Minimax Rank-1 Factorization

Julien M. Hendrickx, Alex Olshevsky and Venkatesh Saligrama. *

Abstract

We consider the problem of recovering a rank-one matrix from a subset of entries subject to arbitrary perturbations, assuming we have no information about the magnitude of perturbation. We propose a weighted log least square based algorithm whose performance for small disturbances matches exactly the fundamental lower bounds that we derive for this problem, and which are related to the spectral gap of a graph representing the revealed entries. We show that for larger disturbances, potentially exponentially growing errors are unavoidable if no additional information is available. We then propose a second algorithm relying on encoding the matrix factorization in the stationary distribution of a Markov chain and leveraging known lower and upper bounds on the entries, allowing to overcome the exponential error growth. Ours is the first work that achieves minimax bounds on rank-one approximation error with algorithms of near-linear time complexity.

1 Introduction

We consider the problem of rank-one approximation, xy^T, of a matrix, $A \in \mathbb{R}^{m\times n}$, when only an indexed subset, Ω, of its entries are revealed. As such our setup is general, imposing neither stochastic assumptions on support set Ω, nor assuming any pre-defined structure on the underlying matrix to be approximated. While rank-one approximation for a fully observed matrix is well-known, the corresponding problem for a partially observed matrix, particularly in such a general setting, is not well understood. We only make the minimal requirements on the support to ensure identifiability for completion of rank-one matrices: We associate the support set with a bipartite graph, G, with node set $\mathcal{I}_x \cup \mathcal{I}_y$ and nodes (i, j) are connected if the ijth entry is an element of Ω. Identifiability for rank-one matrix completion requires that the bipartite graph is connected, see e.g. [1, 2, 3].

Our motivation in considering such a general problem stems from several practical applications ranging from worker skill estimation in crowd-sourcing [1, 3, 4, 5, 6], inferring latent information from limited observations in collaborative filtering and recommender systems [7], and in other matrix completion applications such as global positioning and system identification [8]. To build intuition, we discuss one of these examples in detail in Section 1.1.

We develop stable approximation methods for rank-one estimation of a partially observed matrix that leverage different levels of information of the underlying matrix. Our first scheme is completely agnostic with no knowledge of bias. In this context, we propose a weighted log least square based algorithm whose performance for small disturbances matches exactly the fundamental lower bounds that we derive for this problem, and which are related to the spectral gap of the bipartite graph, G, associated with the revealed entries. We consider lower bounds on a class of consistent algorithms, namely those methods that require consistency if indeed the underlying matrix is rank-one. We show that any consistent scheme must suffer an estimation error that scales exponentially in the size of bias. This negative result leads us to consider algorithms that could possibly leverage knowledge of bias error. We propose a method, which encodes the rank-one factors into a stationary distribution of a suitable Markov chain, whose parameters leverage the known lower and upper bounds on the revealed

*J.H. is with the Department of Mathematical Engineering, ICTEAM, UCLouvain, Belgium the three authors are with the Department of Electrical and Computer Engineering, Boston University, USA

Preprint. Under review.

entries. Its performances are also governed by the spectral gap of G, and mitigate the exponential estimation error growth in terms of bias.

1.1 Motivating Example: A Case for Stable Rank-One Approximation

First, consider the worker skill estimation in *homogeneous* crowd-sourced problems, where a worker, *accepting* a binary task, outputs a binary label, which is assumed to be drawn according to a Bernoulli model parameterized by the worker skill-level. This model is known as the single-coin model and was proposed by Dawid & Skene [4]. In this context, correlations between workers accepting the same task can be estimated, and due to the homogeneity of the model, such correlation matrices are rank-one matrices. Nevertheless, only a partial subset of correlations can be observed [1, 3, 5, 6], due to the nature of crowd-sourced platforms, where a worker-task assignment matrix is sparsely filled and arbitrary, following no particular law (random or otherwise). This leads us to consider rank-one estimation from arbitrarily revealed entries, indexed by Ω, with $|\Omega| \geq m + n$.

On the other hand, the homogeneous model, considered in the several works, parameterizing a worker by a single parameter, is somewhat artificial; it is well known that there could be inherent asymmetries in worker skills [9]. A worker could recognize one class better than another. In these cases the correlation matrices are no longer rank-one. More generally, the situation is more complex in a multi-class classification problem in the same context.

