Simple algorithms and guarantees for low rank matrix completion over $\mathbb{F}_2$

James Saunderson
Dept. of Electrical Engineering
University of Washington
Email: jamesfs@uw.edu

Maryam Fazel
Dept. of Electrical Engineering
University of Washington
Email: mfazel@uw.edu

Babak Hassibi
Dept. of Electrical Engineering
California Institute of Technology
Email: hassibi@caltech.edu

***Abstract*—Let $X^\star$ be a $n_1 \times n_2$ matrix with entries in $\mathbb{F}_2$ and rank $r < \min(n_1, n_2)$ (often $r \ll \min(n_1, n_2)$). We consider the problem of reconstructing $X^\star$ given only a subset of its entries. This problem has recently found numerous applications, most notably in network and index coding, where finding optimal linear codes (over some field $\mathbb{F}_q$) can be reduced to finding the minimum rank completion of a matrix with a subset of revealed entries. The problem of matrix completion over reals also has many applications and in recent years several polynomial-time algorithms with provable recovery guarantees have been developed. However, to date, such algorithms do not exist in the finite-field case. We propose a linear algebraic algorithm, based on inferring low-weight relations among the rows and columns of $X^\star$, to attempt to complete $X^\star$ given a random subset of its entries. We establish conditions on the row and column spaces of $X^\star$ under which the algorithm runs in polynomial time (in the size of $X^\star$) and can successfully complete $X^\star$ with high probability from a vanishing fraction of its entries. We then propose a linear programming-based extension of our basic algorithm, and evaluate it empirically.**

I. Introduction

Let $X^\star$ be a $n_1 \times n_2$ matrix with entries in the binary field $\mathbb{F}_2$ and rank r (often $r \ll \min(n_1, n_2)$). The problem of reconstructing $X^\star$ given only a subset of its entries is referred to as *low rank matrix completion*. When the matrix under consideration is real (or complex), the problem has an extensive literature, with many applications and several polynomial time (convex and non-convex) algorithms with provable recovery guarantees (see, e.g., [1] and the references therein). In the finite-field case the problem has recently found important applications in network and index coding, where finding optimal linear codes can be reduced to finding the minimum rank completion of a matrix with a subset of revealed entries [2], [3], [4], [5]. There are also applications involving decoding rank codes in the presence of erasures [6].

Despite these applications, there is very little in the way of efficient algorithms that guarantee provable recovery in the finite field case. [7] studies complexity issues and shows that the problem is generally NP hard. [8] studies a related problem where, instead of entries, random linear combinations of the entries are observed and gives various information-theoretic bounds on the number of measurements necessary for low rank matrix recovery. [9] uses ideas from graph coloring to obtain efficient algorithms and [10] works with matrices from finite fields but with multiplication over integers. Our work is partially inspired by the work of Feldman [11] who was the first to relax the problem of decoding binary LDPC codes to a linear program over the reals.

We propose a linear algebraic algorithm, based on inferring low-weight relations among the rows and columns of $X^\star$, to attempt to complete $X^\star$ given a *random* subset of its entries. (We focus on a random model of revealed entries primarily for simplicity.) We establish conditions on the row and column spaces of $X^\star$ under which the algorithm runs in polynomial time (in the size of $X^\star$) and can successfully complete $X^\star$ with high probability (in the choice of the random subset) from a vanishing fraction of its entries. This is done in Section IV. Motivated by this, in Section V, we propose a practical, yet more effective (because it can deal *globally and simultaneously* with column and row relations) linear programming-based extension of our basic algorithm. While we currently do not have a complete analysis of this linear program (LP), we do evaluate it empirically in Section VI. Note that the number of $n \times n$ rank r matrices over $\mathbb{F}_2$ is given by $N_{n,r} = \prod_{k=0}^{r-1} \frac{(2^n - 2^k)^2}{(2^r - 2^k)}$, and that therefore identifying any such matrix requires $\log_2 N_{n,r}$ bits. Viewing each revealed entry of $X^\star$ as a bit implies that we need to observe at least $\log_2 N_{n,r}$ entries of $X^\star$ to recover it. Our simulations suggest the LP can complete $n \times n$ random rank r matrices (for $n \leq 100$ and $r \leq 10$) given at most $3 \log_2 N_{n,r}$ random entries.

II. Notation and preliminaries

Let $[n] = \{1, 2, \ldots, n\}$ and let $2^{[n]}$ denote the collection of all subsets of $[n]$. If $p \in [0, 1]$ and T is a finite set we write $S \sim \mathcal{B}(T, p)$ if S is the random subset of T obtained by choosing each element of T independently with probability p.

Let $\mathbb{F}_2^{n_1 \times n_2}$ denote the space of $n_1 \times n_2$ matrices over $\mathbb{F}_2$. The *rank* of $X \in \mathbb{F}_2^{n_1 \times n_2}$ is the smallest positive integer r such that $X = U_1 U_2^T$ for $U_1 \in \mathbb{F}_2^{n_1 \times r}$ and $U_2 \in \mathbb{F}_2^{n_2 \times r}$. If $x \in \mathbb{F}_2^n$ the *Hamming weight*, denoted $\mathrm{wt}(x)$, is the number of non-zero entries of x. If $S \subseteq [n]$ let $e_S \in \mathbb{F}_2^n$ be the vector supported on S, i.e. $[e_S]_i = 1$ if and only if $i \in S$.

