# Consistent Collective Matrix Completion under Joint Low Rank Structure


**Suriya Gunasekar**
suriya@utexas.edu
University of Texas,
Austin, USA

**Makoto Yamada**
makotoy@yahoo-inc.com
Yahoo Labs,
Sunnyvale, USA

**Dawei Yin**
daweiy@yahoo-inc.com
Yahoo Labs,
Sunnyvale, USA

**Yi Chang**
yichang@yahoo-inc.com
Yahoo Labs,
Sunnyvale, USA



## Abstract

We address the collective matrix completion problem of jointly recovering a collection of matrices with shared structure from partial (and potentially noisy) observations. To ensure well–posedness of the problem, we impose a joint low rank structure, wherein each component matrix is low rank and the latent space of the low rank factors corresponding to each entity is shared across the entire collection. We first develop a rigorous algebra for representing and manipulating collective–matrix structure, and identify sufficient conditions for consistent estimation of collective matrices. We then propose a tractable convex estimator for solving the collective matrix completion problem, and provide the first non–trivial theoretical guarantees for consistency of collective matrix completion. We show that under reasonable assumptions stated in Sec. 3.1, with high probability, the proposed estimator exactly recovers the true matrices whenever sample complexity requirements dictated by Theorem 1 are met. The sample complexity requirement derived in the paper are optimum up to logarithmic factors, and significantly improve upon the requirements obtained by trivial extensions of standard matrix completion. Finally, we propose a scalable approximate algorithm to solve the proposed convex program, and corroborate our results through simulated and real life experiments.


## 1 Introduction

Affinity relationships between a pair of *entity types* (e.g. users, movies, documents, explicit features, etc.) are often represented in a matrix form. The standard matrix completion task of predicting the missing entries of a matrix from partial (and potentially noisy) observations is at the core of a wide range of applications including recommendation systems, recovering gene–protein interactions, and modeling text document collections, among others [21, 13, 36]. In many practical applications, data from multiple matrices often share correlated information, and leveraging the shared structure can potentially enhance performance. For example, in e–commerce applications, user preferences in multiple domains such as news, ads, etc., and explicit user/item feature information such as demographics, social network, text description, etc., are made available in the form of a "collection of matrices" sharing interactions among a common set of users/items.

*Collective matrix completion* involves simultaneously completing one or more partially observed matrices by leveraging data from a set of correlated matrices. Each component matrix, also called a *view*, represents pairwise affinity relation among $K$ types of *entities*. We assume a *joint low rank* structure, wherein each entity type $k$ has a low dimensional latent factor representation $U_k$; and each view $v$ representing the affinity between entity types $k_1$ and $k_2$ is a low rank matrix given by $U_{k_1}U_{k_2}^\top$. Leveraging such shared structure is especially attractive in scenarios where standard matrix completion typically fails, such as: (i) *Insufficient Data*: Data sparsity in one view can often be mitigated by augmenting data from related views. For example, in a multiple recommendation systems, user's interests can be better captured by combining data from multiple sources; (ii) *Cold Start*: Recommendation for new users/items with no prior ratings can be partially addressed in collective matrix completion using additional data like explicit user/item features.

However, the problem of collective–matrix completion, like standard matrix completion, is statistically ill–posed as: (a) only a decaying fraction of the number of entries in a matrix are observed; (b) the observations are localized (e.g. individual matrix entries as opposed to random linear measurements). Recent works on matrix completion leverage the developments in high dimensional esti-

---


Appearing in Proceedings of the $18^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

mation [26, 9, 35, 4], and propose statistically consistent tractable estimators under low rank and other structural assumptions [6, 5, 19, 20, 27, 14, 25, 11, 18, 15, 10]. However, to the best of our knowledge, optimal sample complexity requirements for statistically consistent recovery of collective–matrices has not been previously analyzed.

In this paper, we propose a convex estimator for collective matrix completion and provide the first non–trivial theoretical guarantees for consistent recovery of collective–matrices. In a close related work, Bouchard et al. [2] propose the first convex estimator for collective matrix completion without analyzing the consistency of the estimate. In comparison to the analysis for standard matrix completion, several new challenges are encountered in collective matrix completion:

(a) Trivial extensions of sample complexity from existing results on standard matrix completion are suboptimal as they do not consider the shared structure. Thus, fully leveraging the joint low–rank structure in the analysis is the key to obtain optimal sample complexity.

(b) Unlike matrices, for collective matrices with joint low rank structure, the entity factors $U_k$ are not always unique (upto signs and normalization). However, we observe that, under the assumptions in Sec. 3.1, even when $U_k$ are not unique, the $V$ relevant interactions are uniquely captured.

(c) For general collective–matrix structures, a joint factorization may not always exist (even with full rank), and further the proposed convex estimator can be badly behaved, we enforce Assumption 3 to avoid these cases; although this assumption can potentially be relaxed.

To summarize our contributions:

(i) In Sec. 2 and 3, we develop a rigorous algebra for representing and manipulating collective–matrices. We identify sufficient conditions (Assumptions 1–3) under which consistent recovery is feasible, and propose a tractable convex estimator for collective matrix completion.

(ii) We provide the first theoretical guarantee for consistent collective matrix completion (Theorem 1). Specifically, we show that for a subset of collective–matrix structures, with high probability, the proposed estimator exactly recovers the true matrices whenever the sample complexity satisfies $\forall k$, $|\Omega_k| \sim O(n_k R \overline{\log} N)$, where $n_k$ is the number of entities of type $k$, $R$ is the joint rank of the collective matrices, and $|\Omega_k|$ is the expected number of observations corresponding to entity $k$. We note that these rates are optimal upto logarithmic factors. Although our analysis is for a noise–free setting, analogous estimators for recovery under non–adversarial noise are proposed without guarantees.

(iii) Finally, while the proposed convex program can be solved by adapting the Singular Value Thresholding for Collective Matrix Completion (SVT–CMC) algorithm proposed by Bouchard et al. [2, 3, 34], this algorithm is not scalable to large datasets. As a minor contribution, we adapt Hazan's algorithm [16] to provide an approximate solution for the proposed convex program (Sec. 4.2). The proposed algorithm has a significantly better per iteration complexity as compared to SVT–CMC, and can be used to tradeoff accuracy for computation in large datasets. We conclude the paper by corroborating our results through experiments on simulated and real life datasets (Sec. 6).

Besides the convex estimator, related work for collective matrix completion includes various non–convex estimators and probabilistic models. A seminal paper on low rank collective matrix factorization is the work by Singh et al. [32], wherein the views are parameterized by the shared latent factor representation. The latent factors are learnt by minimizing a regularized loss function over the estimates. A Bayesian model for collective matrix factorization was also proposed by the same authors [30, 31]. Collective matrix factorization is also related to applications involving multi–task learning and tensor factorization [23, 22, 1, 37, 38]. For the special case of low rank matrix completion, besides the theoretical guarantees, there are plenty of equally significant work that propose effective and scalable algorithms, including max–margin matrix factorization [33], alternating minimization [21, 39], and probabilistic models [24, 29], among others.

## 2 Collective–Matrix Structure

In this section we introduce equivalent representations for the collective–matrix structure and develop basic algebra for analyzing and manipulating collective–matrices.

