

ROBUST ON-LINE MATRIX COMPLETION ON GRAPHS

Symeon Chouvardas, Mohammed Amin Abdullah, Lucas Claude, Moez Draief

Mathematical and Algorithmic Sciences Lab,
Huawei France R&D,
Paris, France.

ABSTRACT

We study online robust matrix completion on graphs. At each iteration a vector with some entries missing is revealed and our goal is to reconstruct it by identifying the underlying low-dimensional subspace from which the vectors are drawn. We assume there is an underlying graph structure to the data, that is, the components of each vector correspond to nodes of a certain (known) graph, and their values are related accordingly. We give algorithms that exploit the graph to reconstruct the incomplete data, even in the presence of outlier noise. The theoretical properties of the algorithms are studied and numerical experiments using both synthetic and real world datasets verify the improved performance of the proposed technique compared to other state of the art algorithms.

1. INTRODUCTION

The science of data acquisition, processing and inference has been boosted in recent years due to both the abundance of data and the economic benefits associated with understanding it. Modern technologies have produced a vast array of data-generating devices (e.g., smart phones, cameras, sensors) and processes (e.g., web searches, surveys, social media interaction). Frequently, it is known or conjectured that there is an underlying *structure* to the data, and further, that this structure reflects some simpler underlying process. To make this more concrete, consider the notion of *sparsity* in a movie-rating database such as used by Netflix. If the rows of the ratings matrix represent movies and the columns people, then it is reasonable to suppose, and indeed has been found to be the case ([1, 2]), that the ratings matrix is of low rank. This would reflect that ratings vectors really “live” in a small subspace of the ambient space they are generated in. Consequently, understanding this subspace better allows the exploitation of the data, for example, through more tightly focused marketing. On the other hand, the matrix will be incomplete, since any given user will rate only a very small subset of all the movies available. Thus, one would desire *completion* of the matrix: a low-rank matrix with sufficiently many observed entries can be *exactly* reconstructed, and in the last decade, this has been a very active area of research in the signal processing and

machine learning communities. However, low-dimensional subspaces are not the form of structure; indeed, there can also be an underlying *graphical* structure that represents connections/relations between entities. For example, one may generate a graph where movies are nodes and and two movies are linked if they both star a famous actor. Such a structure is sometimes easy to discover, and it immediately begs the question of how (if at all) it can be used to help fill in the missing entries of the matrix.

The above is the subject of this paper: we investigate how graph structure can aid the reconstruction of a low rank matrix with missing entries, and specifically, in the *on-line* setting. In this setting each column of the matrix (with some entries missing) is presented one at a time, and the algorithm must make the best estimation using only what has been presented so far. This is in contrast to the batch setting where the entire non-complete matrix is available to process. The motivations for the online setting are at least two-fold. Firstly, it more realistically reflects many situations. In the Netflix example, one may have a data stream of ratings. Secondly with the massive amount of data being generated, computational and memory limitations present very real challenges to algorithms which operate in batch mode; indeed, it maybe impractical to even hold the entire matrix in memory, let alone perform complex operations on it.

1.1. RELATED WORK

The problem of matrix completion is a well studied one and several solutions have been proposed during the past years, see for example [5, 6, 4]. The online setup has its roots on the so-called subspace tracking problem, e.g., [21], in which the columns of a matrix are revealed sequentially one per iteration step and the goal is the identification of the underlying subspace. Extensions of these works, which deal with the presence of missing entries and/or outliers have been studied in [15, 9, 11, 7, 10, 8]. The batch version of the matrix completion on graphs problem was originally presented in [12] and extended to its robust version, which deals with the presence of outliers, in [17].

1.2. OUR CONTRIBUTION

In this work, we extend the idea presented in [12] and we propose a robust online algorithm for matrix completion exploiting graph information. Here, we propose an online solution, i.e., the columns of the matrix appear and are processed sequentially, one per iteration step. To that direction, at each iteration step we define a proper cost function and we minimize it to produce the updated estimates. Furthermore, we study the case where there is outlier noise, which corrupts a small subset of the observed vector. We propose a robust solution, which estimates the outlier noise and cleans the data before updating the quantities of interest. Our work has two notable differences compared to other online matrix completion works. First, all other works do not exploit any graph information. Second, due to the absence of the graph information, the problem they solve decouples over the rows of the unknown subspace, which is not the case here. This introduces a new difficulty.

Notation: Lowercase and uppercase boldfaced letters stand for vectors and matrices respectively. The stage of discussion will be $\mathbb{R}^{m \times r}$, where the symbol \mathbb{R} stands for the set of real numbers. Furthermore, $\|\mathbf{A}\|$ is the operator norm and $\|\mathbf{A}\|_F$ the Frobenius norm of matrix \mathbf{A} . $\|\mathbf{x}\|$, $\|\mathbf{x}\|_1$ denote the Euclidean and the ℓ_1 norms of vector \mathbf{x} , respectively. The symbol \otimes stands for the Kronecker product. Finally, \mathbf{I}_{mr} is the $mr \times mr$ identity matrix and $\mathbf{O}_{a \times b}$ is the zero matrix of dimension $a \times b$.

2. MATRIX COMPLETION ON GRAPHS

In this paper, we are concerned with the problem of matrix completion (MC) on graphs. The original task of MC, e.g., [5, 19], is the recovery of a data matrix from a sample of its entries. Formally, given a matrix \mathbf{X} of dimension $m \times n$ we have access to $k \ll m \cdot n$ entries and the goal is the prediction of the rest unobserved ones. It has been shown that under certain conditions this can be achieved [6, 5]. Intuitively, MC builds upon the observation that if a certain matrix is structured, in the sense that it is of low rank or of approximate low rank, then it can be recovered exactly, under some mild assumptions regarding the positions of the observed entries. The problem can be summarized as follows: Compute a matrix, \mathbf{A} , which will be of low rank and equal to the observation matrix \mathbf{X} in the set of observed entries, say Ω ; that is $A_{ij} = X_{ij}, \forall i, j \in \Omega$, where X_{ij} , A_{ij} is the i, j -th entry of \mathbf{X} and \mathbf{A} respectively. A way to do so is to solve the following problem:

$$\begin{aligned} \min_{\mathbf{A}} \text{rank}(\mathbf{A}) \\ \text{s.t. } A_{ij} = X_{ij}, \forall i, j \in \Omega. \end{aligned}$$

The rank minimization problem described previously cannot be solved efficiently, since it is NP-hard [5]. However, it has

been shown, [6], that this problem can be relaxed and solved efficiently via convex optimization. The relaxation of the initial problem can be written as follows:

$$\min_{\mathbf{A}} \|\mathbf{A}\|_* \quad (1)$$

$$\text{s.t. } A_{ij} = X_{ij}, \forall i, j \in \Omega, \quad (2)$$

where $\|\mathbf{A}\|_*$ denotes the nuclear norm of the matrix \mathbf{A} with definition: $\|\mathbf{A}\|_* = \sum_{k=1}^{\min(m,n)} \sigma_k(\mathbf{A})$, with $\sigma_k(\cdot)$ being the k -th larger singular value. This model can be further generalized so that to take into account the presence of noise. In that case the equality constraint can be relaxed and the optimization problem becomes:

$$\min_{\mathbf{A}} \lambda_1 \|\mathbf{A}\|_* + \frac{1}{2} \|P_{\Omega}(\mathbf{A} - \mathbf{X})\|_F^2, \quad (3)$$

where P_{Ω} is an operator which sets the entries of its matrix argument not in Ω to zero, and keeps the rest unchanged and $\lambda_1 > 0$ is a regularization term.