If $\mathcal{C} \subseteq \mathbb{F}_2^n$ is a subspace let $\mathcal{C}^\perp = \{x \in \mathbb{F}_2^n : x^T y = 0, \ \forall y \in \mathcal{C}\}$ be its *dual subspace*; let $d(\mathcal{C}) = \min_{x \in \mathcal{C} \setminus \{0\}} \mathrm{wt}(x)$ be the

This is based on work supported by the NSF under grant CCF-1409836.

minimum distance of $\mathcal{C}$; let $[\mathcal{C}]_{\leq s} = \text{span}\{x \in \mathcal{C} : \text{wt}(x) \leq s\}$ be the span of the elements of $\mathcal{C}$ with weight at most s.

If $\mathcal{C}_1 \subseteq \mathbb{F}_2^{n_1}$ and $\mathcal{C}_2 \subseteq \mathbb{F}_2^{n_2}$ are subspaces let $\mathcal{C}_1 \otimes \mathcal{C}_2 \subseteq \mathbb{F}_2^{n_1 \times n_2}$ be the subspace of matrices with column space contained in $\mathcal{C}_1$ and row space contained in $\mathcal{C}_2$.

If $S = \{i_1, \ldots, i_{|S|}\} \subset [n]$ let $P_S : \mathbb{F}_2^n \to \mathbb{F}_2^{|S|}$ be the coordinate projection defined by $[P_S(x)]_\ell = x_{i_\ell}$ for $\ell = 1, 2, \ldots, |S|$. When $\Omega \subseteq [n_1] \times [n_2]$ this definition naturally extends to $P_\Omega : \mathbb{F}_2^{n_1 \times n_2} \to \mathbb{F}_2^{|\Omega|}$. Finally, we record the following simple result because we use it repeatedly.

Lemma 1. *Let $\mathcal{C} \subseteq \mathbb{F}_2^n$ be a subspace. If $S \sim \mathcal{B}([n], p)$ then*

$$\Pr[P_S(x) = 0 \text{ for some } x \in \mathcal{C} \setminus \{0\}] \leq 2^{\dim(\mathcal{C})} e^{-pd(\mathcal{C})}.$$

Proof. For fixed $x \in \mathcal{C} \setminus \{0\}$, $\Pr[P_S(x) = 0] = (1-p)^{\text{wt}(x)}$. Let $r = \dim(\mathcal{C})$. By taking a union bound we see that

$$\begin{aligned}\Pr[P_S(x) = 0 \text{ for some } x \in \mathcal{C} \setminus \{0\}] \\ \leq \sum_{x \in \mathcal{C} \setminus \{0\}} (1-p)^{\text{wt}(x)} \leq (2^r - 1)(1-p)^{d(\mathcal{C})} \leq 2^r e^{-pd(\mathcal{C})}\end{aligned}$$

where we have used the inequalities $d(\mathcal{C}) \leq \text{wt}(x)$ for all $x \in \mathcal{C} \setminus \{0\}$ and $(1-p)^{d(\mathcal{C})} \leq e^{-pd(\mathcal{C})}$ (which follows from $\log(1-p) \leq -p$ for $0 \leq p < 1$). □

III. Basic strategy

Our basic approach to devising algorithms for low rank matrix completion over $\mathbb{F}_2$ is to infer, from the partial information $P_\Omega(X^\star)$, linear relations among the rows and columns of $X^\star$. More formally we aim to find collections $\mathcal{H}_1 \subseteq 2^{[n_1]}$ and $\mathcal{H}_2 \subseteq 2^{[n_2]}$ of sets (or *parity checks*) such that

$$e_{S_1}^T X^\star = 0 \ \ \forall S_1 \in \mathcal{H}_1, \ \text{ and } \ X^\star e_{S_2} = 0 \ \ \forall S_2 \in \mathcal{H}_2.$$

A. Meta-algorithm for low rank completion

The algorithms we propose for matrix completion are all of the following form for different choices of $\mathcal{H}_1$ and $\mathcal{H}_2$.

1) Construct $\mathcal{H}_1 \subseteq 2^{[n_1]}$ and $\mathcal{H}_2 \subseteq 2^{[n_2]}$.
2) For $i = 1, 2$ construct $U_i \in \mathbb{F}_2^{n_i \times k_i}$ with columns that are a basis for $\text{span}\{e_{S_i} : S_i \in \mathcal{H}_i\}^\perp$.
3) Return $U_1 \tilde{X} U_2^T$ for all $\tilde{X} \in \mathbb{F}_2^{k_1 \times k_2}$ satisfying

$$P_\Omega(U_1 \tilde{X} U_2^T) = P_\Omega(X^\star). \tag{1}$$

The following definition is central to our subsequent discussion. It allows us to refer succinctly to the situation in which $\mathcal{H}_1$ and $\mathcal{H}_2$ are such that the meta-algorithm outputs $X^\star$ as the unique completion given $P_\Omega(X^\star)$ and Ω.

Definition 1. *Let $\mathcal{H}_i \subseteq 2^{[n_i]}$ (for $i = 1, 2$) and let $P_\Omega(X^\star) \in \mathbb{F}_2^{|\Omega|}$. We say that $\mathcal{H}_1$ and $\mathcal{H}_2$* are consistent with $P_\Omega(X^\star)$ *if*

$$\begin{aligned}\big\{X \in \mathbb{F}_2^{n_1 \times n_2} : \ P_\Omega(X) = P_\Omega(X^\star), \\ e_{S_1}^T X = 0 \ \forall S_1 \in \mathcal{H}_1, \ X e_{S_2} = 0 \ \forall S_2 \in \mathcal{H}_2\big\} \neq \emptyset\end{aligned} \tag{2}$$

If, in addition, (2) *consists of a single point we say that $\mathcal{H}_1$* and $\mathcal{H}_2$ are uniquely consistent with $P_\Omega(X^\star)$.