### 2.1 Basic Notations

Matrices are denoted by uppercase letters, $X$, $M$, etc. Matrix inner product is given by $\langle X, Y\rangle = \sum_{(i,j)} X_{ij}Y_{ij}$. The set of symmetric matrices of dimension $N$ is denoted as $\mathbb{S}^N$. For $M \in \mathbb{R}^{m\times n}$, with singular values $\sigma_1 \geq \sigma_2 \geq \ldots$, common matrix norms include the *nuclear norm* $\|M\|_* = \sum_i \sigma_i$, the *spectral norm* $\|M\|_2 = \sigma_1$, and the *Frobenius norm* $\|M\|_F = \sum_i \sigma_i^2 = \sum_{ij} M_{ij}^2$.

**Definition 1** (Dual Norm)**.** *Given any norm $\|\cdot\|$ defined on a metric space $\mathcal{V}$, the dual norm, $\|\cdot\|^*$ defined on dual space $\mathcal{V}^*$ is given by $\|X\|^* = sup_{\|Y\|\leq 1}\langle X, Y\rangle$.*

**Definition 2** (Operator Norm)**.** *Given a linear operator $\mathcal{P}: \mathcal{V} \to \mathcal{W}$, the operator norm of $\mathcal{P}$ is given by $\|\mathcal{P}\|_{op} = \underset{X\in\mathcal{V}\backslash\{0\}}{sup} \frac{\|\mathcal{P}(X)\|_{\mathcal{W}}}{\|X\|_{\mathcal{V}}}$, where $\|.\|_{\mathcal{V}}$ and $\|.\|_{\mathcal{W}}$ are the Euclidean norms in the respective spaces.*

### 2.2 Collective–Matrix Representation

A collective–matrix, denoted using script letters, $\mathcal{X}$, $\mathcal{M}$, etc., is a collection of affinity relations among a set of $K$

types of *entities*; and is primarily represented as a list of $V$ matrices, $\mathcal{X} = [X_v]_{v=1}^V = [X_v : v = 1, 2, \ldots, V]$. Each component matrix $X_v$, called a *view*, is the affinity matrix between a pair of entity types denoted by $r_v$ (entity type along rows) and $c_v$ (entity type along columns). We only consider static undirected affinity relations, wherein, for a given pair of entity types $k_1, k_2 \in \{1, 2, \ldots K\}$, there is at most one affinity relation $X_v$ defined between $k_1$ and $k_2$.

The entity–relationship structure defining a collective–matrix is represented by an undirected graph $\mathcal{G}$, with nodes denoting the $K$ entity types, and an edge between nodes $k_1$ and $k_2$ implying that a view $X_v$ with either $(r_v = k_1, c_v = k_2)$ or $(r_v = k_2, c_v = k_1)$ exists in the collective matrix. We assume that the graph $\mathcal{G}$ forms a single connected component, if not, each connected component could be handled separately without loss of generality. An illustration of a collective matrix structure $\mathcal{X}$ and its entity–relationship graph $\mathcal{G}$ is given in Fig. 1 $(a)$–$(b)$.

For $k = 1, 2, \ldots, K$, denote the number of instances of the $k^{\text{th}}$ entity type by $n_k$; let $N = \sum_k n_k$. Then, $\forall v$, $X_v \in \mathbb{R}^{n_{r_v} \times n_{c_v}}$, and collective–matrices with common entity–relationship graph $\mathcal{G}$ belong to the space:

$$\mathfrak{X} = \mathbb{R}^{n_{r_1} \times n_{c_1}} \times \mathbb{R}^{n_{r_2} \times n_{c_2}} \times \ldots \times \mathbb{R}^{n_{r_V} \times n_{c_V}}.$$

Finally, $\forall v$, $\mathcal{I}(v) = \{(i,j) : i \in [n_{r_v}], j \in [n_{c_v}]\} = [n_{r_v}] \times [n_{c_v}]$ denotes the set indices representing the elements in view $v$, where $[N] = \{1, 2, \ldots, N\}$.

### 2.2.1 Equivalent Representations

For mathematical convenience, we introduce two alternate (equivalent) representations for collective–matrices. These are used interchangeably in the rest of the paper.

1. **Entity Matrix Set Representation:** A collective–matrix $\mathcal{X}$, can be equivalently represented as a set of $K$ matrices $\mathbb{X} = [\mathbb{X}_k]_{k=1}^K$, such that $\mathbb{X}_k$ is a matrix formed by concatenating (appropriately transposed) views involving the entity type $k$. Let $\mathbf{1}_E$ denote the indicator variable for statement $E$, and the operator hcat$\{\}$ denote horizontal concatenation of a list. We then have the column dimension of $\mathbb{X}_k$ given by $m_k = \sum_{v=1}^V n_{c_v} \mathbf{1}_{(r_v = k)} + n_{r_v} \mathbf{1}_{(c_v = k)}$, and

$$\mathbb{X}_k := \text{hcat}\big\{[X_v \mathbf{1}_{(r_v=k)}, X_v^\top \mathbf{1}_{(c_v=k)}]_{v=1}^V\big\} \in \mathbb{R}^{n_k \times m_k}.$$

2. **Block Matrix Representation:** Collective–matrices can also be represented as blocks in a symmetric matrix of size $N \times N$, where $N = \sum_k n_k$ [2]. For a symmetric matrix $Z \in \mathbb{S}^N$, we identify $K \times K$ blocks, wherein the $(k_1, k_2)$ block, denoted as $Z[k_1, k_2]$, is of dimension $n_{k_1} \times n_{k_2}$. Block matrix representation for $\mathcal{X}$ is given by:

$$\mathcal{B}(\mathcal{X})[k_1, k_2] = \begin{cases} X_v & \text{if } \exists v, \text{ s.t. } r_v = k_1, c_v = k_2 \\ X_v^\top & \text{if } \exists v, \text{ s.t. } r_v = k_2, c_v = k_1 \\ 0 & \text{otherwise.} \end{cases}$$

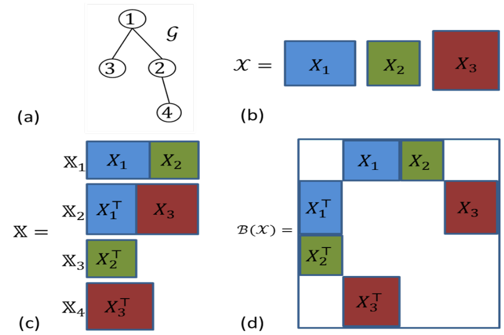



Figure 1: An illustration of the various collective–matrix representations described in Section 2

We define operators $P_v : S^N \to \mathbb{R}^{n_{r_v} \times n_{c_v}}$, such that $P_v(Z) = Z[r_v, c_v]$; and $\forall Z \in \mathbb{S}^N$, $\mathcal{Z} = [P_v(Z)]_{v=1}^V \in \mathfrak{X}$.

These alternate representations for collective–matrix structure are illustrated in Figure 1 $(c)$ and $(d)$, respectively.

## 2.3 Collective–Matrix Algebra

**Collective–Matrix Inner Product and Euclidean Norm**

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{v=1}^V \langle X_v, Y_v \rangle, \text{ and } \|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}.$$

**Note:** We overload the notation for inner product $\langle \cdot, \cdot \rangle$, and the Frobenius norm $\| \cdot \|_F$ for matrices and collective–matrices, with operands providing disambiguation.