In particular, suppose that $W \in \mathbb{N}$ workers are asked to provide labels to a series of M-class classification tasks. Let $\mathcal{G} = \{[1, 0, 0, \ldots, 0]^T, [0, 1, 0 \ldots, 0]^T, \ldots, [0, 0, \ldots, 0, 1]^T\} \subset \mathbb{R}^M$ represent the collection of one-hot encoding of the ground-truth labels. In round t, workers are offered a task, whose ground truth $g_t \in \mathcal{G}$ is unknown. In general, the ith worker's performance is a function of the true class $g_t \in \mathcal{G}$ as well as the label she chooses, but is otherwise assumed independent of other workers. We can consider this effect as a perturbation of a single-coin Dawid-Skene model. We assume that each worker has the same accuracy, p_i, and makes an error with uniform probability, $(1 - p_i)/(M - 1)$, on each class, i.e., a worker i, outputs a prediction, $Y_{i,t}$ such that,

$$\mathbb{P}(Y_{i,t} = \ell \mid g_t) = p_i \mathbb{1}_{\{\ell = g_t\}} + \frac{1 - p_i}{M - 1} \mathbb{1}_{\{\ell = g_t\}}$$

To see how this setup leads to a rank-one matrix approximation from partially revealed entries, we reparameterize p_i and define skill levels $s_i = \frac{M}{M-1} p_i - \frac{1}{M-1}$. Suppose, worker i and worker j label a strictly positive N_{ij} number of common tasks. We can then estimate the cross-correlation as: $\tilde{C}_{ij} = \frac{1}{N_{ij}} \sum_{t \mid (i,t),(j,t) \in A} \langle Y_{i,t}, Y_{j,t} \rangle$. It follows from straightforward algebraic manipulations that,

$$E\left[\frac{M - 1}{M} \tilde{C}_{ij} - \frac{1}{M - 1}\right] = s_i s_j.$$

Our goal reduces to estimating the parameters s_i and p_i from such partially revealed correlations. This is a rank-one matrix completion problem based on partially revealed noisy entries but with the important caveat that we have no control over what entries are revealed. Next, we consider a general model. The worker's probability law follows:

$$\mathbb{P}(Y_{i,t} = \ell \mid g) = (p_i + \eta_{i\ell g}) \mathbb{1}_{\{\ell = g\}} + \left(\frac{1 - p_i}{M - 1} + \eta_{i,\ell,g}\right) \mathbb{1}_{\{\ell \neq g\}}$$

where $\eta_{i,\ell,g}$ is the perturbation of the single coin model. Note that number of possible parameters, in the latter problem, scales both with the number of workers and the number of classes, which could be prohibitively large. It is not clear whether these parameters are even identifiable, since the ground-truth is unknown. Motivated by these issues we attempt to write our observations as a perturbation of rank-one matrix. We will represent this perturbation as Δ_{ij} below.

$$E\left[\frac{M - 1}{M} \tilde{C}_{ij} - \frac{1}{M - 1}\right] = s_i s_j + \Delta_{ij}.$$

Note that unlike the conventional stochastic noise assumptions, we can no longer assume Δ_{ij} to be a random variable, and it represents an unknown bias from the single-coin model. In particular, the correlation matrix is no longer rank-one and bears a complex relationship with worker/class skills. For this reason a natural question that arises is of estimating a good rank-one approximation of the underlying matrix, given only a limited amount of revealed components of the correlation matrix. Hence, for any such rank-one approximation algorithm, a desirable property is the requirement that it be stable, namely the estimation error scales with bias error.

2 Related Work

Our work is broadly related to a number of other works that either utilize rank-one matrix completion in the context of crowd-sourcing, collaborative filtering or deal with low-rank matrix completion [7, 8, 10, 5, 6, 1, 11, 2, 3, 9]. Apart from [2, 3], much of this literature assumes some form of incoherence, a probabilistic model or other structures on what indices of Ω are revealed. Many of these methods [8, 10, 5, 11] described in these contexts reduce to the fact that spectral decomposition is approximately preserved even though the matrices are only partially observed. Unlike, these papers and like [2, 3] we impose no such structure and so the spectral properties can no longer be leveraged.