Note that (2) is exactly the set of solutions to (1), expressed differently. Clearly, if $\mathcal{H}_1$ and $\mathcal{H}_2$ are uniquely consistent with $P_\Omega(X^\star)$ then $X^\star$ is the unique output of the meta-algorithm.

B. Complexity

The complexity of the meta-algorithm depends on

1) constructing $\mathcal{H}_1$ and $\mathcal{H}_2$ (which depends on the subsets);
2) computing U_1 and U_2 (at worst $O(|\mathcal{H}_i|^2 n_i)$, by performing row operations on an $|\mathcal{H}_i| \times n_i$ matrix and reading off a basis for the dual space); and
3) finding solutions to (1) (at worst $O(k_1^2 k_2^2 |\Omega|)$ again by performing row operations on a $k_1 k_2 \times |\Omega|$ matrix).

Since $|\Omega| \leq n_1 n_2$ and $k_1 \leq n_1$ and $k_2 \leq n_2$, the total complexity is polynomial in n_1 and n_2 provided $\mathcal{H}_1$ and $\mathcal{H}_2$ can be constructed in polynomial time.

IV. Linear-algebraic algorithms with guarantees

In this section we describe a simple way to choose collections of parity checks $\mathcal{H}_1 \subseteq 2^{[n_1]}$ and $\mathcal{H}_2 \subset 2^{[n_2]}$ for which we can describe conditions on $X^\star$ and Ω such that the meta-algorithm of Section III-A successfully completes $X^\star$ from $P_\Omega(X^\star)$. Importantly we can construct $\mathcal{H}_1$ and $\mathcal{H}_2$ via simple algorithms that run in time polynomial in n_1 and n_2, giving algorithms and guarantees for matrix completion over $\mathbb{F}_2$.

The following describes those subsets S_1 of rows (resp. columns) that can be completed so that they sum to zero.

Definition 2. *Let $S_1 \subseteq [n_1]$ and $S_2 \subseteq [n_2]$ and let $P_\Omega(X^\star) \in \mathbb{F}_2^{|\Omega|}$. We say that S_1* is apparently consistent with $P_\Omega(X^\star)$ *if*

$$\{X \in \mathbb{F}_2^{n_1 \times n_2} : \ P_\Omega(X) = P_\Omega(X^\star), \ \ e_{S_1}^T X = 0\} \neq \emptyset$$

and that S_2 is apparently consistent with $P_\Omega(X^\star)$ *if*

$$\{X \in \mathbb{F}_2^{n_1 \times n_2} : \ P_\Omega(X) = P_\Omega(X^\star), \ \ X e_{S_2} = 0\} \neq \emptyset.$$

A. Choice of $\mathcal{H}_1$ and $\mathcal{H}_2$

We now define a collection of checks $\mathcal{H}_1$ and $\mathcal{H}_2$ for use in the meta-algorithm of Section III-A. Fix $P_\Omega(X^\star) \in \mathbb{F}_2^{|\Omega|}$ and choose positive integers s_1 and s_2. Define, for $i = 1, 2$,

$$\begin{aligned}\mathcal{H}_{i,s_i} := \{S_i \subseteq [n_i] : |S_i| \leq s_i, \\ S_i \text{ apparently consistent with } P_\Omega(X^\star)\}.\end{aligned} \tag{3}$$

We can construct $\mathcal{H}_{1,s_1}$ via Algorithm 1. An obvious analogue of Algorithm 1 allows us to construct $\mathcal{H}_{2,s_2}$ in a similar way. The following result summarizes the size of $\mathcal{H}_{i,s_i}$, and complexity of constructing these sets, for $i = 1, 2$.

Algorithm 1 Constructing $\mathcal{H}_{1,s_1}$

Input: $\Omega \subseteq [n_1] \times [n_2]$, $P_\Omega(X^\star)$, positive integer s_1

```
H_{1,s_1} ← ∅
for S_1 ⊆ [n_1], |S_1| ≤ s_1 do
    T_1 ← ⋂_{i∈S_1} {j ∈ [n_2] : (i, j) ∈ Ω}
    if Σ_{i∈S_1} X_ij = 0 for all j ∈ T_1 then
        H_{1,s_1} ← H_{1,s_1} ∪ {S_1}
    end if
end for
```

Lemma 2. *For $i = 1, 2$ we have $|\mathcal{H}_i| \leq n_i^{s_i}$. Algorithm 1, for constructing $\mathcal{H}_1$, has complexity $O(s_1 n_1^{s_1} n_2)$. Similarly $\mathcal{H}_2$ can be constructed in $O(s_2 n_2^{s_2} n_1)$ operations.*

If the s_1 and s_2 are constants (w.r.t. n_1 and n_2) then the meta-algorithm of Section III-A runs in polynomial time.

B. Analysis

We next establish conditions on $X^\star$, Ω, and choices of s_1 and s_2, such that the meta-algorithm of Section III-A with $\mathcal{H}_{1,s_1}$ and $\mathcal{H}_{2,s_2}$ completes $X^\star$ with desired probability.