**Standard Orthonormal Basis** The *standard orthonormal basis* for $\mathfrak{X}$ is given by $\{\mathcal{E}^{(v, i_v, j_v)} : v \in [V], (i_v, j_v) \in \mathcal{I}(v)\}$, where $\mathcal{E}^{(v, i_v, j_v)} \in \mathfrak{X}$ has a value of 1 in the $(i_v, j_v)^{\text{th}}$ element of view $v$, and 0 everywhere else. Recall that $[n] = \{1, 2, \ldots, n\}$, and $\mathcal{I}(v) = [n_{r_v}] \times [n_{c_v}]$.

**Joint Factorization and Collective–Matrix Rank** A collective-matrix $\mathcal{X} \in \mathfrak{X}$ is said to possess an $R$–*dimensional joint factorization*, if there exists a set of factors $\{U_k \in \mathbb{R}^{n_k \times R}\}_{k=1}^K$, such that $\forall v$, $X_v = U_{r_v} U_{c_v}^\top$. The set of collective–matrices in $\mathfrak{X}$ that have a joint factorization structure of finite dimension is denoted by $\bar{\mathfrak{X}} \subseteq \mathfrak{X}$. For $\mathcal{X} \in \bar{\mathfrak{X}}$, the *collective–matrix rank* is defined as the minimum value of $R$ such that an $R$–dimensional joint factorization exists for $\mathcal{X}$.

## 2.4 Atomic Decomposition of Collective–Matrices

Consider the following set of rank–1 collective–matrices:

$$\mathscr{A} = \text{ext}(\text{conv}\{[P_v(uu^\top)]_{v=1}^V : u \in \mathbb{R}^N, \|u\|_2 = 1\}), \tag{1}$$

where conv() and ext() return the convex hull and the extreme points of a set, respectively. Recall that $N = \sum_k n_k$, and $P_v : \mathbb{S}^N \to \mathbb{R}^{n_{r_v} \times n_{c_v}}$ extracts the block corresponding to the view $v$ in an $N \times N$ symmetric matrix. From the block matrix representation (Sec. 2.2.1), note that $\mathfrak{X} = \text{aff}(\mathscr{A})$; and the following proposition can be easily verified:

**Proposition 1.** *A collective–matrix has a joint factorization structure if and only if it belongs to the conic hull of $\mathscr{A}$, i.e. $\bar{\mathfrak{X}} = cone(\mathscr{A})$.* □

We define the following quantities of interest:

**Collective–Matrix Atomic Norm**: also the gauge of $\mathscr{A}$,

$$\|\mathcal{X}\|_{\mathscr{A}} := \inf\{t > 0 : \mathcal{X} \in t \cdot \text{conv}(\mathscr{A})\}. \tag{2}$$

**Support function** of $\mathscr{A}$:

$$\|\mathcal{X}\|^*_{\mathscr{A}} := \sup\{\langle \mathcal{X}, \mathcal{A}\rangle : \mathcal{A} \in \mathscr{A}\}. \tag{3}$$

**"sign" collective–matrices** of $\mathcal{X}$:

$$\mathscr{E}(\mathcal{X}) = \{\mathcal{E} : \|\mathcal{X}\|_{\mathscr{A}} = \langle \mathcal{E}, \mathcal{X}\rangle, \|\mathcal{E}\|^*_{\mathscr{A}} = 1\}. \tag{4}$$

**Remarks**

1. $\|\mathcal{X}\|_{\mathscr{A}}$ is not always a norm. It is a norm if $\mathscr{A}$ is centrally symmetric, i.e. if $\mathcal{A} \in \mathscr{A} \Leftrightarrow -\mathcal{A} \in \mathscr{A}$.
2. By convention, $\|\mathcal{X}\|_{\mathscr{A}} = \infty$ if $\mathcal{X} \in \mathfrak{X} \setminus \bar{\mathfrak{X}}$.
3. However, $\|\mathcal{X}\|_{\mathscr{A}}$ is always a convex function and exhibits many norm–like properties. $\forall \mathcal{X} \in \mathfrak{X}$, $\|\mathcal{X}\|_{\mathscr{A}} \geq 0$ and $\|\mathcal{X}\|_{\mathscr{A}} = 0$ iff $\mathcal{X} = 0$; $\forall a \geq 0$, $\|a\mathcal{X}\|_{\mathscr{A}} = a\|\mathcal{X}\|_{\mathscr{A}}$; and $\|\mathcal{X} + \mathcal{Y}\|_{\mathscr{A}} \leq \|\mathcal{X}\|_{\mathscr{A}} + \|\mathcal{Y}\|_{\mathscr{A}}$.
4. If $\|\mathcal{X}\|_{\mathscr{A}}$ is a norm, then $\|\mathcal{X}\|^*_{\mathscr{A}}$ is its dual norm.

#### 2.4.1 Primal Dual representation

For all $\mathcal{X} \in \bar{\mathfrak{X}}$, $\|\mathcal{X}\|_{\mathscr{A}} < \infty$, and the atomic norm defined in (2), can be equivalently defined using the following primal and dual optimization problems.

$$(P) \;\; \|\mathcal{X}\|_{\mathscr{A}} = \min_{\{\lambda_r \geq 0\}} \textstyle\sum_r \lambda_r \;\text{ s.t. } \sum_r \lambda_r \mathcal{A}_r = \mathcal{X}, \tag{5}$$

$$(D) \;\; \|\mathcal{X}\|_{\mathscr{A}} = \max_{\mathcal{Y} \in \mathfrak{X}} \langle \mathcal{X}, \mathcal{Y}\rangle \;\text{ s.t. } \|\mathcal{Y}\|^*_{\mathscr{A}} \leq 1. \tag{6}$$

**Proposition 2.** *$\forall \mathcal{X} \in \bar{\mathfrak{X}}$, convex programs $(P)$ and $(D)$ defined above are equivalent to:*

$$(P) \;\; \|\mathcal{X}\|_{\mathscr{A}} = \min_{Z \in \mathbb{S}^N} tr(Z) \quad s.t.\; P_v[Z] = X_v \forall\, v,$$

$$(D) \;\; \|\mathcal{X}\|_{\mathscr{A}} = \max_{\mathcal{Y} \in \mathfrak{X}} \langle \mathcal{X}, \mathcal{Y}\rangle \quad s.t.\; \frac{1}{2}\mathcal{B}(\mathcal{Y}) \preccurlyeq \mathbb{I}.$$

## 3 Convex Collective–Matrix Completion

Denote the ground truth collective–matrix as $\mathcal{M} \in \bar{\mathfrak{X}}$. The task in collective–matrix completion is to recover $\mathcal{M}$ from a subset of the (potentially noisy) entries of $\mathcal{M}$. Denote the indices of observed entries by $\Omega = \{(v_s, i_s, j_s) : (i_s, j_s) \in \mathcal{I}(v_s), s = 1, 2, \ldots, |\Omega|\}$. For conciseness, we denote the standard basis corresponding to indices in $\Omega$ as $\forall s$, $\mathcal{E}^{(s)} = \mathcal{E}^{(v_s, i_s, j_s)}$. Further, we define the operator $P_\Omega$ as:

$$P_\Omega(\mathcal{X}) = \textstyle\sum_{s=1}^{|\Omega|} \langle \mathcal{X}, \mathcal{E}^{(s)}\rangle \mathcal{E}^{(s)}. \tag{7}$$

We consider two observation models:

1. Noise–free model: $\mathcal{M}$ is observed on $\Omega$ without any noise, i.e. $\forall s, y_s = \langle \mathcal{M}, \mathcal{E}^{(s)}\rangle$.

