wants to deliver it to two users over an unreliable wireless channel. Additionally, there is a terminal in the system that is in the range of the base station, but he is not interested in the file. However, he is willing to help the two users to obtain the file. The base station stops transmitting once all terminals collectively have all the packets, even if individually they have only subsets of the packets. They can then cooperate among themselves to recover the users' missing packets. If the goal is to minimize the total number of communicated bits, helper transmits packet $w_1 + w_3$, while user 2 transmits packet w_4, where the addition is in the field $\mathbb{F}_{q^n}$.

An example of such a setting is shown in Figure 1, where a base station wants to deliver the same file to multiple geographically-close users over an unreliable wireless downlink. We assume that some terminals, which are in the proximity of the users, are not interested in the file, but are able to overhear some of the base station transmissions. Moreover, we assume that these terminals are willing to help in distributing the file to the interested users. Henceforth, we refer to these terminals as *helpers*, and the terminals interested in the file as *users*. In the example of Figure 1 we assume that the file consists of four equally sized packets w_1, w_2, w_3 and w_4 belonging to some finite field $\mathbb{F}_{q^n}$. Suppose that after a few initial transmission attempts by the base station, the three terminals (including one helper) individually receive only parts of the file (see Figure 1), but collectively have the entire file. Now, if all terminals are in close vicinity and can communicate with each other, then, it is much more desirable and efficient, in terms of resource usage, to reconcile the file among users by letting all terminals "talk" to each other without involving the base station. The cooperation among the terminals has the following advantages:

- Local communication among terminals has a smaller footprint in terms of interference, thus allowing one to use the shared resources (code, time or frequency) freely without penalizing the base station's resources, *i.e.*, higher resource reuse factor.
- Transmissions within the close group of terminals is much more reliable than from the base station to any

N. Milosavljevic, S. Pawar, M. Gastpar and K. Ramchandran are with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail:{nebojsa,spawar, gastpar, kannanr}@eecs.berkeley.edu).

S. El Rouayheb is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: salim@princeton.edu).

M. Gastpar is also with the School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland (e-mail: michael.gastpar@epfl.ch).

This research was funded by the NSF grants (CCF-0964018, CCF-0830788), a DTRA grant (HDTRA1-09-1-0032), and in part by an AFOSR grant (FA9550-09-1-0120).

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

terminal due to geographical proximity of terminals.

- This cooperation allows for the file recovery even when the connection to the base station is either unavailable after the initial phase of transmission, or it is too weak to meet the delay requirement.

The problem of reconciling a file among multiple wireless users having parts of it while minimizing the cost in terms of the total number of bits exchanged is known in the literature as the *data exchange problem* and was introduced by El Rouayheb *et al.* in [1]. In the problem formulation of the data exchange problem it is assumed that all the terminals in the system are interested in recovering the entire file, *i.e.*, there are no helpers. For this problem randomized algorithms were proposed in [2] and [3], while deterministic polynomial time algorithms were proposed in [4], [5]. In [6], [7] and [8] the authors used a combinatorial approach, to solve this problem, which exploited the fact that the constraint set in the corresponding optimization problem forms a submodular base-polyhedron. For the problem considered in this paper, where not all terminals are interested in recovering the file, the constraint set does not constitute a submodular base-polyhedron. As a result, here, we take a convex optimization approach to solve the data exchange problem with helpers.

To exemplify the effect of having helpers, consider the example from Figure 1. Let user 1, user 2 and the helper transmit R_1, R_2 and R_3 bits, respectively. The objective is to minimize the weighted sum-rate $\alpha_1 R_1 + \alpha_2 R_2 + \alpha_3 R_3$ such that, user 1 and user 2 can recover the entire file. It can be shown that for the case when $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the minimum communication cost is 2 and can be achieved by the following coding scheme: user 2 transmits packet w_4, and the helper transmits $w_1 + w_3$, where the addition is over the underlying field $\mathbb{F}_{q^n}$. This corresponds to the optimal rate allocation $R_2 = R_3 = 1$ symbol in $\mathbb{F}_{q^n}$. Now, if there was no helper in the system, it would take a total of 3 transmissions to reconcile the file among the two users. That is user 1 has to transmit w_3 and user 2 transmits w_1 and w_4. Thus, the helpers can contribute to lowering the total communication cost in the system.