Low rank implies the linear dependence of rows/columns of \mathbf{X} . However, this dependence is unstructured. In many situations, the rows and/or columns of matrix \mathbf{X} possess additional structure that can be incorporated into the completion problem in the form of a regularization. In this paper, we assume that the rows of \mathbf{X} are given on vertices of graphs. More formally, let us be given an undirected graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ on the rows with vertices $\mathbf{V} = \{1, \dots, m\}$, edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ and non-negative weights on the edges represented by the symmetric $m \times m$ matrix \mathbf{W} . If there is an edge between i, j , then $W_{ij} = W_{ji} = 0$, and we shall assume the graph has no parallel edges or loops. The latter means that the diagonal elements of \mathbf{W} are zero.

The weights capture a strength of association between the row elements. We embed the graph structure into the matrix completion problem using the *Laplacian*. This is the positive semidefinite (PSD) matrix \mathbf{L} defined as $\mathbf{D} - \mathbf{W}$ where \mathbf{D} is the diagonal matrix such that $D_{ii} = \sum_{j=1}^m W_{ij}$.

The problem of matrix completion over graphs can be formulated as follows, [12]:

$$\min_{\mathbf{A}} \lambda_1 \|\mathbf{A}\|_* + \frac{1}{2} \|P_{\Omega}(\mathbf{A} - \mathbf{X})\|_F^2 + \lambda_2 \text{tr}(\mathbf{A}^T \mathbf{L} \mathbf{A}), \quad (4)$$

where $\text{tr}(\mathbf{A}^T \mathbf{L} \mathbf{A})$ is a graph smoothing regularization constraint and $\lambda_2 > 0$ is the regularization parameter associated with it. In fact it holds that

$$\sum_{i,j} W_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|^2 = \text{tr}(\mathbf{A}^T \mathbf{L} \mathbf{A}),$$

with \mathbf{a}_i being the i -th row of the matrix \mathbf{A} . In words we demand that the rows corresponding to neighboring nodes to be ‘‘close’’ (in some sense) to each other. This problem, which was originally been proposed in [12] has been generalized in [17] to tackle scenarios where outliers are present.

Before we turn our focus to the online problem, we present some useful properties of the nuclear norm. The nuclear norm of a matrix \mathbf{M} of rank r can be written as [16]

$$\|\mathbf{M}\|_* = \min_{\mathbf{U} \in \mathbb{R}^{m \times r}, \mathbf{R} \in \mathbb{R}^{r \times n}} \{\|\mathbf{U}\|_F^2 + \|\mathbf{R}\|_F^2\} \quad s.t. \mathbf{M} = \mathbf{UR}. \quad (5)$$

Note that the number of columns of the matrix \mathbf{U} , denoted by r , is also a variable. The problem of estimating r goes beyond the scope of this paper *and from now on we will consider that r will be equal to the rank of \mathbf{X} and will be known*. This assumption was also made in other papers (e.g., [9, 15]) dealing with online matrix completion. Taking this into account and substituting (5) into (4) leads us to:

$$\min_{\mathbf{U}, \mathbf{R}: \mathbf{UR} \in \mathbb{R}^{m \times n}} \lambda_1 (\|\mathbf{U}\|_F^2 + \|\mathbf{R}\|_F^2) + \frac{1}{2} \|P_{\Omega}(\mathbf{UR} - \mathbf{X})\|_F^2 + \lambda_2 \text{tr}(\mathbf{R}^T \mathbf{U}^T \mathbf{LUR}). \quad (6)$$

2.1. ONLINE MATRIX COMPLETION ON GRAPHS

The above deals with the batch problem, i.e., the one in which all the measurements are available a priori and are used in the computations as a whole. However, in many applications, having access to all the data may be impractical and/or infeasible. More specifically, in big data applications, the data might not be able to be stored and the algorithm needs to retrieve them from slow memory devices or to access them over networks. Moreover, in batch operation the unknown subspace has to be re-computed from scratch whenever a new datum becomes available. Our goal here is to present an online solution to the matrix completion over graphs problem.

In our context we consider that at each step, i.e., t , a single column of the matrix \mathbf{X} , say $\mathbf{x}_t \in \mathbb{R}^m$, which also has missing entries, becomes available.

Per (6), each observation vector $\mathbf{x}_t \in \mathbb{R}^m$, $t = 1, \dots, n$ is given by

$$\mathbf{x}_t = P_{\Omega_t}(\mathbf{U}\mathbf{r}_t + \mathbf{v}_t), \quad (7)$$

where \mathbf{U} is an $m \times r$ matrix $\mathbf{r}_t \in \mathbb{R}^r$ and $\mathbf{v}_t \in \mathbb{R}^m$ is the noise process. This formula will be our starting point for the derivation of the online algorithm. Following the exponentially weighted least squares rationale, the online formulation of (6) can be cast as follows:

$$\min_{\mathbf{U}, \{\mathbf{r}_\tau\}} \sum_{\tau=1}^t \left(\frac{1}{2} \|P_{\Omega_\tau}(\mathbf{x}_\tau - \mathbf{U}\mathbf{r}_\tau)\|_2^2 + \frac{\lambda_1}{2} \|\mathbf{r}_\tau\|_2^2 + \frac{\lambda_2}{2} (\mathbf{r}_\tau^T \mathbf{U}^T \mathbf{LUR}_\tau) \right) + \frac{\lambda_1}{2} \|\mathbf{U}\|_F^2, \quad (8)$$

We attempt to solve the above iteratively. In each iteration t , we maintain the last estimate \mathbf{U}_{t-1} of the subspace. We compute an optimal \mathbf{r}_t assuming \mathbf{U}_{t-1} . We then use $\mathbf{r}_1, \dots, \mathbf{r}_t$ to generate \mathbf{U}_t . This two-step procedure is typical in online matrix factorization problems, see for example [14, 18]. Next we derive the minimization for the first step of the algorithm.

To that end, we keep only the terms which depend on \mathbf{r} and we obtain:

$$\min_{\mathbf{r}} \frac{1}{2} \|P_{\Omega_t}(\mathbf{x}_t - \mathbf{U}_{t-1}\mathbf{r})\|_2^2 + \frac{\lambda_1}{2} \|\mathbf{r}\|_2^2 + \frac{\lambda_2}{2} (\mathbf{r}^T \mathbf{U}_{t-1}^T \mathbf{LUR}_{t-1}). \quad (9)$$

Computing the derivative with respect to \mathbf{r} and setting it equal to $\mathbf{0}_r$, we obtain the minimizer of (9) given by:

$$\mathbf{r}_t = \mathbf{A}_t^{-1} \mathbf{U}_{t-1}^T P_{\Omega_t}(\mathbf{x}_t), \quad (10)$$

where

$$\mathbf{A}_t = \lambda_1 \mathbf{I}_r + \mathbf{U}_{t-1}^T (\mathbf{\Omega}_t + \lambda_2 \mathbf{L}) \mathbf{U}_{t-1}$$

and $\mathbf{\Omega}_t \in \mathbb{R}^{m \times m}$ is the diagonal matrix associated with the set Ω_t having in its diagonal 1 if the respective entry is observed and 0 if it is unobserved. Note that λ_1 being positive implies \mathbf{A}_t is positive definite and therefore invertible.