2. Additive noise model: Entries of $\mathcal{M}$ on $\Omega$ are observed with additive random noise, i.e. $\forall s, y_s = \langle \mathcal{M}, \mathcal{E}^{(s)}\rangle + \eta_s$.

### 3.1 Assumptions

Collective–matrix completion is in general an ill–posed problem. However, recent literature on related tasks of compressed sensing [12, 7, 8], matrix estimation [28, 6, 5, 19, 20, 25, 18, 15], and other high dimensional estimation [26, 9, 4, 35] etc. propose tractable estimators with strong statistical guarantees for such high dimensional problems when low dimensional structural constraints are imposed on the ground truth parameters.

**Assumption 1** ($R$–dimensional joint factorization)**.** *We assume that the ground truth collective–matrix $\mathcal{M}$ has a collective–matrix rank of $R \ll N$, i.e. $\exists\{U_k \in \mathbb{R}^{n_k \times R}\}$, such that $\forall v$, $M_v = U_{r_v} U_{c_v}^\top$.* □

Analogous to matrices, $\forall \mathcal{X} \in \bar{\mathfrak{X}}$, we define the following:

$$\begin{aligned} T(\mathcal{X}) = &\text{aff}\{\mathcal{Y} \in \bar{\mathfrak{X}} : \forall\, v, \text{rowSpan}(\mathbb{Y}_{r_v}) \subseteq \text{rowSpan}(\mathbb{X}_{r_v}) \\ &\quad \textit{or } \text{rowSpan}(\mathbb{Y}_{c_v}) \subseteq \text{rowSpan}(\mathbb{X}_{c_v})\}, \end{aligned} \tag{8}$$

$$\begin{aligned} T^\perp(\mathcal{X}) = &\{\mathcal{Y} \in \bar{\mathfrak{X}} : \forall\, v, \text{rowSpan}(Y_v) \perp \text{rowSpan}(M_v) \\ &\quad \textit{and } \text{colSpan}(Y_v) \perp \text{colSpan}(M_v)\}, \end{aligned} \tag{9}$$

where we have used the entity matrix set representation in (8) (See Sec. 2.2.1). In the rest of the paper, we denote $T(\mathcal{M})$ and $T^\perp(\mathcal{M})$ simply as $T$ and $T^\perp$, respectively. Let $P_T$ and $P_{T^\perp}$ be projections onto $T$ and $T^\perp$, respectively.

**Lemma 1.** *$\forall \mathcal{X} \in \bar{\mathfrak{X}}$, $\mathcal{X} \in T^\perp$ iff $\langle \mathcal{X}, \mathcal{Y}\rangle = 0$, $\forall \mathcal{Y} \in T$.*
The lemma is proved in the supplementary material. □

As with matrix completion, in a localized observation setting, consistent recovery is infeasible if any entry in $\mathcal{M}$ is overly significant. Such cases are precluded through the following analogue of *incoherence conditions* [6, 14].

**Assumption 2** (Incoherence)**.** *We assume that $\exists$ $(\mu_0, \mu_1)$ such that the following incoherence conditions with respect to standard basis are satisfied for all $\mathcal{E}^{(v,i,j)}$:*

$$\|P_T(\mathcal{E}^{(v,i,j)})\|_F^2 \leq \frac{\mu_0 R}{m_{r_v}} + \frac{\mu_0 R}{m_{c_v}}, \tag{10}$$

$$\exists \mathcal{E}_{\mathcal{M}} \in \mathscr{E}(\mathcal{M}) \cap T, \; s.t.\; \langle \mathcal{E}^{(v,i,j)}, \mathcal{E}_{\mathcal{M}}\rangle^2 \leq \frac{\mu_1 R}{N^2}. \tag{11}$$

*Recall $\mathscr{E}(\mathcal{M})$ from* (4)*, and $m_k = \sum_{v=1}^V n_{c_v} \mathbf{1}_{(r_v = k)} + n_{r_v} \mathbf{1}_{(c_v = k)}$.*

Note that $\|P_T(\mathcal{E}^{(v,i,j)})\|_F^2$ is upper bounded by a sum of norms of projections of $m_{r_v}$ and $m_{c_v}$ dimensional standard basis (in $\mathbb{R}^{m_{r_v}}$ and $\mathbb{R}^{m_{c_v}}$, respectively) onto the $R$ dimensional latent factor space. Equation (10) ensures that no single latent dimension is overly dominant. □

Further, in Section 2.3 it was noted that in general $\bar{\mathfrak{X}} \subseteq \mathfrak{X}$, and the set of atoms spanning $\bar{\mathfrak{X}}$ defined in (1) need not be centrally symmetric. This poses subtle challenges in analyzing the consistency of collective–matrix completion. To mitigate these difficulties, we consider a restricted set of collective–matrix structures, under which $\mathfrak{X} = \bar{\mathfrak{X}}$, and $\mathscr{A}$ is centrally symmetric.

**Assumption 3** (Bipartite $\mathcal{G}$). *Recall from Section 2 that the entity–relationship structure of $\mathfrak{X}$ is represented through an undirected graph $\mathcal{G}$. We assume that $\mathcal{G}$ is bipartite, or equivalently $\mathcal{G}$ does not contain any odd length cycles.*

Using induction, it can be easily verified that Assumption 3 implies that $\mathfrak{X} = \bar{\mathfrak{X}}$, and that $\mathscr{A}$ is centrally symmetric. Under this assumption, $\|.\|_{\mathscr{A}}$ and $\|.\|_{\mathscr{A}}^*$ are norms, and $\|\mathcal{X}\|_{\mathscr{A}}^* = \frac{1}{2}\lambda_{\max}(\mathcal{B}(\mathcal{X})) \leq \frac{1}{2}\|\mathcal{B}(\mathcal{X})\|_2$. We also note that for the well–posedness of collective–matrix completion, some variation of Assumptions 1, and 2 is necessary. However, it is not clear if Assumption 3 is necessary. □

$\forall k$, we define $\Omega_k = \{(v_s, i_s, j_s) \in \Omega : r_{v_s} = k \text{ or } c_{v_s}\}$. Let $|\Omega_k|$ be the expected number of observations in $\Omega_k$.

**Assumption 4** (Sampling). *For $s \in [|\Omega|]$, independently*
*(a) sample $k_s : k_s = k$ w.p. $\frac{|\Omega_k|}{2|\Omega|}$;*
*(b) sample $i_{k_s} \sim$ uniform$([n_k])$; and*
*(c) sample $j_{k_s} \sim$ uniform$([m_k])$.*
*$(v_s, i_s, j_s)$ is the index of $(i_{k_s}, j_{k_s})$ element in $\mathbb{M}_{k_s}$.*

Given $v \in [V]$ and $(i,j) \in \mathcal{I}(v)$, and $s = 1, 2, \ldots, |\Omega|$:

$$\Pr\big((v,i,j) = \Omega_s\big) = \frac{|\Omega_{r_v}|}{2|\Omega| n_{r_v} m_{r_v}} + \frac{|\Omega_{c_v}|}{2|\Omega| n_{c_v} m_{c_v}}. \quad (12)$$

**Remarks:**

1. Note that we overload the notation for cardinality of the set. $|\Omega_k|$ in the sampling scheme is the expected cardinality of $\Omega_k$, not the true cardinality of $\Omega_k$. However, Hoeffdings's inequality can be used to show that the cardinality of $\Omega_k$ concentrates sharply around the expectation, $|\Omega_k|$.