The discussion above considers only a simple form of side-information, where different terminals observe partial uncoded "raw" packets of the original file. Content distribution networks are increasingly using coding, such as Fountain codes or linear network codes, to improve the system efficiency [9]. In such scenarios, the side-information representing the partial knowledge gained by the terminals would be coded and in the form of linear combinations of the original file packets, rather than the raw packets themselves. The previous two cases of side-information ("raw" and coded) can be regarded as special cases of the more general problem where the side-information has arbitrary correlation among the data observed by the different terminals and where the goal is to minimize the weighted total communication cost. In [10] Csiszár and Narayan posed a related security problem referred to as the "multi-terminal key agreement" problem. They showed that obtaining the file among the users in minimum number of bits exchanged over the public channel is sufficient to maximize the size of the secret key shared between the users. This result establishes a connection between the Multi-party key agreement and the Data exchange problem with helpers. The authors in [10] solved the key agreement problem by formulating it as a linear program (LP) with an exponential number of rate-constraints, corresponding to all possible cut-sets that need to be satisfied.

In this paper, we make the following contributions. First, we provide a *deterministic polynomial time* algorithm for finding an optimal rate allocation, w.r.t. a linear weighted sum-rate cost needed to deliver the file to all users when all terminals have arbitrarily correlated side-information. Second, for the the data exchange problem with helpers, for raw or linearly coded side-information, we propose an efficient *communication scheme* design based on the algebraic network coding framework [11], [12].

II. System Model and Preliminaries

In this paper, we consider a setup with m terminals out of which some subset of them is interested in gaining access to a file or a random process. Let $X_1, X_2, \ldots, X_m$, $m \geq 2$, denote the components of a discrete memoryless multiple source (DMMS) with a given joint probability mass function. Each user $i \in \mathcal{M} \triangleq \{1, 2, \ldots, m\}$ observes n i.i.d. realizations of the corresponding random variable X_i.

Let $\mathcal{A} = \{1, 2, \ldots, k\} \subseteq \mathcal{M}$ be the subset of terminals, called users, who are interested in gaining access to the file, *i.e.*, learning the joint process $X_{\mathcal{M}} \triangleq (X_1, \ldots, X_m)$. The remaining terminals $\{k+1, \ldots, m\}$ serve as helpers, *i.e.*, they are not interested in recovering the file, but they are willing to help users in the set $\mathcal{A}$ to obtain it. In [10], Csiszár and Narayan showed that to deliver the file to all users in a setup with general DMMS, *interactive communication is not needed.* As a result, in the sequel WLOG we can assume that the transmission of each user is only a function of its own initial observations. Let $F_i \triangleq f_i(X_i^n)$ represent the transmission of the user $i \in \mathcal{M}$, where $f_i(\cdot)$ is any desired mapping of the observations X_i^n. For each user in $\mathcal{A}$ in order to recover the entire file, transmissions F_i, $i \in \mathcal{M}$, should satisfy,

$$\lim_{n \to \infty} \frac{1}{n} H(X_{\mathcal{M}}^n | \mathbf{F}, X_l^n) = 0, \quad \forall l \in \mathcal{A}, \tag{1}$$

where $\mathbf{F} \triangleq (F_1, F_2, \ldots, F_m)$.