The next step is the minimization with respect to \mathbf{U} . Computing the gradient of (8) with respect to \mathbf{U} , and equating it with the zero matrix, we obtain:

$$\lambda_1 \mathbf{U} + \lambda_2 \mathbf{LUR}_t + \sum_{\tau=1}^t \mathbf{\Omega}_\tau \mathbf{U} \mathbf{r}_\tau \mathbf{r}_\tau^T = \mathbf{P}_t \quad (11)$$

where

$$\mathbf{R}_t = \sum_{\tau=1}^t \mathbf{r}_\tau \mathbf{r}_\tau^T \quad (12)$$

$$\mathbf{P}_t = \sum_{\tau=1}^t \mathbf{\Omega}_\tau \mathbf{x}_\tau \mathbf{r}_\tau^T. \quad (13)$$

A drawback of this formulation is that the matrix \mathbf{U} is coupled with $\mathbf{\Omega}_\tau$ and \mathbf{r}_τ so solving directly (11) with respect to \mathbf{U} becomes difficult or infeasible. However, we can use properties of Kronecker products and bypass this difficulty. First, we vectorize (11) and we obtain:

$$\text{vec} \left\{ \sum_{\tau=1}^t \mathbf{\Omega}_\tau \mathbf{U} \mathbf{r}_\tau \mathbf{r}_\tau^T \right\} + \lambda_1 \text{vec} \{ \mathbf{U} \} + \lambda_2 \text{vec} \{ \mathbf{LUR}_t \} = \text{vec} \{ \mathbf{P}_t \},$$

where $\text{vec}\{\cdot\}$ is the vectorization operator that vectorizes a matrix by stacking the columns so as to form a supervector. The first term of the left hand side can be equivalently written [13]:

$$\text{vec} \left\{ \sum_{\tau=1}^t \mathbf{\Omega}_\tau \mathbf{U} \mathbf{r}_\tau \mathbf{r}_\tau^T \right\} = \left(\sum_{\tau=1}^t \mathbf{r}_\tau \mathbf{r}_\tau^T \otimes \mathbf{\Omega}_\tau \right) \mathbf{u}, \quad (14)$$

where $\mathbf{u} := \text{vec}\{\mathbf{U}\}$. The third term of the left hand side of (11) can be written as:

$$\text{vec} \{ \mathbf{LUR}_t \} = (\mathbf{R}_t \otimes \mathbf{L}) \mathbf{u} \quad (15)$$

So, the solution of (11) (in a vectorized form) is given by:

$$\mathbf{u} = \left(\sum_{\tau=1}^t \mathbf{r}_\tau \mathbf{r}_\tau^T \otimes \boldsymbol{\Omega}_\tau + \lambda_1 \mathbf{I}_{mr} + \mathbf{R}_t \otimes \mathbf{L} \right)^{-1} \mathbf{p}_t, \quad (16)$$

where $\mathbf{p}_t = \text{vec}\{\mathbf{P}_t\}$. The steps of the algorithm are summarized as Algorithm 1

Algorithm 1: Online Matrix Completion on Graphs

Input: $\lambda_1, \lambda_2, \mathbf{L}$

Output: Computed Subspaces \mathbf{U}_t and vectors \mathbf{r}_t

- 1 **Initialize:** \mathbf{U}_0
 - 2 **for** $t = 1, 2, \dots$ **do**
 - 3 Compute \mathbf{r}_t by solving (10)
 - 4 Update $\mathbf{R}_t, \mathbf{P}_t$ by (12) (13) respectively
 - 5 Compute \mathbf{U}_t by solving (16) and devectorizing
-

A crucial point regarding the computational aspects of Algorithm 1 is that the memory and time complexities do not grow with time: The update in line 4 only needs to use the previous values $\mathbf{R}_{t-1}, \mathbf{P}_{t-1}$, and the new quantities $\mathbf{r}_t, \mathbf{x}_t, \boldsymbol{\Omega}_t$. Hence, only a bounded amount of memory (and computation time) is required.

Full Observability

In the special case where the entries of each \mathbf{x}_t are fully observable (and so $\boldsymbol{\Omega}_t$ becomes the identity matrix), we can take a more direct approach. This is the *subspace tracking problem*. Since, $\lambda_2 > 0$ and \mathbf{L} is positive semidefinite, from (11) we can write

$$\lambda_1 (\mathbf{I}_m + \lambda_2 \mathbf{L})^{-1} \mathbf{U} + \mathbf{U} \mathbf{R}_t = (\mathbf{I}_m + \lambda_2 \mathbf{L})^{-1} \mathbf{P}_t.$$

This belongs to the family of the so-called *Sylvester's equations* (see, e.g., [22]), and can be solved efficiently. The general form of Sylvester's equation is:

$$\mathbf{A} \mathbf{X} + \mathbf{X} \mathbf{B} = \mathbf{C},$$

and has a unique solution when there are no common eigenvalues of \mathbf{A} and $-\mathbf{B}$. For our case, this is assured because \mathbf{R}_t is PSD.

3. ROBUSTIFICATION

A drawback of the matrix completion techniques, which rely on the Frobenious norm minimization is that they are sensitive to heavy tailed noise. In the batch scenario, Robust PCA (RPCA) originally proposed in [4] overcomes this limitation. In particular, the model generating the matrix comprising missing entries is the following:

$$\mathbf{M} = \mathbf{A} + \mathbf{S}, \quad (17)$$

where \mathbf{A} is a low rank matrix and \mathbf{S} is a sparse matrix, the entries of which have arbitrarily large amplitude; the latter matrix models the outlier noise. The optimization problem for the matrix completion takes the following form:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{S}} \|\mathbf{A}\|_* + \lambda_s \|\mathbf{S}\|_1, \\ \text{s.t. } \mathbf{M} = \mathbf{A} + \mathbf{S}, \end{aligned}$$

where $\|\cdot\|_1$ promotes sparsity and has the following definition $\|\mathbf{S}\|_1 = \sum_{i,j} |S_{i,j}|$, i.e., the sum of absolute values of the entries of \mathbf{S} .

The aforementioned problem has been also extended to the online scenario, e.g., [9, 15]. This will be our starting point for deriving the online robust MC algorithm on graphs. To be more specific, the model generating the columns of the matrix becomes:

$$\mathbf{x}_t = P_{\Omega_t} (\mathbf{U} \mathbf{r}_t + \mathbf{s}_t + \mathbf{v}_t),$$

where \mathbf{s}_t stands for the outlier vector. For this, we assume that the ℓ_0 (pseudo) norm, which counts the number of non-zero coefficients, is bounded and smaller than m , i.e., $\|\mathbf{s}_t\|_0 \leq m' < m$.¹ Furthermore, similarly to what we have done before, we assume that there exists an underlying graph structure, which is assumed to be known. The problem we want to solve becomes:

$$\begin{aligned} \min_{\mathbf{U}, \{\mathbf{r}_\tau\}, \{\mathbf{s}_\tau\}} \frac{\lambda_1}{2} \|\mathbf{U}\|_F^2 + \sum_{\tau=1}^t \left(\frac{1}{2} \|P_{\Omega_\tau} (\mathbf{x}_\tau - \mathbf{U} \mathbf{r}_\tau - \mathbf{s}_\tau)\|_2^2 \right. \\ \left. + \frac{\lambda_1}{2} \|\mathbf{r}_\tau\|_2^2 + \frac{\lambda_2}{2} (\mathbf{r}_\tau^T \mathbf{U}^T \mathbf{L} \mathbf{U} \mathbf{r}_\tau) + \frac{\lambda_3}{2} \|\mathbf{s}_\tau\|_1 \right), \end{aligned} \quad (18)$$

where $\lambda_3 > 0$.