2. *Why* $|\Omega_k|$*?*: For consistent recovery of $\mathcal{M}$, the low dimensional factors of $\mathcal{M}$, $\{U_k \in \mathbb{R}^{n_k \times R}\}$ need to be learnt. Given $k$, information on $U_k$ is entirely contained in $\mathbb{M}_k$. Thus, the optimal sample complexity for consistent recovery depends on individual $|\Omega_k|$. The assumed sampling scheme is convenient for deriving bounds in terms of $|\Omega_k|$.

### 3.2 Atomic Norm Minimization

Collective–matrix rank of $\mathcal{M} \in \bar{\mathfrak{X}}$ is given by:

$$\text{rank}(\mathcal{M}) = \min_{\{\lambda_r \geq 0\}} \sum_r \mathbf{1}_{\lambda_r \neq 0} \quad s.t. \sum_r \lambda_r \mathcal{A}_r = \mathcal{M},$$

where $\mathcal{A}_r \in \mathscr{A}$. However, minimizing the rank of a collective–matrix is intractable. We use the atomic norm (2) as a convex surrogate for the rank function and propose the following convex estimator for the noise–free model:

$$\widehat{\mathcal{M}} = \operatorname*{argmin}_{\mathcal{X} \in \bar{\mathfrak{X}}} \|\mathcal{X}\|_{\mathscr{A}} \quad \text{s.t. } P_\Omega(\mathcal{X}) = P_\Omega(\mathcal{M}). \quad (13)$$

For the additive–noise model, we suitably modify the above convex program to propose three equivalent estimators:

$$\widehat{\mathcal{M}} = \operatorname*{argmin}_{\mathcal{X} \in \bar{\mathfrak{X}}} \|\mathcal{X}\|_{\mathscr{A}} \text{ s.t. } \|P_\Omega(\mathcal{X} - \mathcal{M})\|_F^2 \leq \omega^2, \quad (14)$$

$$\widehat{\mathcal{M}} = \operatorname*{argmin}_{\mathcal{X} \in \bar{\mathfrak{X}}} \|P_\Omega(\mathcal{X} - \mathcal{M})\|_F^2 \text{ s.t. } \|\mathcal{X}\|_{\mathscr{A}} \leq \eta, \quad (15)$$

$$\widehat{\mathcal{M}} = \operatorname*{argmin}_{\mathcal{X} \in \bar{\mathfrak{X}}} \|P_\Omega(\mathcal{X} - \mathcal{M})\|_F^2 + \gamma \|\mathcal{X}\|_{\mathscr{A}}. \quad (16)$$

The estimators are theoretically equivalent in the sense that for some combination of $\omega$, $t$, and $\gamma$ we obtain the same estimate from the three convex programs. In practice, the parameters are set through cross validation, and the choice of a convex program for noisy collective–matrix completion is often made by the algorithmic considerations.

## 4 Main Results

The main result of the paper states that under the assumptions stated in Sec. 3.1, the convex program in (13), exactly recovers the ground truth collective–matrix with high probability. We then propose a scalable greedy algorithm with convergence guarantees for solving noisy collective–matrix completion using (15).

### 4.1 Consistency under Noise–Free Model

Recall: $|\Omega_k|$ is the expected cardinality of $\Omega_k = \{(v,i,j) \in \Omega : r_v = k \text{ or } c_v = k\}$, with the true cardinality concentrating sharply under the sampling scheme (Assumption 4), and $|\Omega|$ is the cardinality of $\Omega$; $n_k$ is the number of instances of type $k$, and $N = \sum_k n_k$; $R$ is the collective–matrix rank of $\mathcal{M}$; and $\mu_0$ and $\mu_1$ are the incoherence parameters (Assumption 2).

**Theorem 1.** *Assume that the following sample complexity requirements are met,*

*(i)* $\forall k,\ |\Omega_k| > c_0 \mu_0 n_k R \beta \log N \log(N\kappa_\Omega(N))$,
*(ii)* $|\Omega| > c_1 \max\{\mu_0, \mu_1\} N R \beta \log N \log(N\kappa_\Omega(N))$,
*(iii)* $\forall k,\ \frac{|\Omega_k|}{n_k m_k} \geq c\frac{|\Omega|}{N^2}$ *for some constant c,*

*where* $\kappa_\Omega(N) = \frac{3|\Omega|\sqrt{\max_k \frac{|\Omega_k|}{n_k m_k}}}{\min_k \frac{|\Omega_k|}{n_k m_k}}$, *which scales at most as $N^4$ for general $\Omega$ and as $N^2$ under the above requirements. Then, under the assumptions in Sec. 3.1, for large enough $c_0$, and $c_1$, and $\beta > 1$, and noise–free observation model, the convex program in* (13) *exactly recovers the true collective–matrix $\mathcal{M}$ with probability greater than* $1 - N^{1-\beta} - c_2 N^{1-\beta}\log(N\kappa_\Omega(N))$ *for some constant $c_2$.*

### 4.2 Algorithm

Recently, Jaggi et. al. [17] proposed a scalable approximate algorithm for solving nuclear norm regularized matrix estimation, by adapting the approximate SDP solver of Hazan [16]. We observe that using the alternate formulation of collective–matrix atomic norm stated in Proposition 2, the convex program for noisy collective–matrix completion in (15) can be cast as the following SDP:

$$\min_{Z\succcurlyeq 0,} \sum_{v=1}^{V} \|P_{\Omega_v}(M_v - P_v(Z))\|_F^2 \quad \text{s.t. } \mathrm{tr}(Z) \leq \eta, \quad (17)$$

where $\Omega_v = \{(v_s, i_s, j_s) \in \Omega : v_s = v\}$. Hazan's algorithm for solving (17) is given in Algorithm 1.