Definition 1. A rate vector $\mathbf{R} = (R_1, R_2, \ldots, R_m)$ is an *achievable data exchange (DE)-rate vector* if there exists a communication scheme with transmitted messages $\mathbf{F} = (F_1, F_2, \ldots, F_m)$ that satisfies (1), and is such that

$$R_i = \lim_{n \to \infty} \frac{1}{n} H(F_i), \quad \forall i \in \mathcal{M}. \tag{2}$$

It is easy to show using cut-set bounds that all the achievable *DE*-rate vectors necessarily belong to the following region

$$\mathcal{R} \triangleq \{\mathbf{R} : R(\mathcal{S}) \geq H(X_{\mathcal{S}} | X_{\mathcal{S}^c}),\ \forall \mathcal{S} \subset \mathcal{M},\ \mathcal{A} \nsubseteq \mathcal{S}\}, \tag{3}$$

where $R(\mathcal{S}) \triangleq \sum_{i \in \mathcal{S}} R_i$. Also, using a random coding argument, it can be shown that the rate region $\mathcal{R}$ is an achievable rate region [10].

In this work, we aim to design a polynomial complexity algorithm that delivers the file to all users in $\mathcal{A}$ while simultaneously minimizing a linear communication cost function $\sum_{i=1}^{m}\alpha_i R_i$, where $\underline{\alpha} \triangleq (\alpha_1,\cdots,\alpha_m), 0 \leq \alpha_i < \infty$, is an $m-$dimensional vector of non-negative finite weights. We allow α_i's to be arbitrary non-negative constants, to account for the case when communication of some terminals is more expensive compared to the others, *e.g.*, setting α_1 to be a large value compared to the other weights minimizes the rate allocated to the user 1. This goal can be formulated as the following linear program:

$$\begin{aligned}&\min_{\mathbf{R}} \sum_{i=1}^{m} \alpha_i R_i, \\ &\text{s.t. } \; R(\mathcal{S}) \geq H(X_{\mathcal{S}}|X_{\mathcal{S}^c}), \; \forall \mathcal{S} \subset \mathcal{M}, \; \mathcal{A} \not\subseteq \mathcal{S}.\end{aligned} \tag{4}$$

A. Finite Linear Source Model

In this section we introduce a source-model to characterize *linearly correlated* side-information. The source model we use is called *Finite linear source* as defined in [13]. It captures the fact that the content distribution networks using linear codes result in terminals' side-information to be in the form of linear combinations of the original packets instead of the uncoded packets, as is the case in conventional "Data Exchange problem".

Next, we briefly describe the finite linear source model. Let q be some power of a prime. Consider the N-dimensional random vector $\mathbf{W} \in \mathbb{F}_{q^n}^N$ whose components are independent and uniformly distributed over the elements of $\mathbb{F}_{q^n}$. Then, in the linear source model, the observation of i^{th} user is simply given by

$$\mathbf{X}_i = \mathbf{A}_i \mathbf{W}, \; i \in \mathcal{M}, \tag{5}$$

where $\mathbf{A}_i \in \mathbb{F}_q^{\ell_i \times N}$ is an observation matrix for the user i.

It is easy to verify that for the finite linear source model,

$$\frac{H(X_i)}{\log q^n} = \text{rank}(\mathbf{A}_i). \tag{6}$$

Henceforth for the finite linear source model we will use the entropy of the observations and the rank of the observation matrix interchangeably.

III. Deterministic Algorithm

We begin this section by exploring the case when the set $\mathcal{A}$ consists of only one user. Then, by using the methodology of [14], we extend our solution to the case when the set $\mathcal{A}$ has arbitrary number of users.