The basic issue with our choice of $\mathcal{H}_{i,s_i}$ for $i = 1, 2$ is that there is no mechanism to ensure that these collections of parity checks are actually consistent with $P_\Omega(X^\star)$. Nevertheless, if Ω is large enough, we might hope that $\text{span}\{e_{S_1} : \mathcal{H}_{1,s_1}\} = [\mathcal{C}_{\text{col}}^\perp]_{\le s_1}$ and $\text{span}\{e_{S_2} : \mathcal{H}_{2,s_2}\} = [\mathcal{C}_{\text{row}}^\perp]_{\le s_2}$. If both of these occur then $\mathcal{H}_{1,s_1}$ and $\mathcal{H}_{2,s_2}$ are consistent with $P_\Omega(X^\star)$.

Lemma 3. *Let $X^\star \in \mathbb{F}_2^{n_1 \times n_2}$ have rank r and row and column spaces $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$. If $\Omega \sim \mathcal{B}([n_1] \times [n_2], p)$ and $p \ge \max\{p_1, p_2\}$ where*

$$p_1 := \left(\frac{r \log(2) + s_1 \log(n_1) + \log(1/\epsilon)}{d(\mathcal{C}_{\text{row}})} \right)^{1/s_1} \quad \textit{and} \quad (4)$$

$$p_2 := \left(\frac{r \log(2) + s_2 \log(n_2) + \log(1/\epsilon)}{d(\mathcal{C}_{\text{col}})} \right)^{1/s_2} \quad (5)$$

then with probability at least $1-2\epsilon$, $\text{span}\{e_{S_1} : S_1 \in \mathcal{H}_{1,s_1}\} = [\mathcal{C}_{\text{col}}^\perp]_{\le s_1}$ and $\text{span}\{e_{S_2} : S_2 \in \mathcal{H}_{2,s_2}\} = [\mathcal{C}_{\text{row}}^\perp]_{\le s_2}$.

Proof. If we fix $S_1 \subseteq [n_1]$ then (in Algorithm 1) $T_1 \sim \mathcal{B}([n_2], p^{|S_1|})$ since any $j \in T_1$ if and only if $(i, j) \in \Omega$ for all $i \in S_1$. Let $x = e_{S_1}^T X^\star \in \mathcal{C}_{\text{row}}$. Suppose $P_{T_1}(x) = 0$, or equivalently that $S_1 \in \mathcal{H}_{1,s_1}$. Then the probability that $e_{S_1} \notin [\mathcal{C}_{\text{col}}^\perp]_{\le s_1}$ (i.e. $x \ne 0$) is bounded above by

$$\Pr[P_{T_1}(x) = 0 \text{ for some } x \in \mathcal{C}_{\text{row}} \setminus \{0\}] \le 2^r e^{-p^{|S_1|} d(\mathcal{C}_{\text{row}})}.$$

Taking a union bound over all $S_1 \subseteq [n_1]$ with $|S_1| \le s_1$, the probability that $\text{span}\{e_{S_1} : S_1 \in \mathcal{H}_{1,s_1}\} \ne [\mathcal{C}_{\text{col}}^\perp]_{\le s_1}$ is at most

$$\sum_{k=1}^{s_1} \binom{n_1}{k} 2^r e^{-p^k d(\mathcal{C}_{\text{row}})} \le n_1^{s_1} 2^r e^{-p^{s_1} d(\mathcal{C}_{\text{row}})} \le \epsilon.$$

Similarly, $\text{span}\{e_{S_2} : S_2 \in \mathcal{H}_{2,s_2}\} \ne [\mathcal{C}_{row}^\perp]_{\le s_2}$ with probability at most ϵ. A union bound over the two error events completes the proof. □

Combining this with an estimate of the probability that $\mathcal{H}_{1,s_1}$ and $\mathcal{H}_{2,s_2}$ are uniquely consistent with $P_\Omega(X^\star)$, gives our main technical result.

Theorem 1. *Let $X^\star \in \mathbb{F}_2^{n_1 \times n_2}$ have rank r and row and column spaces $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$. let $\mathcal{C}_1 = [\mathcal{C}_{\text{col}}^\perp]_{\le s_1}^\perp$ and $\mathcal{C}_2 = [\mathcal{C}_{\text{col}}^\perp]_{\le s_2}^\perp$. If $\Omega \sim \mathcal{B}([n_1] \times [n_2], p)$ and $p \ge \max\{p_0, p_1, p_2\}$ (where p_1 and p_2 are defined in (4) and (5)) and*

$$p_0 = \frac{\dim(\mathcal{C}_1)\dim(\mathcal{C}_2)\log(2) + \log(1/\epsilon)}{d(\mathcal{C}_1) d(\mathcal{C}_2)}$$

then $\mathcal{H}_{1,s_1}$ and $\mathcal{H}_{2,s_2}$ are uniquely consistent with $P_\Omega(X^\star)$ with probability at least $1 - 3\epsilon$.

Proof. Since $p \ge \max\{p_1, p_2\}$ we know from Lemma 3 that with probability at least $1 - 2\epsilon$, we have $\text{span}\{e_{S_1} : S_1 \in \mathcal{H}_{1,s_1}\} = \mathcal{C}_1^\perp$ and $\text{span}\{e_{S_2} : S_2 \in \mathcal{H}_{2,s_2}\} = \mathcal{C}_2^\perp$. (Hence $\mathcal{H}_{1,s_1}$ and $\mathcal{H}_{2,s_2}$ are consistent with $P_\Omega(X^\star)$.) Since $\mathcal{C}_1 \otimes \mathcal{C}_2$ has dimension $\dim(\mathcal{C}_1)\dim(\mathcal{C}_2)$ and minimum distance $d(\mathcal{C}_1)d(\mathcal{C}_2)$ (see, e.g., [6]), if $p \ge p_0$ then (by Lemma 1)

$$\{X \in \mathbb{F}_2^{n_1 \times n_2} : X \in \mathcal{C}_1 \otimes \mathcal{C}_2, \;\; P_\Omega(X) = P_\Omega(X^\star)\}$$

has one element with probability at least $1 - \epsilon$. □

We now simplify Theorem 1 to the setting where s_1 and s_2 are large enough that $[\mathcal{C}_{\text{col}}^\perp]_{s_1} = \mathcal{C}_{\text{col}}^\perp$ and $[\mathcal{C}_{\text{row}}^\perp]_{s_2} = \mathcal{C}_{\text{col}}^\perp$.