**Algorithm 1** Hazan's Algorithm for Convex Collective–Matrix Completion (17) (Hazans–CMC)

Rescale loss: $\widehat{f}_\eta(Z) = \sum_v \|P_{\Omega_v}(M_v - P_v(\eta Z))\|_F^2$
Initialize $Z^{(1)}$
**for all** $t = 1, 2 \ldots, T = \frac{4}{\epsilon}$ **do**
  Compute $u^{(t)} = \mathrm{approxEV}\big(-\nabla \widehat{f}_\eta(Z^{(t)}), \frac{1}{t^2}\big)^1$
  $\alpha_t := \frac{2}{2+t}$
  $Z^{(t+1)} = Z^{(t)} + \alpha_t u^{(t)} u^{(t)\top}$
**return** $[P_v(Z^{(T)})]_{v=1}^V$

**Lemma 2.** *Algorithm 1 returns an $\epsilon$ approximate solution to* (15) *in time* $O\big(\frac{|\Omega|}{\epsilon^2}\big)$

*Proof:* From Theorem 2 of Hazan's work [16], the proposed algorithm returns an estimate for a SDP with primal–dual error of at most $\epsilon$ in $\frac{4C_f}{\epsilon}$ iterations, where $C_f$ is a curvature constant of the loss function. For squared loss, $C_f \leq 1$ (Lemma 4 in [17]). Iteration $t$ in Algorithm 1 involves computing an $\frac{1}{t^2}$–approximate largest eigen value of a sparse matrix with $|\Omega|$ non–zero elements, which requires $O(\frac{|\Omega|}{t})$ computation using Lanczos algorithm. □

In comparison, the SVT–CMC algorithm proposed by Bouchard et. al. [2] converges faster in $O(\frac{1}{\sqrt{\epsilon}})$ iterations; however, each iteration in SVT–CMC requires computing all the non–zero eigen vectors of a $N \times N$ matrix, which does not scale well with $N$. Hazan's algorithm can be used to trade–off computation for accuracy in large datasets.

### 4.3 Discussion and Directions for Future Work

A collective–matrix $\mathcal{M}$ of collective–matrix rank $R$ lies in a lower dimensional model space spanned by the entity factors, $\{U_k \in \mathbb{R}^{n_k \times R}\}$. Given $k$, $U_k$ is estimated entirely from $P_{\Omega_k}(\mathbb{M}_k)$. Thus, an immediate lower bound on the sample complexity for well–posedness is given by $|\Omega_k| \sim O(n_k R)$. The results presented in the paper are optimal upto a poly–logarithmic factor.

A trivial estimate for collective–matrix completion is to estimate each component matrices independently. Since a joint low rank structure also imposes low rank structure on the component matrices, this is feasible if each component matrix satisfies the sample complexity requirements of standard matrix completion, i.e. $|\Omega_v| > C \max\{\mu_0, \mu_1\} R(n_{r_v} + n_{c_v}) \log(n_{r_v} + n_{c_v})$. Another, estimate from standard matrix completion can be obtained by completing each matrix $\{\mathbb{M}_k\}$ in the entity–matrix set representation independently, this requires a sample complexity of $|\Omega_k| > C \max\{\mu_0, \mu_1\} R(n_k + m_k) \log(n_k + m_k)$ for consistent recovery. In comparison to the sample complexity in Theorem 1, these results are sub–optimal as they do not completely leverage the shared structure introduced by the jointly factorizability of collective–matrices.

Finally, the collective–matrix completion problem can also be cast as standard matrix completion problem of completing an incomplete $N \times N$ symmetric matrix, in which blocks corresponding to the collective–matrix are partially observed. However, the existing theoretical results on the consistency of matrix completion algorithms require either uniform random sampling [6, 19, 18], or coherent sampling [10] of the entries of the matrix; and these results fail for blockwise random sampled matrix. Thus, our results provide a strict generalization to existing matrix completion results for the task of collective–matrix completion.

The key challenge in the analysis is to optimally leverage the shared structure. In high dimensional recovery, sample complexity depends on some complexity measure of the model space $T$. Compared to trivial extensions, $T$ defined in (8) exploits the structure to give a narrow subspace for optimal sample complexity.

As a part of future work, we would like to extend the analysis in this paper to general structures on $\mathcal{G}$ (i.e. eliminate Assumption 3). Extension of the analysis to noisy–observation models is also of interest.

## 5 Proof Sketch

Detailed proofs of lemmata are included in the supplementary material. The proof technique is analogous to the analysis for matrix completion.

Let $\widehat{\mathcal{M}} = \mathcal{M} + \Delta$ be the output of the convex program in (13). The key steps in the proof are:

1. Show that under the sample complexity requirements of Theorem 1, $\|P_T(\Delta)\|_F$ can be upper bounded by a finite multiple of $\|P_{T^\perp}(\Delta)\|_F$. ($T$ and $T^\perp$ are defined in (8)).

2. Show optimality of $\mathcal{M}$ for (13) if a *dual certificate* $\mathcal{Y}$ satisfying certain conditions exists.

3. Adapt the *golfing scheme* introduced by Gross et al. [14] to construct $\mathcal{Y}$.

We define $p(v,i,j) = \frac{|\Omega_{r_v}|}{2n_{r_v}m_{r_v}} + \frac{|\Omega_{c_v}|}{2n_{c_v}m_{c_v}}$, and note that for $s = 1,2,\ldots,|\Omega|$, $\Pr((v,i,j) = \Omega_s) = \frac{p(v,i,j)}{|\Omega|}$. We also define the following operators for $s = 1,2,\ldots,|\Omega|$:

$$\mathcal{R}_s : \mathcal{X} \to \frac{1}{p(v_s,i_s,j_s)}\langle\mathcal{X},\mathcal{E}^{(s)}\rangle\,\mathcal{E}^{(s)}, \text{ and} \tag{18}$$

$$\mathcal{R}_\Omega : \mathcal{X} \to \textstyle\sum_{s=1}^{|\Omega|}\mathcal{R}_s(\mathcal{X}) \text{ with } E[\mathcal{R}_\Omega] = \mathcal{I}, \tag{19}$$

where $\mathcal{I}$ is the identity operator, and $\mathcal{E}^{(s)} = \mathcal{E}^{(v_s,i_s,j_s)}$

**Lemma 3.** *Let* $\forall\ k,\ |\Omega_k| \geq c_0\mu_0 n_k R\beta\log N$ *for a large constant enough* $c_0$*. Then, under the assumptions in Sec. 3.1, the following holds w. p. greater than* $1 - N^{1-\beta}$*,*

$$\|P_T\mathcal{R}_\Omega P_T - P_T\|_{op} \leq \frac{1}{2}.$$

Proof in the supplementary material. □

Let $M_\Omega(v,i,j)$ denote the multiplicity of $(v,i,j)$ in $\Omega$, i.e. $M_\Omega(v,i,j) = \sum_s \mathbf{1}_{(v,i,j)=(v_s,i_s,j_s)}$; we have $M_\Omega(v,i,j) \leq |\Omega|$. Also, note that $\min_k \frac{|\Omega_k|}{n_k m_k} \leq p(v,i,j) \leq \max_k \frac{|\Omega_k|}{n_k m_k}$. Thus, for all $\mathcal{X}$,

$$\begin{aligned}\|\mathcal{R}_\Omega(\mathcal{X})\|_F &= \Big\|\sum_{\substack{v\in[V],\\(i,j)\in\mathcal{I}(v)}}\frac{M_\Omega(v,i,j)}{p(v,i,j)}\langle\mathcal{X},\mathcal{E}^{(v,i,j)}\rangle\mathcal{E}^{(v,i,j)}\Big\|_F\\ &\leq \frac{|\Omega|}{\min_k\frac{|\Omega_k|}{n_k m_k}}\|\mathcal{X}\|_F,\end{aligned} \tag{20}$$

Further, using Lemma 3 we have the following w.h.p,

$$\begin{aligned}\|\mathcal{R}_\Omega P_T(\Delta)\|_F^2 &\geq \frac{1}{\max_k\frac{|\Omega_k|}{n_k m_k}}\langle\mathcal{R}_\Omega P_T(\Delta), P_T(\Delta)\rangle\\ &= \frac{1}{\max_k\frac{|\Omega_k|}{n_k m_k}}\langle P_T\mathcal{R}_\Omega P_T(\Delta), P_T(\Delta)\rangle\\ &\geq \frac{1}{2\max_k\frac{|\Omega_k|}{n_k m_k}}\|P_T(\Delta)\|_F^2.\end{aligned} \tag{21}$$

Combining (20) and (21), along with $0 = \|\mathcal{R}_\Omega(\Delta)\|_F \geq \|\mathcal{R}_\Omega P_T(\Delta)\|_F - \|\mathcal{R}_\Omega P_{T^\perp}(\Delta)\|_F$, we have

$$\|P_T(\Delta)\|_F \leq \frac{1}{2}\kappa_\Omega(N)\|P_{T^\perp}(\Delta)\|_F, \tag{22}$$

where $\kappa_\Omega(N) = \frac{3|\Omega|\sqrt{\max_k|\Omega_k|/n_k m_k}}{\min_k|\Omega_k|/n_k m_k}$.