A. Deterministic Algorithm when $|\mathcal{A}| = 1$

Let there be only one user interested in obtaining the file, *i.e.*, $\mathcal{A} = \{1\}$. This is known as a multi-terminal Slepian-Wolf problem [15] for which the achievable rate region has the following form:

$$\mathcal{R}_1 = \{\mathbf{R} : R(\mathcal{S}) \geq H(X_{\mathcal{S}}|X_{\mathcal{S}^c}, X_1), \; \forall \mathcal{S} \subseteq \mathcal{M} \setminus \{1\}\}.$$

Hence, the underlying optimization problem has the following form

$$\min_{\mathbf{R}} \sum_{i \in \mathcal{M}\setminus\{1\}} \alpha_i R_i, \quad \text{s.t. } \mathbf{R} \in \mathcal{R}_1. \tag{7}$$

Optimization problem (7) can be solved analytically due to the fact that the set function

$$f(\mathcal{S}) = H(X_{\mathcal{S}}|X_{\mathcal{S}^c}, X_1), \quad \forall \mathcal{S} \subseteq \mathcal{M} \setminus \{1\} \tag{8}$$

is supermodular (see [16] for the formal definition). Therefore, optimization problem (7) is over a supermodular polyhedron $\mathcal{R}_1$. From the combinatorial optimization theory it is known that Edmonds' greedy algorithm [17] provides an analytical solution to this problem (see Algorithm 1).

Algorithm 1 Edmonds' algorithm applied to our problem

```
Set $j_1, j_2, \ldots, j_{m-1}$ to be an ordering of $\mathcal{M} \setminus \{1\}$ such
  that $\alpha_{j_1} \leq \alpha_{j_2} \leq \cdots \leq \alpha_{j_{m-1}}$.
for $i = 1$ to $m - 1$ do
    $R^*_{j_i} = H(X_{j_i}|X_1, X_{j_1}, X_{j_2}, \ldots, X_{j_{i-1}})$.
end for
```

Example 1. Consider a system with $m = 6$ terminals $\mathcal{M} = \{1,2,3,4,5,6\}$. For convenience, we express the underlying data vector as $\mathbf{W} = \begin{bmatrix} a & b & c \end{bmatrix}^T \in \mathbb{F}_{q^n}^3$, where a, b, c are independent uniform random variables in $\mathbb{F}_{q^n}$. Let us consider the case where each node has the following observations: $\mathbf{X}_1 = \{a+b\}$, $\mathbf{X}_2 = \{a+c\}$, $\mathbf{X}_3 = \{b+c\}$, $\mathbf{X}_4 = \{a\}$, $\mathbf{X}_5 = \{b\}, \mathbf{X}_6 = \{c\}$. Let us assume that user 1 is interested in recovering the vector $\mathbf{W}$ such that underlying communication cost is $\sum_{i=2}^{6} R_i$.

It immediately follows from Algorithm 1 that a solution to this problem is $R^*_4 = R^*_6 = 1$, and $R^*_2 = R^*_3 = R^*_5 = 0$. In other words, user 1 is missing 2 linear equations in order to be able to decode all 3 data packets.

B. Deterministic Algorithm when $|\mathcal{A}| > 1$

When $|\mathcal{A}| > 1$, the set of constraints of the optimization problem (4) does not constitute a polyhedron of a supermodular function, as it was the case when $|\mathcal{A}| = 1$. The reason is that the possible set function is not defined for all subsets of $\mathcal{M}$. For that reason, in this section, we take a convex optimization approach to solve the problem (4), where we use the single user solution as a key building block.

First, we observe that an achievable *DE*-rate vector $\mathbf{Z}$ has to simultaneously belong to the rate regions $\mathcal{R}_l$ of each individual user $l \in \mathcal{A}$, where

$$\mathcal{R}_l = \{\mathbf{R} : R(\mathcal{S}) \geq H(X_{\mathcal{S}}|X_{\mathcal{S}^c}, X_l), \; \forall \mathcal{S} \subseteq \mathcal{M} \setminus \{l\}\}. \tag{9}$$

Thus, the rate region $\mathcal{R}$ defined in (3) can be equivalently represented as

$$\begin{aligned}\mathcal{R} = \{\mathbf{Z} : Z_i \geq R_i^{(l)}, \quad \forall i \in \mathcal{M} \setminus \{l\}, \\ \text{s.t. } \mathbf{R}^{(l)} \in \mathcal{R}_l, \; \forall l \in \mathcal{A}\}.\end{aligned} \tag{10}$$