For given $\boldsymbol{\Omega}, \mathbf{U}, \mathbf{x}$ and \mathbf{s} , the expression

$$\|\boldsymbol{\Omega} (\mathbf{x} - \mathbf{U} \mathbf{r} - \mathbf{s})\|_2^2 + \lambda_1 \|\mathbf{r}\|_2^2 + \lambda_2 (\mathbf{r}^T \mathbf{U}^T \mathbf{L} \mathbf{U} \mathbf{r}) + \lambda_3 \|\mathbf{s}\|_1 \quad (19)$$

is minimized when

$$\mathbf{r} = \mathbf{B} (\mathbf{x} - \mathbf{s}), \quad (20)$$

where

$$\mathbf{B} = \mathbf{A}^{-1} \mathbf{U}^T \boldsymbol{\Omega}$$

and

$$\mathbf{A} = \lambda_1 \mathbf{I}_r + \mathbf{U}^T (\boldsymbol{\Omega} + \lambda_2 \mathbf{L}) \mathbf{U}.$$

Treating \mathbf{r} as a function of \mathbf{s} and plugging it back into (19), the joint minimization of \mathbf{r} and \mathbf{s} for given \mathbf{U} is formulated as

$$\begin{aligned} \min_{\mathbf{s}} \|\boldsymbol{\Omega} (\mathbf{I}_m - \mathbf{U} \mathbf{B}) (\mathbf{x} - \mathbf{s})\|_2^2 + \|\sqrt{\lambda_1} \mathbf{B} (\mathbf{x} - \mathbf{s})\|_2^2 \\ + \|\sqrt{\lambda_2} \mathbf{L}^{\frac{1}{2}} \mathbf{U} \mathbf{B} (\mathbf{x} - \mathbf{s})\|_2^2 + \lambda_3 \|\mathbf{s}\|_1, \end{aligned}$$

¹In practice if $m' = O(\log m)$ then we can recover the sparse vector.

where we have used that \mathbf{L} is PSD. This, in turn can be formulated as the following lasso estimator:

$$\min_{\mathbf{s}} \|\mathbf{C}(\mathbf{x} - \mathbf{s})\|_2^2 + \lambda_3 \|\mathbf{s}\|_1, \quad (21)$$

where \mathbf{C} is the $(m + r + m) \times m$ matrix such that

$$\mathbf{C} = \left[(\boldsymbol{\Omega}(\mathbf{I}_m - \mathbf{U}\mathbf{B}))^T, \sqrt{\lambda_1} \mathbf{B}^T, \sqrt{\lambda_2} \left(\mathbf{L}^{\frac{1}{2}} \mathbf{U}\mathbf{B} \right)^T \right]^T.$$

This is a convex optimization problem and therefore efficiently solvable. We use the above to compute \mathbf{r}_t and \mathbf{s}_t using $\boldsymbol{\Omega}_t$, \mathbf{x}_t and \mathbf{U}_{t-1} . Similar to the above algorithm, we use these computed values to compute \mathbf{U}_t .

Taking partial derivative of (18) with respect to \mathbf{U} and setting it to zero, we get

$$\mathbf{Q}_t = \lambda_1 \mathbf{U} + \lambda_2 \mathbf{L}\mathbf{U}\mathbf{R}_t + \sum_{\tau=1}^t \boldsymbol{\Omega}_\tau \mathbf{U} \mathbf{r}_\tau \mathbf{r}_\tau^T$$

where $\mathbf{Q}_t = \sum_{\tau=1}^t \boldsymbol{\Omega}_\tau (\mathbf{x}_\tau - \mathbf{s}_\tau) \mathbf{r}_\tau^T$ and, as before, $\mathbf{R}_t = \sum_{\tau=1}^t \mathbf{r}_\tau \mathbf{r}_\tau^T$.

As before, we vectorize and solve, thereby getting

$$\mathbf{u} = \left(\sum_{\tau=1}^t \mathbf{r}_\tau \mathbf{r}_\tau^T \otimes \boldsymbol{\Omega}_\tau + \lambda_1 \mathbf{I}_{mr} + \mathbf{R}_t \otimes \mathbf{L} \right)^{-1} \mathbf{q}_t, \quad (22)$$

where $\mathbf{q}_t = \text{vec}\{\mathbf{Q}_t\}$.

The algorithm is summarized as Algorithm 6.

Algorithm 2: Online Robust Matrix Completion on Graphs

Input: $\lambda_1, \lambda_2, \mathbf{L}$

Output: Computed Subspaces \mathbf{U}_t and vectors $\mathbf{r}_t, \mathbf{s}_t$

- 1 **Initialize:** \mathbf{U}_0
 - 2 **for** $t = 1, 2, \dots$ **do**
 - 3 Compute \mathbf{s}_t by solving the lasso (21) using \mathbf{U}_{t-1} and $\boldsymbol{\Omega}_t$
 - 4 Compute \mathbf{r}_t by applying equation (20)
 - 5 Update \mathbf{R}_t and \mathbf{Q}_t using $\mathbf{x}_t, \mathbf{r}_t$ and \mathbf{s}_t
 - 6 Compute \mathbf{U}_t using (22)
-

4. CONVERGENCE

In this section we will discuss the convergence of the proposed scheme, in particular, the robust scheme with missing entries. The convergence proofs for the other schemes follow similar steps.

Define the following: $g_t(\mathbf{U}, \mathbf{r}, \mathbf{s}) := \left(\frac{1}{2} \|P_{\boldsymbol{\Omega}_t}(\mathbf{x}_t - \mathbf{U}\mathbf{r} - \mathbf{s})\|_2^2 + \frac{\lambda_1}{2} \|\mathbf{r}\|_2^2 + \frac{\lambda_2}{2} (\mathbf{r}^T \mathbf{U}^T \mathbf{L}\mathbf{U}\mathbf{r}) + \lambda_3 \|\mathbf{s}\|_1 \right)$, and $g_t(\mathbf{U}) :=$

$\min_{\mathbf{r}, \mathbf{s}} g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$. The proposed algorithm effectively aims to minimize the following²:

$$C_t(\mathbf{U}) = \frac{1}{t} \sum_{\tau=1}^t g_\tau(\mathbf{U}) + \frac{\lambda_1}{2t} \|\mathbf{U}\|_F^2.$$

It is worth pointing out that as time increases, minimization of $C_t(\mathbf{U})$ becomes computationally demanding since it involves solving t least squares and t ℓ_1 minimization problems for the estimation of \mathbf{r} and \mathbf{s} respectively. For this reason, the algorithm actually minimizes the following approximation of the above cost function:

$$\widehat{C}_t(\mathbf{U}) = \frac{1}{t} \sum_{\tau=1}^t g_\tau(\mathbf{U}, \mathbf{r}_\tau, \mathbf{s}_\tau) + \frac{\lambda_1}{2t} \|\mathbf{U}\|_F^2, \quad (23)$$

where

$$\{\mathbf{r}_t, \mathbf{s}_t\} = \arg \min_{\mathbf{r}, \mathbf{s}} g_t(\mathbf{U}_{t-1}, \mathbf{r}, \mathbf{s}).$$

For the analysis of convergence, we make the following assumptions:

- **A1:** $\{\boldsymbol{\Omega}_t\}_t$ and $\{\mathbf{x}_t\}_t$ are i.i.d. random processes.
- **A2:** Each \mathbf{x}_t and \mathbf{U}_t is in fixed compact set $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{C} \subset \mathbb{R}^{m \times r}$, respectively.
- **A3:** $\widehat{C}_t(\mathbf{U})$ is strongly convex, i.e., $\lambda_{\min}(\nabla^2 \widehat{C}_t(\mathbf{U})) \geq \epsilon$ for a positive constant ϵ .
- **A4:** The lasso given in equation (21) has a unique solution.