Corollary 1. *Let $X^\star \in \mathbb{F}_2^{n_1 \times n_2}$ have rank r and row and column spaces $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$. Suppose s_1 and s_2 are such that $\mathcal{C}_{\text{col}}^\perp = [\mathcal{C}_{\text{col}}]_{\le s_1}$ and $\mathcal{C}_{\text{row}}^\perp = [\mathcal{C}_{\text{row}}]_{\le s_2}$. If $\Omega \sim \mathcal{B}([n_1] \times [n_2], p)$ and $p \ge \max\{p_1, p_2\}$ (where p_1 and p_2 are defined in (4) and (5)) then $\mathcal{H}_{1,s_1}$ and $\mathcal{H}_{2,s_2}$ are uniquely consistent with $P_\Omega(X^\star)$ with probability at least $1 - 3\epsilon$.*

Proof. We first apply Theorem 1 with $\mathcal{C}_1 = \mathcal{C}_{\text{col}}$ and $\mathcal{C}_2 = \mathcal{C}_{\text{row}}$. In this case $p_0 = (r^2 \log(2) + \log(1/\epsilon))/(d_{\text{row}} d_{\text{col}})$. If $\max\{p_1, p_2\} \ge 1$ the conclusion follows. Otherwise,

$$1 \ge \max\{p_1, p_2\} \ge p_1^{s_1/(s_1+s_2)} p_2^{s_2/(s_1+s_2)} \ge p_1^{s_1} p_2^{s_2} \ge p_0$$

where the last inequality is straightforward to verify. □

C. Applications of Corollary 1

We now illustrate the use of Corollary 1. First we consider completing general rank r matrices from a random subset of entries. Then we consider completing large ($n > 2^r$) rank r matrices with 2^r distinct rows and columns. Finally we specialize our results to random rank r matrices.

1) General rank r matrices: The following result tells us that for a general rank r matrix, $\mathcal{C}_{\text{col}}^\perp$ and $\mathcal{C}_{\text{row}}^\perp$ are spanned by elements of weight at most $r + 1$.

Lemma 4. *Let $\mathcal{C} \subseteq \mathbb{F}_2^n$ be a subspace of dimension $0 \le r \le n - 1$. Then $\mathcal{C}^\perp = [\mathcal{C}^\perp]_{\le r+1}$.*

Sketch of proof. Up to permutation, $\mathcal{C}^\perp$ is spanned by the rows of $H = \begin{bmatrix} -Z & I_{n-r} \end{bmatrix}$ for some $Z \in \mathbb{F}_2^{n-r \times r}$. □

By combining Corollary 1 with Lemma 4 and our observations about the complexity of the meta-algorithm in Section III-B, we obtain the following result.

Theorem 2. *Let $X^\star \in \mathbb{F}_2^{n \times n}$ have rank r and row and column spaces $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$. If $\Omega \sim \mathcal{B}([n] \times [n], p)$ with*

$$p \ge \left(\frac{r \log(2) + (r+1)\log(n) + \log(1/\epsilon)}{\min\{d(\mathcal{C}_{\text{row}}), d(\mathcal{C}_{\text{col}})\}} \right)^{1/(r+1)}$$

then the meta-algorithm with $\mathcal{H}_i = \mathcal{H}_{i,r+1}$ for $i = 1, 2$ recovers $X^\star$ with probability at least $1 - 3\epsilon$ in time $O(n^{2r+3})$.

2) Rank r matrices with 2^r distinct rows and columns: If $X^\star$ is a rank r matrix with 2^r distinct rows and columns then every element of its row space appears as a row of $X^\star$ (and similarly for every element of the column space). In this case $\mathcal{C}_{\text{col}}^\perp$ and $\mathcal{C}_{\text{row}}^\perp$ are spanned by elements of weight at most 3.

Lemma 5. *If $X^\star$ has rank r and 2^r distinct rows then $\mathcal{C}_{\text{col}}^\perp = [\mathcal{C}_{\text{col}}^\perp]_{\leq 3}$. If, in addition, $X^\star$ has 2^r distinct columns then $\mathcal{C}_{\text{row}}^\perp = [\mathcal{C}_{\text{row}}^\perp]_{\leq 3}$.*

Sketch of proof. The sum of any pair of rows (resp. columns) of $X^\star$ is another row (resp. column) of $X^\star$. Hence any element of $\mathcal{C}_{\text{col}}^\perp$ (resp. $\mathcal{C}_{\text{row}}^\perp$) of weight k is the sum of an element of weight three and an element of weight at most $k-1$. The result follows by induction. □

By combining Corollary 1 with Lemma 5 and our observations about the complexity of the meta-algorithm in Section III-B, we obtain the following result.