Therefore, the optimal *DE*-rate vector $\mathbf{Z}^*$ can be obtained as follows

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{R}} \sum_{i=1}^{m} \alpha_i Z_i, \qquad (11)\\ &\text{s.t. } Z_i \geq R_i^{(l)}, \quad \forall l \in \mathcal{A}, \ \forall i \in \mathcal{M} \setminus \{l\},\\ &\quad \mathbf{R}^{(l)} \in \mathcal{R}_l, \quad \forall l \in \mathcal{A}. \end{aligned}$$

Optimization problem (11) has an exponential number of constraints, which makes it challenging to solve in polynomial time. To obtain a polynomial time solution we consider the Lagrangian dual of problem (11).

$$\begin{aligned} &\max_{\mathbf{\Lambda}} \sum_{l=1}^{k} g^{(l)}(\mathbf{\Lambda}^{(l)}), \qquad (12)\\ &\text{s.t. } \sum_{l=1}^{k} \lambda_i^{(l)} = \alpha_i, \ \lambda_i^{(l)} \geq 0, \quad \forall l \in \mathcal{A}, \ \forall i \in \mathcal{M} \setminus \{l\}, \end{aligned}$$

where

$$g^{(l)}(\mathbf{\Lambda}^{(l)}) = \min_{\mathbf{R}^{(l)}} \sum_{i \in \mathcal{M} \setminus \{l\}} \lambda_i^{(l)} R_i^{(l)}, \quad \text{s.t. } \mathbf{R}^{(l)} \in \mathcal{R}_l. \qquad (13)$$

Dual variable $\mathbf{\Lambda}$ in the above problem is represented in matrix form as follows.

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1^{(1)} & \lambda_2^{(1)} & \cdots & \lambda_m^{(1)} \\ \lambda_1^{(2)} & \lambda_2^{(2)} & \cdots & \lambda_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{(k)} & \lambda_2^{(k)} & \cdots & \lambda_m^{(k)} \end{bmatrix}. \qquad (14)$$

We denote by $\mathbf{\Lambda}_i$ and $\mathbf{\Lambda}^{(l)}$, the i^{th} column vector and l^{th} row vector of the matrix $\mathbf{\Lambda}$, respectively. Moreover, we denote by

$$\tilde{\mathbf{R}} = \begin{bmatrix} \tilde{R}_1^{(1)} & \tilde{R}_2^{(1)} & \cdots & \tilde{R}_m^{(1)} \\ \tilde{R}_1^{(2)} & \tilde{R}_2^{(2)} & \cdots & \tilde{R}_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{R}_1^{(k)} & \tilde{R}_2^{(k)} & \cdots & \tilde{R}_m^{(k)} \end{bmatrix} \qquad (15)$$

the rate matrix whose l^{th} row, here denoted by $\tilde{\mathbf{R}}^{(l)}$, represents an optimizer of the problem (13) w.r.t. the weight vector $\mathbf{\Lambda}^{(l)}$. In order to ensure consistency with the optimization problem (12) observe that $\lambda_l^{(l)} = 0$, and $\tilde{R}_l^{(l)} = 0$, $\forall l = 1, \ldots, k$.

For any given user $l \in \mathcal{A}$, the objective function (13) of the dual problem (12) can be computed analytically using Algorithm 1. The optimization problem (12) is a linear program (LP) with $\mathcal{O}(m \cdot k)$ number of constraints, which makes it possible to solve it in polynomial time (w.r.t. number of terminals). To solve the optimization problem (12) we apply a subgradient method, as described below.