Our work further differs from [1], who utilize propagation technique to estimate worker skills. Unlike these works we do not assume that the underlying matrix is either a rank-one matrix or that the observed entries are stochastic and unbiased. We allow for bias (rank-one approximation error) and consider algorithms that are agnostic to the size of the bias. This is an important difference as one can readily construct simple examples where propagation and gradient based techniques fail. While [9] also consider the possibility that the observations are not rank-one, they introduce other assumptions such as that the entries are observed at random and that various moments can be estimated among the different observed components can be estimated.

Other techniques proposed for matrix completion include nuclear norm minimization [8]. Unfortunately, nuclear norm minimization fails to solve our problem, as it will in almost all cases output a higher-rank matrix, even when there is no disturbance and sufficiently many entries are revealed, as shown in [2]. Ridge-regression based approaches have also been considered [3, 12], and appear natural for our setting, since Tikhonov regularization typically provides stable solutions. Nevertheless, as pointed out in [2], even these approaches are unstable. Moreover, they require solving non-convex optimization problems with a potentially high number of local minima, of the form

$$\min_{x,y} \left|\left|(xy^T - A^R)_\Omega\right|\right|_F + \lambda(||x||_2 + ||y||_2),$$

where Ω selects the entries for which data is available, and A^R denotes the revealed entries.

Our work is closely related to [2]. Like their work we require our algorithm be stable. On the other hand, we differ from [2]'s method in several ways. First, [2] is based on solving an SDP relaxation involving a matrix whose size grows quadratically with that of the matrix to be recovered leading in the best-case scenario to a fourth-order complexity. In addition, [2] provides guarantees on the relative error on a matrix of moments that is related to, but different from, the initial matrix to be recovered, and assumes knowing some bound on the magnitude of the perturbation. In contrast, ours is the first work that guarantees minimax bounds on estimation error with near-linear time algorithms.

3 Problem Statement

Given **a positive rank-1 matrix** $A^0 \in \mathbb{R}_+^{m\times n}$, we are revealed corrupted versions of certain entries

$$A^R_{ij} = A^0_{ij} + \Delta_{ij}, \qquad \forall (i,j) \in \Omega,$$

for a certain mask $\Omega \subset I_x \times I_y := \{1,\dots,m\} \times \{1,\dots,n\}$, and arbitrary perturbations Δ_{ij}. Our goal is to build a rank-1 estimate $\hat{A}$ of A^0, aiming at minimizing the error

$$||\hat{A} - A^0||_F^2 = \sum_{i\in I_x, j\in I_y} (\hat{A}_{ij} - A^0_{ij})^2.$$

We will bound this error as a function of $||\Delta||_F^2 = \sum_{(i,j)\in\Omega} \Delta_{ij}^2$, and of other characteristics of the matrices and of the mask. In one regime, we will further assume that $A^0_{ij} \in [\underline{\alpha}, \bar{\alpha}]$ for all i, j and that these bounds can be used in the algorithm. The structure of the mask Ω can conveniently be described by an undirected bipartite graph whose partitions correspond to the rows and columns of A. Slightly overloading the notations, we consider the set of $m + n$ nodes $I_x \cup I_y$ and define the graph G by connecting $i \in I_x$ and $j \in I_y$ if $(i, j) \in \Omega$, i.e. if A^R_{ij} is available. There is thus no edge inside I_x or I_y. It is known that the problem cannot be solved, even if the absence of disturbance, if the graph G is not connected, see e.g. [2]. Hence we make the following standing assumption.

Assumption 1. *The bipartite graph G is connected.*

Positivity of the matrix: The assumption that $A^0_{ij} > 0$ is in most cases not very restrictive. Indeed, if a not necessarily positive matrix A^0 is rank-1, then so is $|A^0|$ where the absolute value is taken entry-wise. Moreover, we have $||A^R_{ij}| - |A^0_{ij}|| \leq |A^R_{ij} - A^0_{ij}| = |\Delta_{ij}|$. So the disturbance if we consider $|A^R|$ as the perturbed revealed entries of $|A^0|$ is no larger than on the initial problem. We can therefore first recover $|A^0|$, and identify the sign pattern in a post-processing step. This is immediate if the disturbance does not change the sign of the revealed entries, and would remain simple if the number of sign changes is limited. A full analysis of this issue is out of the scope of this paper.