Before we proceed to the proof a few words on the assumptions are due. Assumption A1 is typically adopted in several online learning problems, e.g., [19], and has been made in the online matrix completion problem, e.g., [15]. For \mathbf{x}_t , A2 naturally holds in many applications, e.g., media, data transmission. For \mathbf{U}_t is a technical assumption which simplifies the proof and has been verified through extensive simulations; however, it is also reasonable in many cases to suppose that the principle vectors of an underlying subspace are bounded. This is especially the case where the application forces it, e.g., you can only rate a movie one to five stars. Regarding assumption A3, we assume that the Hessian of the cost function is bounded. This is also considered in [14, 15] and essentially implies that $\frac{1}{t} \left(\sum_{\tau=1}^t \mathbf{r}_\tau \mathbf{r}_\tau^T \otimes \boldsymbol{\Omega}_\tau + \lambda_1 \mathbf{I}_{mr} + \mathbf{R}_t \otimes \mathbf{L} \right) \succcurlyeq \epsilon \mathbf{I}_{mr}$. An additional regularization term can be added to ensure that this assumption holds, but here for simplicity we won't consider such a case. Assumption A4 is reasonable since it is helped by the uniformly random matrices $\boldsymbol{\Omega}_t$ not affecting too much the incoherence of the subspace estimates \mathbf{U}_{t-1} .

We wish to show the following:

²We normalize with t so as to prevent the existence of unbounded values. It can be readily seen that the solution at each time step doesn't depend on the normalization.

Theorem 1 *If assumptions A1 – A4 hold, then Algorithm 6 converges to a stationary point of the objective function, i.e., $\lim_{t \rightarrow \infty} \nabla C_t(\mathbf{U}_t) = \mathbf{O}_{r \times m}$.*

In a nutshell, this theorem states that asymptotically the estimated subspace minimize the *original* cost function, despite the fact that the estimates occur from the minimization of an *approximate* cost function.

Mardani et al. [15] study an online matrix completion-type problem in the context of tracking network anomalies. Application aside, and framed in our notation, the algorithms they present essentially try to compute the same low-rank \mathbf{U}_t , matrices and sparse vectors \mathbf{s}_t as our algorithms do, but they make no use of graph structure in the sense we have done via the Laplacian. They prove a version of Theorem 1, but for us to apply their proof technique (which is, in turn, based on [14]), we must ensure the following lemma holds:

Lemma 1 *If assumptions A2 and A4 hold, then for \mathbf{U} in a compact set $\mathcal{C} \subset \mathbb{R}^{m \times r}$, the following are Lipschitz continuous functions of \mathbf{U} with constants independent of t : (i) $\{\mathbf{r}_t(\mathbf{U}), \mathbf{s}_t(\mathbf{U})\} = \arg \min_{\mathbf{r}, \mathbf{s}} g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$, (ii) $g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$, for fixed \mathbf{r}, \mathbf{s} , (iii) $g_t(\mathbf{U})$, and (iv) $\nabla g_t(\mathbf{U})$.*

Lemma 1 above is the equivalent of Lemma 1 in [15], and we modify their proof to cope with the terms arising from the Laplacian. Define

$$\mathbf{M}_t(\mathbf{U}) := \left[\begin{array}{c} \boldsymbol{\Omega}_t(\mathbf{I}_m - \mathbf{U}\mathbf{B}_t(\mathbf{U})) \\ \sqrt{\lambda_1}\mathbf{B}_t(\mathbf{U}) \\ \sqrt{\lambda_2}\mathbf{L}^{\frac{1}{2}}\mathbf{U}\mathbf{B}_t(\mathbf{U}) \end{array} \right],$$

where $\mathbf{B}_t(\mathbf{U}) := \mathbf{A}_t(\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Omega}_t$ and $\mathbf{A}_t(\mathbf{U}) := \lambda_1\mathbf{I}_r + \mathbf{U}^T(\boldsymbol{\Omega}_t + \lambda_2\mathbf{L})\mathbf{U}$. Note, $\mathbf{A}_t(\mathbf{U})$ is positive definite, and therefore invertible.

For simplicity, we omit the subscript t below where it does not aid the argument.

Proof of Lemma 1 (i) As in Section 3, \mathbf{r} can first be expressed as an affine function of \mathbf{s} (see (20)), and after the Lipschitz continuity of $\mathbf{s}(\mathbf{U})$ is demonstrated, the Lipschitz continuity of $\mathbf{r}(\mathbf{U})$ follows easily. This is the approach taken in [15], and we apply a modified version of it below. Thus, defining

$$u(\mathbf{s}, \mathbf{U}_1, \mathbf{U}_2) := \|\mathbf{M}(\mathbf{U}_1)(\mathbf{x} - \mathbf{s})\|_2^2 - \|\mathbf{M}(\mathbf{U}_2)(\mathbf{x} - \mathbf{s})\|_2^2$$

(cf. (21)), it is shown that

$$u(\mathbf{s}(\mathbf{U}_2), \mathbf{U}_1, \mathbf{U}_2) - u(\mathbf{s}(\mathbf{U}_1), \mathbf{U}_1, \mathbf{U}_2) \geq c_0 \|\mathbf{s}(\mathbf{U}_2) - \mathbf{s}(\mathbf{U}_1)\|_2^2$$

for some constant $c_0 > 0$ independent of t . This holds for our case as well. It is then shown that $u(\cdot, \mathbf{U}_1, \mathbf{U}_2)$ is Lipschitz continuous, and we can follow the same steps of the proof until it is required to show that $\mathbf{M}(\mathbf{U})$ is Lipschitz continuous. Here we have to cater for the terms related to the Laplacian.

It is quite possible to apply the same technique as in [15], but we use a shorter argument thus:

$$\begin{aligned} \|\mathbf{M}(\mathbf{U}_1) - \mathbf{M}(\mathbf{U}_2)\| &\leq \|\boldsymbol{\Omega}[\mathbf{U}_1\mathbf{B}(\mathbf{U}_1) - \mathbf{U}_2\mathbf{B}(\mathbf{U}_2)]\| \\ &+ \sqrt{\lambda_1}\|\mathbf{B}(\mathbf{U}_1) - \mathbf{B}(\mathbf{U}_2)\| + \sqrt{\lambda_2}\|\mathbf{L}^{\frac{1}{2}}[\mathbf{U}_1\mathbf{B}(\mathbf{U}_1) - \mathbf{U}_2\mathbf{B}(\mathbf{U}_2)]\| \\ &\leq \left(1 + \sqrt{\lambda_2}\|\mathbf{L}^{\frac{1}{2}}\|\right)\|\mathbf{U}_1\mathbf{A}(\mathbf{U}_1)^{-1}\mathbf{U}_1^T - \mathbf{U}_2\mathbf{A}(\mathbf{U}_2)^{-1}\mathbf{U}_2^T\| \\ &+ \sqrt{\lambda_1}\|\mathbf{A}(\mathbf{U}_1)^{-1}\mathbf{U}_1^T - \mathbf{A}(\mathbf{U}_2)^{-1}\mathbf{U}_2^T\|. \end{aligned} \quad (24)$$

Now consider the function $f(\mathbf{U}) := \mathbf{A}(\mathbf{U})^{-1}\mathbf{U}^T$. This is differentiable with respect to \mathbf{U} and since \mathbf{U} is assumed to be constrained to a fixed compact space $\mathcal{C} \subset \mathbb{R}^{m \times r}$, Lipschitz continuity of $f(\mathbf{U})$ follows by the mean value theorem. Similarly for $\mathbf{U}f(\mathbf{U})$. It follows that there is a constant $c_2 > 0$ independent of t such (24) is bounded by $c_2\|\mathbf{U}_1 - \mathbf{U}_2\|$. The rest of the rest of the proof for part (i) follows as in [15].