Starting with a feasible iterate $\mathbf{\Lambda}[0]$ w.r.t. the optimization problem (12), every subsequent iterate $\mathbf{\Lambda}[\eta]$ can be recursively represented as an Euclidian projection of the vector

$$\mathbf{\Lambda}_i[\eta] = \mathbf{\Lambda}_i[\eta - 1] + \theta[\eta - 1] \cdot \tilde{\mathbf{R}}_i[\eta - 1], \quad \forall i \in \mathcal{M} \qquad (16)$$

onto the hyperplane $\left\{\mathbf{\Lambda}_i \geq \mathbf{0} | \sum_{l=1}^{k} \lambda_i^{(l)} = \alpha_i\right\}$, where $\tilde{\mathbf{R}}_i[\eta - 1]$ is the i^{th} column of the rate matrix $\tilde{\mathbf{R}}[\eta - 1]$. The Euclidian projection ensures that every iterate $\mathbf{\Lambda}[\eta]$ is feasible w.r.t. the optimization problem (12). It is not hard to verify that the following initial choice of $\mathbf{\Lambda}[0]$ is feasible.

$$\lambda_i^{(l)}[0] = \begin{cases} \frac{\alpha_i}{k} & \text{if } i \notin \mathcal{A} \\ \frac{\alpha_i}{k-1} & \text{if } i \in \mathcal{A} \setminus \{l\} \\ 0 & \text{if } i = l \end{cases}, \quad \forall i \in \mathcal{M}, \ \forall l \in \mathcal{A}. \qquad (17)$$

By appropriately choosing the step size $\theta[\eta]$ in each iteration (16), it is guaranteed that the subgradient method described above converges to the optimal solution of the dual problem (12). To recover the primal optimal solution from the iterates $\mathbf{\Lambda}[\eta]$ we use results from [18], where at each iteration of (16), the primal iterate is constructed as follows.

$$\hat{\mathbf{R}}[\eta] = \sum_{j=1}^{\eta} \mu_j^{(\eta)} \tilde{\mathbf{R}}[j], \qquad (18)$$

where

$$\sum_{j=1}^{n} \mu_j^{(\eta)} = 1, \ \mu_j^{(\eta)} \geq 0, \ \text{for } j = 1, 2, \ldots, \eta. \qquad (19)$$

By carefully choosing the step size $\theta[\eta]$, $\forall \eta$ in (16) and the convex combination coefficients $\mu_j^{(\eta)}$, $\forall j = 1, \ldots, \eta$, $\forall \eta$, it is guaranteed that (18) converges to the minimizer of (11), and therefore to the minimizer of the original problem (4). In [18], the authors proposed several choices for $\{\theta[\eta]\}$ and $\{\mu_j^{(\eta)}\}$ which lead to the primal recovery. Here we list some of them.

1) $\theta[\eta] = \frac{a}{b + c\eta}$, $\forall \eta$, where $a > 0$, $b \geq 0$, $c > 0$, $\mu_j^{(\eta)} = \frac{1}{\eta}$, $\forall j = 1, \ldots, \eta$, $\forall \eta$,
2) $\theta[\eta] = \eta^{-a}$, $\forall \eta$, where $0 < a < 1$, $\mu_j^{(\eta)} = \frac{1}{\eta}$, $\forall j = 1, \ldots, \eta$, $\forall \eta$.

Now, it is only left to compute an optimal rate allocation w.r.t to the problem defined in (4). Let $\mathbf{R}^*$ and $\mathbf{Z}^*$ be the optimal rate vectors of the problems (4) and (11), respectively. As we pointed out earlier $\mathbf{R}^* = \mathbf{Z}^*$, where $\mathbf{Z}^*$ can be computed from the matrix $\hat{\mathbf{R}}[\eta]$ for a sufficiently large η, as follows

$$Z_i^* = \max\left\{\hat{R}_i^{(1)}[\eta], \hat{R}_i^{(2)}[\eta], \ldots, \hat{R}_i^{(k)}[\eta]\right\}, \quad \forall i \in \mathcal{M}. \qquad (20)$$

C. Code Construction for the Linear Source Model

In this Section we briefly address the question of the optimal code construction for the linear source model. For that matter, let us consider the following example.