Theorem 3. *Let $X^\star \in \mathbb{F}_2^{n\times n}$ have rank r, row and column spaces $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$, and 2^r distinct rows and 2^r distinct columns. If $\Omega \sim \mathcal{B}([n]\times[n], p)$ with*

$$p \geq \left(\frac{r\log(2) + 3\log(n) + \log(1/\epsilon)}{\min\{d(\mathcal{C}_{\text{row}}), d(\mathcal{C}_{\text{col}})\}}\right)^{1/3}$$

then the meta-algorithm with $\mathcal{H}_i = \mathcal{H}_{i,3}$ for $i = 1, 2$ recovers $X^\star$ with probability at least $1 - 3\epsilon$ in time $O(n^7)$.

3) Random rank r matrices: The results of the previous two sections are most interesting in the case where $X^\star = U_1 U_2^T$ where $U_1, U_2 \in \mathbb{F}_2^{n\times r}$ are indepdent random matrices with i.i.d. Bernoulli entries, and r is fixed and n is growing. In this setting, with high probability $d(\mathcal{C}_{\text{col}}) \geq \delta n$ and $d(\mathcal{C}_{\text{row}}) \geq \delta n$ for some constant δ (see, e.g., [12]). In this setting Theorem 2 shows that as long as the number of random revealed entries is about $n^2 p \geq \tilde{\Omega}(n^{2-\frac{1}{r+1}})$ (ignoring logarithmic factors and quantities independent of n), a *vanishing fraction* of the total number of entries, we can recover $X^\star$ with high probability (inverse polynomial in n) in time $O(n^{2r+3})$.

We now consider the case $n > \log(2) r 2^r$ in which it is very likely that $X^\star = U_1 U_2^T$ has 2^r distinct rows and colunms.

Lemma 6. *Let $U_1, U_2 \in \mathbb{F}_2^{n\times r}$ be independent with i.i.d. Bernoulli entries. If $n \geq \log(2) r 2^r + \log(\epsilon^{-1}) 2^r$ then $U_1 U_2^T$ has 2^r distinct rows and 2^r distinct columns with probability at least $1 - 2\epsilon$.*

Proof. It is enough that U_1 (resp. U_2) has 2^r distinct rows (resp. columns) with probability at least $1 - \epsilon$. Each row of U_1 is independent and takes on 2^r different values each with equal probability, so we estimate the probability that n is at least the stopping time for the coupon collector problem with 2^r coupons via a standard tail bound [13, Proposition 2.4]. □

It follows from Theorem 3 that as long as $n > cr2^r$ (for some constant c) and the number of random revealed entries is about $n^2 p \geq \tilde{\Omega}(n^{2-1/3})$ we can recover $X^\star$ with high probability in time polynomial in n.

V. Linear programming-based algorithms

Let $\mathcal{H}_1 \subseteq 2^{[n_1]}$ and $\mathcal{H}_2 \subseteq 2^{[n_2]}$ be collections of parity checks. For instance, we could take $\mathcal{H}_1 = \mathcal{H}_{1,s_1}$ and $\mathcal{H}_2 = \mathcal{H}_{2,s_2}$ from Section IV. In this section we describe one way to select large subsets of $\mathcal{H}_1$ and $\mathcal{H}_2$ such that the selected subsets are consistent with $P_\Omega(X^\star)$. By removing enough elements from $\mathcal{H}_1$ and $\mathcal{H}_2$, we can always make them consistent with $P_\Omega(X^\star)$, but removing too many makes it unlikely they will be uniquely consistent. We formulate this subset selection problem as a combinatorial optimization problem and then relax it to an LP.

Our combinatorial formulation has decision variables $X \in \{0,1\}^{n_1\times n_2}$ (intended to represent a completion of $P_\Omega(X^\star)$), and $Y_{S_1} \in \{0,1\}$ for $S_1 \in \mathcal{H}_1$ and $Z_{S_2} \in \{0,1\}$ for $S_2 \in \mathcal{H}_2$ that indicate which elements of $\mathcal{H}_1$ and $\mathcal{H}_2$ remain in our consistent set. The objective aims to maximize the number of parity checks we keep. The normalization factors $N_{i,j} := |\{S : S \in \mathcal{H}_i,\ |S| = j\}|$ correct for the fact that low-weight checks can combine to form higher weight checks so their inclusion should be penalized more. The formulation then is

$$\max_{X,Y,Z} \sum_{j=1}^{n_1} \sum_{S\in\mathcal{H}_1:|S|=j} \frac{Y_S}{N_{1,j}} + \sum_{j=1}^{n_2} \sum_{S\in\mathcal{H}_2:|S|=j} \frac{Z_S}{N_{2,j}} \tag{6}$$

$$\text{s.t. } X_{ij} = X^\star_{ij}\ \ \forall (i,j)\in\Omega,\ X_{ij}\in\{0,1\}\ \ \forall (i,j)\notin\Omega$$

$$Y_S \in \{0,1\} \text{ for } S\in\mathcal{H}_1,\ \ Z_S\in\{0,1\} \text{ for } S\in\mathcal{H}_2$$

$$Y_S = 1 \implies \textstyle\sum_{i\in S} X_{ij} = 0 \pmod 2 \text{ for } j\in[n_2] \tag{7}$$

$$Z_S = 1 \implies \textstyle\sum_{j\in S} X_{ij} = 0 \pmod 2 \text{ for } i\in[n_1]. \tag{8}$$