(ii) $g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$ is a quadratic function of \mathbf{U} on a compact set and so clearly Lipschitz.

(iii) Using $g_t(\mathbf{U}) = g_t(\mathbf{U}, \mathbf{r}_t(\mathbf{U}), \mathbf{s}_t(\mathbf{U}))$ where $\{\mathbf{r}_t(\mathbf{U}), \mathbf{s}_t(\mathbf{U})\} = \arg \min_{\mathbf{r}, \mathbf{s}} g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$, we have (omitting t subscripts)

$$\begin{aligned} g(\mathbf{U}_2) - g(\mathbf{U}_1) &= \frac{1}{2}\|P_{\Omega}(\mathbf{U}_2\mathbf{r}(\mathbf{U}_2) + \mathbf{s}(\mathbf{U}_2))\|_2^2 \\ &- \frac{1}{2}\|P_{\Omega}(\mathbf{U}_1\mathbf{r}(\mathbf{U}_1) + \mathbf{s}(\mathbf{U}_1))\|_2^2 \\ &+ \langle \boldsymbol{\Omega}\mathbf{x}, \mathbf{U}_1\mathbf{r}(\mathbf{U}_1) + \mathbf{s}(\mathbf{U}_1) - \mathbf{U}_2\mathbf{r}(\mathbf{U}_2) - \mathbf{s}(\mathbf{U}_2) \rangle \\ &+ \frac{\lambda_1}{2}(\|\mathbf{r}(\mathbf{U}_2)\|_2^2 - \|\mathbf{r}(\mathbf{U}_1)\|_2^2) + \lambda_3(\|\mathbf{s}(\mathbf{U}_2)\|_1 - \|\mathbf{s}(\mathbf{U}_1)\|_1) \\ &+ \frac{\lambda_2}{2}(\mathbf{r}(\mathbf{U}_2)^T\mathbf{U}_2^T\mathbf{L}\mathbf{U}_2\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1)^T\mathbf{U}_1^T\mathbf{L}\mathbf{U}_1\mathbf{r}(\mathbf{U}_1)). \end{aligned}$$

As demonstrated in [15], the first term is bounded as

$$\begin{aligned} &\|P_{\Omega}(\mathbf{U}_2\mathbf{r}(\mathbf{U}_2) + \mathbf{s}(\mathbf{U}_2))\|_2^2 - \|P_{\Omega}(\mathbf{U}_1\mathbf{r}(\mathbf{U}_1) + \mathbf{s}(\mathbf{U}_1))\|_2^2 \\ &\leq c_3\left(\|\mathbf{U}_2 - \mathbf{U}_1\|\|\mathbf{r}(\mathbf{U}_2)\|_2 + \|\mathbf{U}_1\|\|\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1)\|_2\right. \\ &\quad \left. + \|\mathbf{s}(\mathbf{U}_2) - \mathbf{s}(\mathbf{U}_1)\|_2\right) \end{aligned}$$

for some constant $c_3 > 0$, the second is bounded as

$$\begin{aligned} &\langle \boldsymbol{\Omega}\mathbf{x}, \mathbf{U}_1\mathbf{r}(\mathbf{U}_1) + \mathbf{s}(\mathbf{U}_1) - \mathbf{U}_2\mathbf{r}(\mathbf{U}_2) - \mathbf{s}(\mathbf{U}_2) \rangle \\ &\leq \left(\|\mathbf{U}_2 - \mathbf{U}_1\|\|\mathbf{r}(\mathbf{U}_2)\|_2 + \|\mathbf{U}_1\|\|\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1)\|_2\right. \\ &\quad \left. + \|\mathbf{s}(\mathbf{U}_2) - \mathbf{s}(\mathbf{U}_1)\|_2\right)\|P_{\Omega}(\mathbf{x})\|_2, \end{aligned}$$

and the third term is bounded as

$$\begin{aligned} &\frac{\lambda_1}{2}(\|\mathbf{r}(\mathbf{U}_2)\|_2^2 - \|\mathbf{r}(\mathbf{U}_1)\|_2^2) + \lambda_3(\|\mathbf{s}(\mathbf{U}_2)\|_1 - \|\mathbf{s}(\mathbf{U}_1)\|_1) \\ &\leq \frac{\lambda_1}{2}\|\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1)\|_2(\|\mathbf{r}(\mathbf{U}_2)\|_2 + \|\mathbf{r}(\mathbf{U}_1)\|_2) \\ &\quad + \lambda_3\sqrt{r}\|\mathbf{s}(\mathbf{U}_2) - \mathbf{s}(\mathbf{U}_1)\|_2. \end{aligned}$$

By previous results, all the above terms are Lipschitz continuous. This was shown in [15]. It remains to show Lipschitz

continuity for the final term.

$$\begin{aligned}
& \mathbf{r}(\mathbf{U}_2)^T \mathbf{U}_2^T \mathbf{L} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1)^T \mathbf{U}_1^T \mathbf{L} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1) \\
&= \|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_2)\|_2^2 - \|\sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2^2 \\
&= \left(\|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_2)\|_2 - \|\sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2 \right) \\
&\quad \times \left(\|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_2)\|_2 + \|\sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2 \right).
\end{aligned}$$

By virtue of compactness the last term is bounded from above by some positive constant independent of t , and

$$\begin{aligned}
& \|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_2)\|_2 - \|\sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2 \\
&\leq \|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_1)\|_2 - \|\sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2 \\
&\quad + \|\sqrt{\mathbf{L}} \mathbf{U}_2 (\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1))\|_2.
\end{aligned}$$

By the submultiplicativity property of the operator norm, we have

$$\|\sqrt{\mathbf{L}} \mathbf{U}_2 (\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1))\|_2 \leq \|\sqrt{\mathbf{L}} \mathbf{U}_2\| \|\mathbf{r}(\mathbf{U}_2) - \mathbf{r}(\mathbf{U}_1)\|.$$

Furthermore,

$$\begin{aligned}
&\leq \|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_1)\|_2 - \|\sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2 \\
&\leq \|\sqrt{\mathbf{L}} \mathbf{U}_2 \mathbf{r}(\mathbf{U}_1) - \sqrt{\mathbf{L}} \mathbf{U}_1 \mathbf{r}(\mathbf{U}_1)\|_2 \\
&\leq \|\sqrt{\mathbf{L}}\| \|\mathbf{U}_2 - \mathbf{U}_1\| \|\mathbf{r}(\mathbf{U}_1)\|_2
\end{aligned}$$

where the last inequality follows by two applications of the submultiplicativity property of the operator norm. By compactness, $\|\sqrt{\mathbf{L}} \mathbf{U}_2\|$ and $\|\sqrt{\mathbf{L}}\| \|\mathbf{r}(\mathbf{U}_1)\|_2$ are both bounded from above by some positive constant independent of t .