Example 2. Let us consider the same source model as in Example 1, where $\mathcal{A} = \{1, 2, 3\}$, and the objective function is $\sum_{i=1}^{6} R_i$. Applying the algorithm described above, we obtain

$$R_1^* = R_2^* = R_3^* = \frac{1}{4}, \quad R_4^* = R_5^* = R_6^* = \frac{1}{2}. \qquad (21)$$

This solution suggests that in order to design a scheme that performs optimally, it is necessary to split all the packets into 4

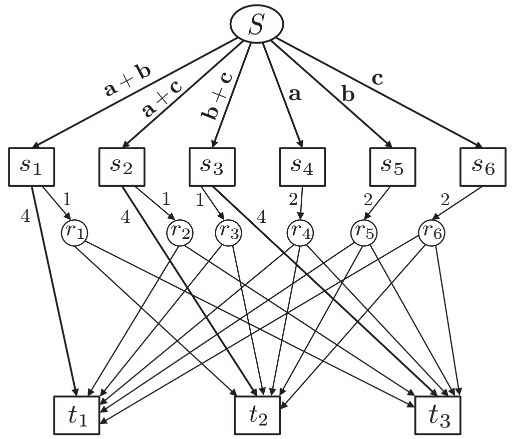

Fig. 2. Multicast network constructed for Example 2 with the optimal rate tuple $R_1^* = R_2^* = R_3^* = 2$, $R_4^* = R_5^* = R_6^* = 1$. Each user receives side-information from "itself" through links (s_i, r_i), $i = 1, 2, 3$, and from the other terminals through links (r_i, t_j), $i = 1, \ldots, 6$, $j = 1, 2, 3$, $i \neq j$.

equally sized chunks. In other words, terminals' observations can be written as $\mathbf{X}_1 = \mathbf{a} + \mathbf{b} = \{a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4\}$, $\mathbf{X}_2 = \mathbf{a} + \mathbf{c} = \{a_1 + c_1, a_2 + c_2, a_3 + c_3, a_4 + c_4\}$, etc., where all a_i's, b_i's and c_i's belong to $\mathbb{F}_{q^{n/4}}$. For this "extended" source model we have that the optimal rate allocation is $R_1^* = R_2^* = R_3^* = 1$, $R_4^* = R_5^* = R_6^* = 2$.

Next question we need to address is how to design transmissions of each user? Starting from an optimal (integer) rate allocation, we first construct the corresponding multicast network (see Figure 2). In this construction, notice that there are several types of nodes. First, there is a super node S that possesses all the packets. Each user in the set $\mathcal{A}$ plays the role of a transmitter and a receiver, while the helpers act only as transmitters. To model this, we denote $s_1, \ldots, s_6$ to be the "sending" nodes, and t_1, t_2 and t_3 to be the receiving nodes. To model the side-information at users 1, 2 and 3, we introduce links (s_i, t_i), $i = 1, 2, 3$, of capacity 4, which are routing the users' observations to the corresponding receiving nodes. To model the broadcast nature of each transmission, we introduce "dummy" nodes $r_1, \ldots, r_6$, such that the capacity of the links (s_i, r_i) is the same as link capacity (r_i, t_j), $i \neq j$, and is equal to R_i^*, $\forall i \in \mathcal{M}$.

To solve for actual transmissions of each terminal, we apply the algebraic network coding approach [11], with appropriately designed source matrix $\mathbf{A}$ which corresponds to the side-information of all terminals. Finally, the network code for the data exchange problem with helpers can be constructed in polynomial time from the algorithms provided in [12] which are based on a simultaneous transfer matrix completion.