4 A weighted log-least square algorithm

4.1 Algorithm Description

Pre-processing: Our algorithm involves the logarithm of the revealed entries, which must thus be positive. We replace negative revealed entries by their absolute values, which decreases the magnitude of the perturbation as just explained in Section 3. Zero entries are more problematic, and indeed carry no information in a disturbance agnostic context. Hence they will be simply ignored. This decreases $||\Delta||_F$, but also modifies Ω and thus G, so one should verify it remains connected. Arguments similar to those used in the proof of Theorem 3 in Section 5 show that the matrix A^0 cannot be recovered with any guarantee of accuracy in case the connectivity is lost after removing edges whose corresponding revealed entries are zero. Therefore we assume $A^R_{ij} > 0$ for all $(i, j) \in \Omega$ in the sequel of this section.

A positive rank-1 matrix A can be rewritten as $A = xy^T$ for some $x \in \mathbb{R}^m, y \in \mathbb{R}^n$. Defining $z_i = x_i$ for every $i \in I_x$ and $z_j = 1/y_j$ for every $j \in I_y$ we obtain $A_{ij} = z_i/z_j$, and thus

$$\log A_{ij} = \log z_i - \log z_j \tag{1}$$

Our algorithm will identify the z_i by solving in the least square sense a system containing a version of this equation for each revealed entry. But we will weight the different equations to take into account the difference of effects of disturbances. Indeed, the equation involving the revealed entries are

$$\log z_i - \log z_j = \log A^0_{ij} + D_{ij} := \log A^0_{ij} + (\log(A^0_{ij} + \Delta_{ij}) - \log A^0_{ij}).$$

Hence a disturbance Δ_{ij} on A^0_{ij} will result in a larger disturbance D_{ij} on the corresponding linear equation if A^0_{ij} is small than if it is large. An adversary should thus favor perturbing small entries. We want to re-weight the equations in such a way that an adversary could not take advantage of more "sensitive" entries. For this purpose, we observe that based on a first order approximation,

$$\frac{D_{ij}}{\Delta_{ij}} = \frac{\log(A^0_{ij} + \Delta_{ij}) - \log A^0_{ij}}{\Delta_{ij}} \simeq \frac{1}{A^0_{ij}} \simeq \frac{1}{A^0_{ij} + \Delta_{ij}}.$$

Hence to obtain approximately similar sensitivities we should multiply each equation by a value similar to A^0_{ij} or A^R_{ij} for each $(i, j) \in \Omega$. We chose A^R_{ij} as they are directly available. So our algorithm consists in solving

$$A^R_{ij}(\log \hat{z}_i - \log \hat{z}_j) = A^R_{ij} \log A^R_{ij}, \qquad \forall (i, j) \in \Omega \tag{2}$$

in the least square sense. We can rewrite (2) in a more compact form by introducing a weighted version of the bipartite graph G: G_{W^R} has node set $I_x \cup I_y$, and $i \in I_x$ is connected to $j \in I_y$ if $(i, j) \in \Omega$ exactly as in G, and the weight of the edge (i, j) is $(A^R_{ij})^2$. The Laplacian L_{W^R} of this graph can be expressed as BW^RB^T, where B is the incidence matrix and $W^R \in \mathbb{R}^{|\Omega|\times|\Omega|}$ a diagonal matrix collecting all the weights $W^R{}_{(i,j)} = (A^R_{i,j})^2$ in an order consistent with the incidence matrix. Both B and L_{W^R} can be constructed from A^R in a time scaling linearly with $|\Omega|$. We also note that $L = BB^T$ is the Laplacian of G. The system (2) can be re-expressed as

$$(W^R)^{\frac{1}{2}} B^T \log \hat{z} = (W^R)^{\frac{1}{2}} \log A^R_\Omega$$

where A^R_Ω denotes the vector collecting the revealed entries A^R_{ij} in the same order as for B and W^R. Least-square solutions of this system are solutions of