Putting it all together proves the Lipschitz continuity of $g_t(\mathbf{U})$.

(iv) Since by assumption $\{\mathbf{r}_t(\mathbf{U}), \mathbf{s}_t(\mathbf{U})\} = \arg \min_{\mathbf{r}, \mathbf{s}} g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$ is unique as a minimizer of $g_t(\mathbf{U}, \mathbf{r}, \mathbf{s})$ for a given \mathbf{U} , a theorem of Danskin (see, e.g., [3]) allows us to say

$$\begin{aligned}
\nabla g_t(\mathbf{U}) &= \mathbf{r}_t(\mathbf{U}) \left(\mathbf{U} \mathbf{r}_t(\mathbf{U}) + \mathbf{s}_t(\mathbf{U}) - \mathbf{x}_t(\mathbf{U}) \right) \Omega \\
&\quad + \lambda_2 \mathbf{r}_t(\mathbf{U}) \mathbf{r}_t(\mathbf{U})^T \mathbf{U}^T \mathbf{L}.
\end{aligned}$$

To prove $g_t(\mathbf{U})$ is Lipschitz continuous, [15] has already shown $\|g_t(\mathbf{U}_2) - g_t(\mathbf{U}_1)\|_F \leq c_4 \|\mathbf{U}_2 - \mathbf{U}_1\|$ for the first term, and the proof applies just as well for the above. It remains to deal with the second term.

Writing \mathbf{R}_i for $\mathbf{r}_t(\mathbf{U}_i) \mathbf{r}_t(\mathbf{U}_i)^T$, we have

$$\begin{aligned}
& \|(\mathbf{R}_2 \mathbf{U}_2 - \mathbf{R}_1 \mathbf{U}_1) \mathbf{L}\|_F \leq \|\mathbf{R}_2 \mathbf{U}_2 - \mathbf{R}_1 \mathbf{U}_1\|_F \|\mathbf{L}\|_F \\
&\leq \|\mathbf{L}\|_F \left(\|\mathbf{R}_2 - \mathbf{R}_1\|_F \|\mathbf{U}_2\|_F + \|\mathbf{R}_1\|_F \|\mathbf{U}_2 - \mathbf{U}_1\|_F \right)
\end{aligned}$$

and from here it is straightforward to see that Lipschitz continuity follows. ■

Having proved Lemma 1, Lemma 2 below can be proved exactly as done in [15] (we omit the proof here). The lemma is used in the proof of Theorem 1, summarized below.

Lemma 2 ([15]) *If Assumptions A1 – A4 hold then $\widehat{C}_t(\mathbf{U}_t)$ converges and $\widehat{C}_t(\mathbf{U}_t) - C(\mathbf{U}_t) \rightarrow 0$ almost surely.*

Proof overview of Theorem 1 [15]: First, since the \mathbf{U}_t belong to a compact subset, then one can choose a convergent subsequence for which $\lim_{t_i \rightarrow \infty} \mathbf{U}_{t_i} = \mathbf{U}_*$. With a slight abuse of notation t_i will be substituted by t . Choose a sequence $\alpha_t > 0$ for which $\alpha_t \rightarrow 0$, $t \rightarrow \infty$. It holds that $\widehat{C}_t(\mathbf{U}_t + \alpha_t \mathbf{U}_o) \geq C_t(\mathbf{U}_t + \alpha_t \mathbf{U}_o)$, $\forall \mathbf{U}_o$, since the approximate cost always overestimates C_t . Exploiting the mean value theorem and Lemma 2, it can be shown that

$$\begin{aligned}
& \lim_{t \rightarrow \infty} \text{tr}(\mathbf{U}_o^T (\nabla \widehat{C}_t(\mathbf{U}_t) - \nabla C_t(\mathbf{U}_t))) \\
&+ \lim_{t \rightarrow \infty} \frac{1}{2} \alpha_t \text{tr}(\mathbf{U}_o^T (\nabla^2 \widehat{C}_t(\mathbf{U}_t^1) - \nabla^2 C(\mathbf{U}_t^2))) \geq 0,
\end{aligned} \tag{25}$$

for some matrices $\mathbf{U}_t^1, \mathbf{U}_t^2$. It can be readily shown that the second term tends to zero, since all the involved quantities apart from α_t are bounded and $\alpha_t \rightarrow 0$. So, we have that

$$\lim_{t \rightarrow \infty} \text{tr}(\mathbf{U}_o^T (\nabla \widehat{C}_t(\mathbf{U}_t) - \nabla C_t(\mathbf{U}_t))) \geq 0. \tag{26}$$

Since \mathbf{U}_o is arbitrarily chosen, (26) can be true iff $\lim_{t \rightarrow \infty} \nabla \widehat{C}_t(\mathbf{U}_t) - \nabla C_t(\mathbf{U}_t) = 0$. ■

5. EXPERIMENTS

5.1. SYNTHETIC NETFLIX DATASET

5.1.1. Generating the data

We conduct several experiments to confirm that our online algorithm exhibits better results when we utilize the Laplacian. In particular, we first generate a synthetic Netflix dataset, similarly as in [12]; the matrix that springs from this dataset obeys both the low-rank as well as the graph structure properties. The rows of the matrix represent users and the columns represent movies; the corresponding entries denote the rating. We consider a number $m_c = 10$ of communities, forming a partition for the rows. The underlying graph is constructed as follows: two individuals are adjacent in the graph if and only if they belong to the same community. Similarly we assume that we have $n_c = 20$ communities for the columns. The data matrix is then constructed by assigning a random value from $\{1, \dots, 5\}$ to each couple (movies community, users community).

It can be readily seen that this ideal matrix is of rank $r = \min(m_c, n_c)$. However, in practice it is very unlikely that we are dealing with low rank matrices, since the movie ratings are not necessarily consistent inside one users' community. And neither do we observe the communities one after another, because the order of appearance of the individuals is randomized. Therefore, to get a more realistic situation, we add noise and we also permute all the columns.

5.1.2. Generating the noise

The process to generate the noise in the Netflix framework is the following. Assuming that an user is likely to have a different opinion on a movie than the rest of his community, we define $N_{prob} \in [0 \dots 1]$ the probability of a rating to be affected by the noise, and $N_{level} \in \{1 \dots 5\}$ the maximum level of noise. Then, for each entry X_{ij} of the data matrix, we pick the parameter a according to a Bernoulli $\mathcal{B}(1, N_{prob})$ distribution and the parameter b according to the uniform $\mathcal{U}(\{-N_{level}, -N_{level} + 1, \dots, N_{level} - 1, N_{level}\})$ distribution. The entry of the corresponding corrupted matrix is then defined as:

$$\tilde{X}_{ij} = \max(\min(X_{ij} + ab, 5), 1)$$

One can easily verify that this definition preserves the fact that the occurring noisy entry will belong to the $\{1, \dots, 5\}$ set.

5.1.3. Error measurement

We run the online algorithm and compute for each time step the euclidean distance between the predicted vector, $\hat{\mathbf{x}}_i = \mathbf{U}_i \mathbf{r}_i$, and the true one, divided by the norm of the latter. Afterwards, we compute the mean over time and the resulting metric is given by: $err(t) = 20 \log_{10} \left(\frac{1}{t} \sum_{i=1}^t \frac{\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2}{\|\mathbf{x}_i\|_2} \right)$.