### 5.1 Optimality of $\mathcal{M}$

**Lemma 4.** *Under the assumptions in Sec. 3.1, let* $\forall k$, $|\Omega_k| \geq c_0\mu_0 n_k R\beta\log N$ *for a sufficiently large constant* $c_0$*. If there exists a dual certificate* $\mathcal{Y}$ *satisfying the following conditions, then* $\mathcal{M}$ *is the unique minimizer to* (13) *w.p. greater than* $1 - N^{1-\beta}$*:*

1. $\|P_T(Y) - \mathcal{E}_\mathcal{M}\|_F \leq \frac{1}{\kappa_\Omega(N)}$*, and*
2. $\|P_{T^\perp}(Y)\|_\mathscr{A}^* \leq 1/2$*,*

*where recall* $\mathcal{E}_\mathcal{M}$ *from Assumption 2.*

Proof is in the supplementary material. □

### 5.2 Constructing Dual Certificate

The proof is completed by constructing a dual certificate satisfying the conditions in Lemma 4. We begin by partitioning each $\Omega$ into $p = \mathcal{O}(\log(N\kappa_\Omega(N)))$ partitions denoted by $\Omega^{(j)}$, for $j = 1,2,\ldots,p$, such that for all $j$:

(a) $\forall k$, $|\Omega_k^{(j)}| > c_0\mu_0\beta R n_k\log N$ and $\frac{|\Omega_k^{(j)}|}{n_k m_k} \leq c\frac{|\Omega^{(j)}|}{N^2}$,
(b) $|\Omega^{(j)}| > c_2\max\{\mu_0,\mu_1\}\beta RN\log N$,
where $\Omega_k^{(j)} = \{(v,i,j) \in \Omega^{(j)} : r_v = k \text{ or } c_v = k\}$.

Define $\mathcal{W}_0 = \mathcal{E}_\mathcal{M}$ where $\mathcal{E}_\mathcal{M}$ is the sign matrix from Assumption 2. We define a process for $j = 1,2,\ldots$ s.t. :

$$\begin{aligned}&\mathcal{Y}_j = \textstyle\sum_{j'=1}^{j}\mathcal{R}_{\Omega^{(j')}}\mathcal{W}_{j'-1} = \mathcal{R}_{\Omega^{(j)}}\mathcal{W}_{j-1} + \mathcal{Y}_{j-1},\\ &\mathcal{W}_j = \mathcal{E}_\mathcal{M} - P_T(\mathcal{Y}_j).\end{aligned} \tag{23}$$

Note that $\forall\ j$, $P_\Omega(\mathcal{Y}_j) = \mathcal{Y}_j$, and $P_T(\mathcal{W}_j) = \mathcal{W}_j$. We show that $\mathcal{Y}_p$ for $p = \mathcal{O}(\log(N\kappa_\Omega(N)))$ satisfies the first condition required in Lemma 4. The proof for second condition follows directly from the analogous proof for standard matrix completion by Recht [27] and is provided in the supplementary material.

It is easy to verify that $\frac{1}{2}\mathcal{E}^{(v,i,j)} \in \mathscr{A}$ for all $(v,i,j)$, and by Assumption 3, $-\frac{1}{2}\mathcal{E}^{(v,i,j)} \in \mathscr{A}$. Thus, $\forall\mathcal{X} \in \bar{\mathfrak{X}}$,

$$\|\mathcal{X}\|_\mathscr{A}^* \geq \frac{1}{2}\max_{\substack{v\in[V]\\(i,j)\in\mathcal{I}(v)}}|\langle\mathcal{X},\mathcal{E}^{(v,i,j)}\rangle| \geq \frac{1}{2N}\|\mathcal{X}\|_F.$$

Also, $1 = \|\mathcal{E}_\mathcal{M}\|_\mathscr{A}^* \geq \frac{1}{2N}\|\mathcal{E}_\mathcal{M}\|_F$, and $P_T(\mathcal{Y}_p) - \mathcal{E}_\mathcal{M} = \mathcal{W}_p$. Using the above inequalities, we have:

$$\begin{aligned}\|P_T(\mathcal{Y}_p) - \mathcal{E}_\mathcal{M}\|_F &= \|\mathcal{W}_{p-1} - P_T\mathcal{R}_{\Omega^{(p)}}\mathcal{W}_{p-1}\|_F\\ &\overset{(a)}{\leq} \frac{1}{2}\|\mathcal{W}_{p-1}\|_F \leq \frac{1}{2^p}\|\mathcal{E}_\mathcal{M}\|_F \overset{(b)}{<} \frac{1}{\kappa_\Omega(N)}\end{aligned} \tag{24}$$

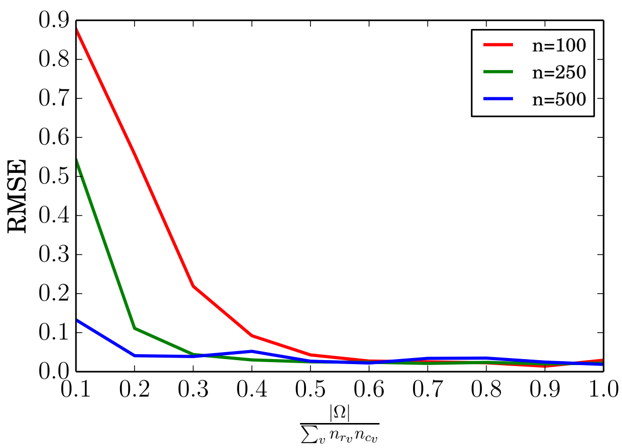
where $(a)$ follows from Lemma 3, and $(b)$ follows for large enough $c_1$ s.t. $p = c_1\log(N\kappa_\Omega(N))$. Note that we use union bound to bound the probability of failure in $\mathcal{O}(\log(N\kappa_\Omega(N)))$ partitions.