IV. Conclusion and Extensions

In this paper, we studied the data exchange problem with helpers. We provided a deterministic polynomial time algorithm for minimizing the weighted sum-rate cost of communication. We showed that the data exchange problem with only one user and many helpers can be solved analytically using Edmonds' algorithm. Further, using the single user solution as a building block we showed how one can solve the more general problem with arbitrary number of users. Several extensions are of interest. For instance, we can consider a modification of the original data exchange problem where only helpers are allowed to transmit. Starting from a single user case, it is easy to see that an achievable rate vector must satisfy all the cut-set constraints over the helper set such that the user is always on the receiving side of the cut. Minimizing the weighted sum-rate cost over all achievable rate tuples can again be done using Edmonds' algorithm (see Algorithm 1). Finally, extension to the multiple user case corresponds to the weighted sum-rate minimization over all rate tuples that are simultaneously achievable for all the users. This optimization problem can be solved in polynomial time using the same approach as described in Section III-B.

References

[1] S. El Rouayheb, A. Sprintson, and P. Sadeghi, "On coding for cooperative data exchange," in *Proceedings of ITW*, 2010.

[2] A. Sprintson, P. Sadeghi, G. Booker, and S. El Rouayheb, "A randomized algorithm and performance bounds for coded cooperative data exchange," in *Proceedings of ISIT*, 2010, pp. 1888–1892.

[3] D. Ozgul and A. Sprintson, "An algorithm for cooperative data exchange with cost criterion," in *Proceedings of ITA*, 2011, pp. 1–4.

[4] T. Courtade, B. Xie, and R. Wesel, "Optimal Exchange of Packets for Universal Recovery in Broadcast Networks," in *Proceedings of Military Communications Conference*, 2010.

[5] S. Tajbakhsh, P. Sadeghi, and R. Shams, "A generalized model for cost and fairness analysis in coded cooperative data exchange," in *Proceedings of NetCod*, 2011, pp. 1–6.

[6] T. Courtade and R. Wesel, "Weighted universal recovery, practical secrecy, and an efficient algorithm for solving both," in *Proceedings of Allerton Conference on Communication, Control, and Computing*, sept. 2011, pp. 1349 –1357.

[7] N. Milosavljevic, S. Pawar, S. El Rouayheb, M. Gastpar, and K. Ramchandran, "Deterministic algorithm for the cooperative data exchange problem," in *Proceedings of ISIT*, 2011, pp. 410–414.

[8] N. Milosavljevic, S. Pawar, S. Rouayheb, M. Gastpar, and K. Ramchandran, "Optimal deterministic polynomial-time data exchange for omniscience," *Arxiv preprint arXiv:1108.6046*, 2011.

[9] M. Luby, "LT codes," in *Proceedings of Foundations of Computer Science*. IEEE, 2002, pp. 271–280.

[10] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3047–3061, 2004.

[11] R. Koetter and M. Medard, "An Algebraic Approach to Network Coding," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 782 – 795, 2003.

[12] N. Harvey, D. Karger, and K. Murota, "Deterministic network coding by matrix completion," in *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 2005, pp. 489–498.

[13] C. Chan, "Generating Secret in a Network," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[14] D. Lun, N. Ratnakar, M. Médard, R. Koetter, D. Karger, T. Ho, E. Ahmed, and F. Zhao, "Minimum-cost multicast over coded packet networks," *Information Theory, IEEE Transactions on*, vol. 52, no. 6, pp. 2608–2623, 2006.

[15] T. Cover and J. Thomas, "Elements of information theory 2nd edition," 2006.

[16] S. Fujishige, *Submodular functions and optimization*. Elsevier Science, 2005.

[17] J. Edmonds, "Submodular functions, matroids, and certain polyhedra," *Combinatorial structures and their applications*, pp. 69–87, 1970.

[18] H. Sherali and G. Choi, "Recovery of primal solutions when using subgradient optimization methods to solve lagrangian duals of linear programs," *Operations Research Letters*, vol. 19, no. 3, pp. 105–113, 1996.