5.1.4. Results

In the following we study the realistic case of 20% missing entries in the observations. We assume the sampling of the entries is uniform, which may not be true in practice. However the case of non-uniform sampling goes beyond the scope of this paper.

We compare our proposed algorithm with the one presented in [15]. This methodology is suitable for robust online matrix completion, albeit no graph information is included. It is worth pointing out that, in both algorithms the regularization parameter related to the sparse outlier noise is set equal to zero, since in this experiment we do not assume outliers. The rest parameters are chosen via cross validation so that all the algorithms exhibit the best trade-off between convergence speed and steady state error floor. Contenting ourselves with a small level of noise ($N_{prob} = 0.3$ and $N_{level} = 1$), we obtain the results presented in Figure 1. It can be readily observed that the Laplacian regularization improves the performance, as expected. In fact, Algorithm 1 converges faster to a lower steady state error floor, compared to the algorithm in [15]. Moreover, setting $\lambda_2 = 10$ exhibits a slightly improved performance compared to the $\lambda_2 = 1$ case.

5.2. CONTINUOUS VALUES DATASET: THE ROBUST CASE

In the previous experiment, the entries of the data matrix are integers taking values between 1 and 5. Such experiments do

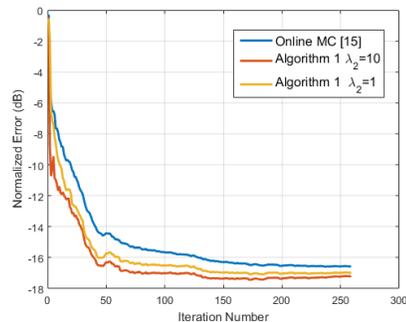


Fig. 1. Errors for Netflix dataset

not permit us to evaluate if the robust algorithm deals well with very large (outlier) values. Therefore, we turn our focus now on another dataset generated in a similar way as in the previous experiment, albeit the entries now are allowed to take continuous values. To that end, they are drawn from a zero-mean normal distribution with variance equal to 1. We add i.i.d. Gaussian noise, with standard deviation equal to $\sigma = 0.2$. On top of that, we add an “outlier” sparse matrix, the non-zero entries of which have a high magnitude compared to the data matrix. The sparse matrix is generated randomly and 1% of its entries are non-zero. These non-zeros entries are constructed so that their magnitude is at least 10 times the maximum value of the data matrix. Doing so, we have significant outliers. We compare the proposed robust algorithm (Algorithm 2) with: a) the non-robust one (Algorithm 1), b) a grassmannian manifold based algorithm suitable for online robust matrix completion, [11] and c) the algorithm of [15]. Again, the parameters are chosen via cross validation. Figure 2 presents the evolution of the error at each time step. It can be readily seen that, Algorithm 1 converges to a high error floor, since the presence of outliers is not taken into account. Furthermore, the proposed algorithm outperforms the other robust based schemes, since it exploits the underlying graph structure.

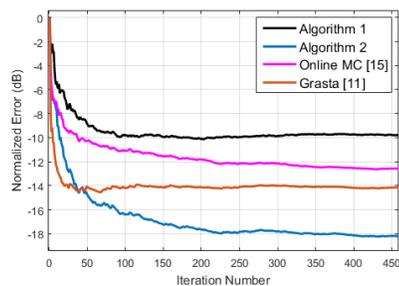


Fig. 2. Comparison of the standard and robust methods. Continuous values.

5.3. REAL NETWORK DATA

Let us now evaluate our proposed algorithm using data collected from a real network. In particular, we use the dataset captured in 2006, [20] on GEANT, the high bandwidth pan-European research and education backbone. The network comprises 22 nodes and 36 links. We consider that at each time step, the load from a subset of the links becomes available to us, whereas the load for the rest of them is unknown. Our goal is to estimate the load for these links. To that direction, we employ the proposed algorithm (Algorithm 1), for different values of the Laplacian related regularization parameter λ_2 , as well as the online matrix completion algorithm of [15]. In all the algorithms, we fix λ_1 to be equal to 0.1, since this particular choice leads to a fast convergence speed and a low steady state error floor at the same time. Moreover, in both algorithms the regularization parameter associated to the sparse outlier term is set equal to zero, since in that context there are no outliers. The results are shown in Fig. 3. First, it should be highlighted that the online matrix completion algorithm is able to provide a decent estimate of the missing entries due to the low-rank property of the link load traffic matrix. To be more specific, the network traffic pattern is highly correlated both temporally and spatially (i.e., across different links). This amounts to claiming that the data exhibit a low rank structure. Nevertheless, the results can be enhanced significantly if we exploiting the network graph topology, via the Laplacian smoothing.

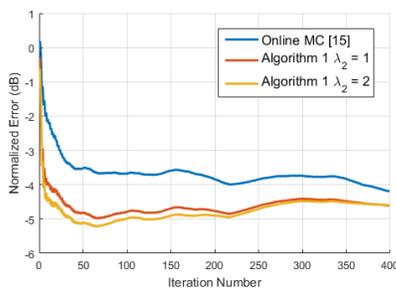


Fig. 3. Comparison of the proposed algorithm using the GEANT database

6. REFERENCES

- [1] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.
- [2] James Bennett and Stan Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007.
- [3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.
- [4] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
- [5] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- [6] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- [7] Yuejie Chi, Yonina C Eldar, and Robert Calderbank. Peltres: Parallel subspace estimation and tracking by recursive least squares from partial observations. *IEEE Transactions on Signal Processing*, 61(23):5947–5959, 2013.
- [8] Symeon Chouvardas, Yannis Kopsinis, and Sergios Theodoridis. Robust subspace tracking with missing entries: The set-theoretic approach. *IEEE Transactions on Signal Processing*, 63(19):5060–5070, 2015.
- [9] Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust pca via stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 404–412, 2013.
- [10] Han Guo, Chenlu Qiu, and Namrata Vaswani. An online algorithm for separating sparse and low-dimensional signal sequences from their sum. *IEEE Transactions on Signal Processing*, 62(16):4284–4297, 2014.
- [11] Jun He, Laura Balzano, and John Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011.
- [12] Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix completion on graphs. *arXiv preprint arXiv:1408.1717*, 2014.
- [13] Alan J Laub. *Matrix analysis for scientists and engineers*. Siam, 2005.

- [14] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.
- [15] Morteza Mardani, Gonzalo Mateos, and Georgios Giannakis. Dynamic anomalography: Tracking network anomalies via sparsity and low rank. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):50–66, 2013.
- [16] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- [17] Nauman Shahid, Nathanael Perraudin, Vassilis Kalofolias, and Pierre Vandergheynst. Fast robust pca on graphs. *arXiv preprint arXiv:1507.08173*, 2015.
- [18] Konstantinos Slavakis, Georgios Giannakis, and Gonzalo Mateos. Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. *IEEE Signal Processing Magazine*, 31(5):18–31, 2014.
- [19] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.
- [20] Steve Uhlig, Bruno Quoitin, Jean Lepropre, and Simon Balon. Providing public intradomain traffic matrices to the research community. *ACM SIGCOMM Computer Communication Rev*, 36(1), 2006.
- [21] Bin Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, 43(1):95–107, Jan 1995.
- [22] Bin Zhou and Guang-Ren Duan. An explicit solution to the matrix equation $AX - XF = BY$. *Linear Algebra and its applications*, 402:345–366, 2005.