## 6 Experiments

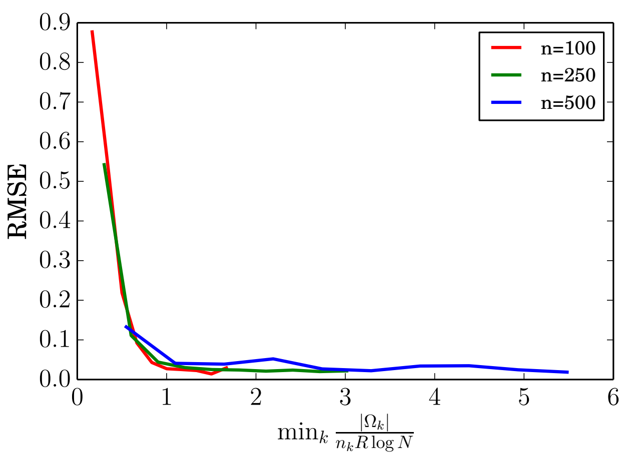
The simulated experiments are intended to corroborate our theoretical results in Sec. 4. Experiments with commercial news recommendation dataset is provided to show the efficacy collective matrix completion over standard matrix completion for cold–start scenarios.

### 6.1 Simulated Experiments

We create low–rank ground truth collective–matrices with $K = 4$, $V = 3$, where view 1 is a relation between entity types 1 and 2, view 2 is a relation between entity types 1 and 3, and view 3 is a relation between entity types 2 and 4 respectively. For simplicity we assumed a common $n_k = n$. We create collective matrices with $n \in \{100, 250, 500\}$ and set the rank to $R = 2\log n$. The matrices are partially observed with the fraction of observed entries, $\frac{|\Omega|}{\sum_v n_{r_v} n_{c_v}}$ varying as $[0.1, 0.2, \ldots, 1]$. We plot the convergence of the errors against the unnormalized fraction of observations, $\frac{|\Omega|}{\sum_v n_{r_v} n_{c_v}}$ in Fig. 2a, and against the normalized sample complexity provided by the theoretical analysis, $\min_k \frac{|\Omega_k|}{n_k R \log N}$ in Fig. 2b. It can be seen from the plots that the error uniformly decays with increasing normalized sample size, indeed $|\Omega_k| > 1.5 n_k R \log N,\ \forall k$ samples suffice for the errors to decay to a very small value. The aligning of the curves (for different $n$) given the normalized sample size corroborates the theoretical sample complexity requirements.




(a) RMSE vs unnormalized sample size




(b) RMSE vs normalized sample size

Figure 2: Convergence of error measured against normalized and unnormalized sample size

### 6.2 Experiments with Commercial News Recommendation Dataset

We work with two datasets from a commercial news recommendation engine. The entities include users, news articles, and news–categories. The datasets consists of two views (a) user–article click information in a 3hr time window, (b) a dense and complete user–category preference obtained by an aggregation of the categories clicked by users.

The first dataset "News–Cold–Start", consists of $\sim 180K$ users, $\sim$ 750 articles, and 34 categories. In this dataset, $\sim$ 25000 users have only one click. Randomly chosen negative samples were added to give dataset of $\sim$ 1.25 million user–article ratings, and $\sim$ 1.4 million user–category annotations. The dataset was split in 70 : 10 : 20 proportion as training, validation and test set. The 20% of the test dataset contains cold start users with no rating information. In the second dataset "News–No–Cold–Start", we remove the cold start users in the test dataset. This leads to a much smaller datasets consisting of $\sim$ 6500 users, $\sim$ 750 articles and 34 categories, with $\sim 150K$ user–article ratings (including the randomly chosen negatives) and $\sim 50K$ user–category ratings. The negatives in each dataset were sampled independently in each cross–validation iteration to remove bias.

Mean absolute error (MAE) on the test dataset obtained from the proposed Hazans algorithm for Collective–Matrix Completion (CMF–Hazans) and Standard Matrix Factorization (SMF) are reported in Table 1.

| Method | News–Cold–Start | News–No-Cold-Start |
|---|---|---|
| CMF–Hazans | $0.2741 \pm 0.0002$ | $0.2156 \pm 0.0014$ |
| SMF | $0.2905 \pm 0.0007$ | $0.2149 \pm 0.0008$ |

Table 1: MAE of the predictors on the two news recommendation datasets

It is observed that collective matrix factorization does not add much value for warm–start cases as the ratings give accurate prediction. On the other hand, for test dataset consisting on both warm–start and cold–start test cases, the proposed joint estimation potentially leverages the information in the user–category affinities and shows statistically significant improvement.

**Acknowledgement:** Suriya Gunasekar acknowledges funding from NSF grant IIS–1116656.

## References

[1] D. Agarwal, B. C. Chen, and B. Long. Localized factor models for multi-context recommendation. In *Proceedings of KDD*, 2011.

[2] G. Bouchard, S. Guo, and D. Yin. Convex collective matrix factorization. In *AISTATS*, 2013.

[3] J. F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.

[4] E. Candes. Mathematics of sparsity (and a few other things). Plenary Lectures, International Congress of Mathematicians, 2014.

[5] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 2010.

[6] E. J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 2009.

[7] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 2006.

[8] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 2006.

[9] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 2012.

[10] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. In *ICML*, 2014.

[11] M. A. Davenport, Y. Plan, E. Berg, and M. Wootters. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.

[12] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 2006.

[13] D. Dueck, Q. D. Morris, and B. J. Frey. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, 2005.

[14] D Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 2011.

[15] S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. In *ICML*, 2014.

[16] E. Hazan. Sparse approximate solutions to semidefinite programs. In *LATIN 2008: Theoretical Informatics*. Springer, 2008.

[17] M. Jaggi and M. Sulovsk. A simple algorithm for nuclear norm regularized problems. In *ICML*, 2010.

[18] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.

[19] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 2010.

[20] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *JMLR*, 2010.

[21] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 2009.

[22] C. Lippert, S. H. Weber, Y. Huang, V. Tresp, M. Schubert, and H. P. Kriegel. Relation prediction in multi-relational domains using matrix factorization. In *NIPS 2008 Workshop: Structured Input-Structured Output*. Citeseer, 2008.

[23] B. Long, Z. M. Zhang, X. Wu, and P. S. Yu. Spectral clustering for multi-type relational data. In *ICML*. ACM, 2006.

[24] A. Mnih and R. Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, 2007.

[25] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *JMLR*, 2012.

[26] S. Negahban, B. Yu, M. J. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. In *NIPS*, 2009.

[27] B. Recht. A simpler approach to matrix completion. *JMLR*, 2011.

[28] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 2010.

[29] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *ICML*. ACM, 2008.

[30] A. P. Singh. Efficient matrix models for relational learning. Technical report, DTIC Document, 2009.

[31] A. P. Singh and G. Gordon. A bayesian matrix factorization model for relational data. *arXiv preprint arXiv:1203.3517*, 2012.

[32] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of KDD*. ACM, 2008.

[33] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *NIPS*, 2004.

[34] K. C Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 2010.

[35] R. Vershynin. Estimation in high dimensions: a geometric perspective. *ArXiv e-prints*, May 2014.

[36] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of ACM SIGIR conference on R&D in informaion retrieval*, 2003.

[37] K. Y. Yilmaz, A. T. Cemgil, and U. Simsekli. Generalised coupled tensor factorisation. In *NIPS*, 2011.

[38] Y. Zhang, B. Cao, and D. Y. Yeung. Multi-domain collaborative filtering. *arXiv preprint arXiv:1203.3535*, 2012.

[39] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Algorithmic Aspects in Information and Management, LNCS 5034*, 2008.