A. Linear programming relaxation

We now describe a simple LP relaxation of this combinatorial optimization problem. We relax $X_{ij}, Y_S, Z_S \in \{0,1\}$ to $X_{ij}, Y_S, Z_S \in [0,1]$. If $S = \{i_1, i_2, \ldots, i_s\} \subseteq [n_1]$ we relax (7) to

$$(Y_S, X_{i_1 j}, \ldots, X_{i_s j}) \in \widetilde{P}_s := \text{conv}\{\text{integer solutions of (7)}\}.$$

Similarly, if $S = \{j_1, j_2, \ldots, j_s\} \subseteq [n_2]$ we relax (8) to

$$(Z_S, X_{ij_1}, \ldots, X_{ij_s}) \in \widetilde{P}_s := \text{conv}\{\text{integer solutions of (8)}\}.$$

These constraints generalize those used for LP decoding of binary linear codes [11]. Here the additional variables Y_S and Z_S can turn 'on' and 'off' the constraint for a given parity check. The polytopes $\tilde{P}_s$ have compact linear inequality descriptions. For $s = 1, 2, 3$ these are

$$\widetilde{P}_1 = \{(t,x)\in[0,1]^2 : x \leq 1-t\}$$

$$\widetilde{P}_2 = \{(t,x,y)\in[0,1]^3 : t-1 \leq x-y \leq 1-t\}$$

$$\widetilde{P}_3 = \big\{(t,x,y,z)\in[0,1]^2 : x+y+z \leq 3-t,$$
$$x-y-z\leq 1-t,\ y-x-z\leq 1-t,\ z-x-y\leq 1-t\big\}.$$

Overall our LP relaxation is as follows:

$$\max_{X,Y,Z} \sum_j \sum_{S\in\mathcal{H}_1:|S|=j} \frac{Y_S}{N_{1,j}} \sum_j \sum_{S\in\mathcal{H}_2:|S|=j} \frac{Z_S}{N_{2,j}} \tag{9}$$

$$\text{s.t. } X_{ij} = X^\star_{ij}\ \ \forall (i,j)\in\Omega,\ \ X_{ij}\in[0,1]\ \ \forall (i,j)\notin\Omega$$

$$Y_S\in[0,1]\ \ \forall S\in\mathcal{H}_1,\ \ Z_S\in[0,1]\ \ \forall S\in\mathcal{H}_2$$

$$\text{for all } S=\{i_1,\ldots,i_s\}\in\mathcal{H}_1 \text{ and all } j\in[n_2]$$
$$(Y_S, X_{i_1 j}, \ldots, X_{i_s,j})\in\widetilde{P}_s$$

$$\text{for all } S=\{j_1,\ldots,j_s\}\in\mathcal{H}_2 \text{ and } i\in[n_1]$$
$$(Z_S, X_{ij_1}, \ldots, X_{ij_s})\in\widetilde{P}_s.$$

This relaxation is quite weak, keeping only a small set of linear inequalities valid for the integer program. Nevertheless, it seems to perform well in the numerical experiments reported in Section VI.

B. Modifications for implementation

To reduce the computational effort required, for our experiments we actually solve three LPs in sequence to construct the sets $\mathcal{H}_1$ and $\mathcal{H}_2$ that we will pass to the meta-algorithm. We now briefly sketch this sequence of LPs.

First we solve (9) with $\mathcal{H}_1 = \mathcal{H}_{1,1}$ (i.e. $s_1 = 1$) and $\mathcal{H}_2 = \mathcal{H}_{2,1}$ (i.e. $s_2 = 1$). If $Y_{\{i\}} = 1$ (resp. $Z_{\{j\}} = 1$) then $P_\Omega(X^\star)$ has a completion with row i (resp. column j) being zero.

We then assume these are correct relations for $X^\star$ and restrict to the $n_1' \times n_2'$ submatrix $X^{\star\prime}$ corresponding to the non-zero rows and columns (and accordingly restrict Ω to the appropriate $\Omega' \subseteq [n_1'] \times [n_2']$). We now seek the weight two relations among the rows and columns that are consistent with $P_{\Omega'}(X^{\star\prime})$. To do this we solve (9) for this smaller problem with $\mathcal{H}_1 = \mathcal{H}_{1,2}$ and $\mathcal{H}_2 = \mathcal{H}_{2,2}$. If $Y_{\{i_1,i_2\}} = 1$ (resp. $Z_{\{j_1,j_2\}} = 1$) then $P_{\Omega'}(X^{\star\prime})$ has a completion with rows i_1 and i_2 (resp. columns j_1 and j_2) being *identical*.

As before we assume these are correct relations for $X^\star$ and restrict to an $n_1'' \times n_2''$ submatrix $X^{\star\prime\prime}$ of $X^{\star\prime}$ indexed by taking one copy of each repeated row (resp. column). The corresponding set of revealed entries Ω'' is obtained by taking $(i,j) \in \Omega''$ if and only if $(k,\ell) \in \Omega'$ for k indexing a row identical to i, and ℓ indexing a row identical to j. We now seek weight three relations among the rows and columns that are consistent with $P_{\Omega''}(X^{\star\prime\prime})$. To do this we solve (9) with $\mathcal{H}_1 = \mathcal{H}_{1,3}$ and $\mathcal{H}_2 = \mathcal{H}_{2,3}$. If $Y_{\{i_1,i_2,i_3\}} = 1$ (resp. $Z_{\{j_1,j_2,j_3\}} = 1$) then $P_{\Omega''}(X^{\star\prime\prime})$ has a completion with rows i_1, i_2 and i_3 (resp. columns j_1, j_2 and j_3) summing to zero.

From these three stages (after appropriate re-indexing) we have constructed (in polynomial time) a collection $\mathcal{H}_1$ of row parity checks and $\mathcal{H}_2$ of column parity checks. We pass these to the meta-algorithm of Section III-A.

VI. Numerical experiments

We conclude by describing two experiments to evaluate the LP-based method from Section V-B. Each investigates for which p we can recover a random $n \times n$ matrix of rank r from a random observed set of entries $\Omega \sim \mathcal{B}([n] \times [n], p)$.

A. Fixed n, varying r

For fixed $n = 100$, each $r = 3, 4, \ldots, 10$, and each $\kappa = 1.5, 1.75, \ldots, 3.5$ we carried out the following 15 times:

1) Let $U_1, U_2 \in \mathbb{F}_2^{n \times r}$ be independent with i.i.d. Bernoulli entries and let $X^\star = U_1 U_2^T$.
2) Sample $\Omega \sim \mathcal{B}([n] \times [n], \kappa r(2n-r)/n^2)$.
3) Run the algorithm from Section V-B and record whether or not it successfully completes $X^\star$ from $P_\Omega(X^\star)$.

The results of this experiment are shown on the left in Figure 1. In particular it appears that as long as $\kappa \geq 3$ (i.e. we observe about $3r(200-r)$ entries) then the LP-based method is typically successful.

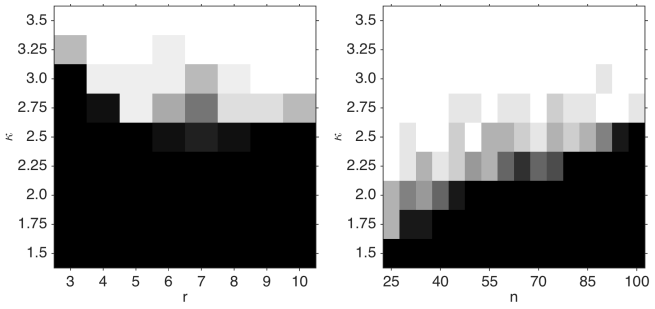

Fig. 1. Phase-transition plots for the success of the method of Section V-B for completing $n \times n$ binary matrices of rank r given entries revealed independently with probability $p = \kappa r(2n-r)/n^2$. The pixel corresponding to (r, κ) is black if the method failed on all attempts and white if it succeeded on all attempts. The grayscale intensity indicates the proportion of successful trials. On the left is the result of 15 trials with $n = 100$ and $r = 3, 4, \ldots, 10$. On the right is the result of 10 trials with $r = 5$ and $n = 25, 30, \ldots, 100$.

B. Fixed r, varying n

For fixed $r = 5$, each $n = 25, 30, \ldots, 100$, and each $\kappa = 1.5, 1.75, \ldots, 3.5$ we carried out, 10 times, the three steps of Section VI-A. The results of this experiment are shown on the right in Figure 1. It appears that as long as κ grows mildly, perhaps logarithmically, with n (i.e. we observe about $5c\log(n)(2n-5)$ entries for some constant c), then the LP-based method is typically successful. Since $\log_2(N_{n,r}) \geq r(2n-r)$ [8] our results suggest that (for $r \leq 10$ and $n \leq 100$) our LP-based method is close to optimal.

References

[1] E. J. Candés and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 4, pp. 717–772, 2012.

[2] Y. Birk and T. Kol, "Informed-source coding-on-demand (ISCOD) over broadcast channels," in *Proc. 17th Joint Conf. IEEE Computer and Comm. Societies (INFOCOM'98)*, vol. 3. IEEE, 1998, pp. 1257–1264.

[3] Z. Bar-Yossef, Y. Birk, T. S. Jayram, and T. Kol, "Index coding with side information," *IEEE Trans. Information Theory*, vol. 57, no. 3, pp. 1479–1494, 2011.

[4] S. El Rouayheb, A. Sprintson, and C. Georghiades, "On the index coding problem and its relation to network coding and matroid theory," *IEEE Trans. Information Theory*, vol. 56, no. 7, pp. 3187–3195, 2010.

[5] H. Esfahanizadeh, F. Lahouti, and B. Hassibi, "A matrix completion approach to linear index coding problem," *arXiv:1408.3046*, 2014.

[6] F. R. Kschischang, "Product codes," *Encyclopedia of Telecommunications*, 2003.

[7] N. Harvey, D. Karger, and S. Yekhanin, "The complexity of matrix completion," *Proc. 17th ACM-SIAM Symp. Discrete Algorithms (SODA)*, pp. 1103–1111, 2006.

[8] V. Y. F. Tan, L. Balzano, and S. C. Draper, "Rank minimization over finite fields: Fundamental limits and coding-theoretic interpretations," *IEEE Trans. Information Theory*, vol. 58, no. 4, pp. 2018–2039, 2012.

[9] K. Shanmugam, A. G. Dimakis, and M. Langberg, "Graph theory versus minimum rank for index coding," in *Proc. Intl. Symp. Information Theory (ISIT 2014)*. IEEE, 2014, pp. 291–295.

[10] S. Vishwanath, "Information theoretic bounds for low-rank matrix completion," in *Proc. Intl. Symp. Information Thoery (ISIT 2010)*. IEEE, 2010, pp. 1508–1512.

[11] J. Feldman, "Decoding error-correcting codes via linear programming," Ph.D. dissertation, MIT, 2003.

[12] A. Barg and G. D. Forney, "Random codes: minimum distances and error exponents," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2568–2573, 2002.

[13] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Math. Soc., 2009.