$$L_{W^R} \log \hat{z} = BW^R \log A^R_\Omega,$$

where we have used $L_{W^R} := BW^R B^T$, and include thus

$$\log \hat{z} = L^{\dagger}_{W^R} B W^R \log A^R_{\Omega} \tag{3}$$

where † denotes the Monroe-Penrose inverse. It follows from [13] that (3) can be solved in near linear time in terms of $|\Omega|$: a solution with precision ϵ can be obtained in $O(|\Omega| \log^{\kappa}(n+m) \log(\epsilon^{-1}))$ operations for some constant $\kappa > 0$. One can then reconstruct $\hat{A}_{ij} = \hat{z}_i / \hat{z}_j = \hat{x}_i \hat{y}_j$ for all $i \in I_x, j \in I_y$ in $O(mn)$ operation for the whole matrix.

4.2 Accuracy Results

To state our main accuracy result, we need to introduce an additional matrix K_{W^0}, the Laplacian of the fully connected bipartite graph on $I_x \cup I_y$ whose weights are the $(A^0_{ij})^2$, i.e. the real values, not known to the user. Later we will also use the Laplacian K of the unweighted version of this graph.

Theorem 1. *Suppose that the disturbance satisfies $\Delta_{ij} \leq (c-1)A^0_{ij}$ for all revealed entries for some $c \geq 1$. Then the estimate* (3) *satisfies*

$$\left|\left|\hat{A} - A^0\right|\right|^2_F \leq c^2 \lambda_{\max}(K_{W^0} L^{\dagger}_{W^R}) \, ||\Delta||^2_F \, e^{2c\sqrt{R_{W^R,\max}}||\Delta||_F},$$

where $R_{W^R,\max}$ is the largest resistance distance in the weighted bipartite graph G_{W^R}.

For small values of $||\Delta||^2_F$, both the exponential factor and the constant c approach 1, so that the error is driven by $\lambda_{\max}(K_{W^0} L^{\dagger}_{W^R})$. The value of that factor depends on the specific interplay between the set of revealed entries, their values, and the disturbance. The following corollary provides a bound in term of more usual graph characteristics, albeit a more conservative one.

Corollary 1. *Let $\bar{\alpha}^0$ be an upper bound on the entries of the matrix A^0, and $\underline{\alpha}^R$ a lower bounds on the revealed entries A^R_{ij}. Under the same assumption as in Theorem 1, the estimate* (3) *satisfies*

$$\begin{aligned}\left|\left|\hat{A} - A^0\right|\right|^2_F &\leq c^2 \left(\frac{\bar{\alpha}^0}{\underline{\alpha}^R}\right)^2 \frac{m+n}{\lambda_2(L)} \, ||\Delta||^2_F \, e^{2c\sqrt{R_{\max}}||\Delta||_F/\underline{\alpha}^R} \\ &\leq \frac{c^2}{4} \left(\frac{\bar{\alpha}^0}{\underline{\alpha}^R}\right)^2 (m+n)^3 \, ||\Delta||^2_F \, e^{2c\sqrt{m+n}||\Delta||_F/\underline{\alpha}^R},\end{aligned}$$

where $R_{\max}$ is the maximal resistance in the unweighted bipartite graph G representing the revealed entries, $\lambda_2(L)$ is its algebraic connectivity, and m, n are the number of rows and columns of A^0.

5 Fundamental lower bounds

We now derive lower bounds on the performances achievable by any algorithm.

Theorem 2. *For any algorithm computing an estimate $\hat{A}$ of A^0 based solely on A^R, and any entry-wise positive rank-1 matrix A^R and mask Ω, one can find a matrix A^0 such that*

$$\left|\left|\hat{A} - A^0\right|\right|^2_F \geq \lambda_{\max}(K_{W^0} L^{\dagger}_{W^R}) \, ||\Delta||^2_F + o\left(||\Delta_{\Omega}||^2_F\right),$$

with $\Delta_{ij} = A^R_{ij} - A^0_{ij}$ for $(i,j) \in \Omega$. This result also holds for algorithms using explicit constraints on possible A^0 and/or bounds $\bar{\Delta}$ on $||\Delta||_F$ provided that A^0 lies in the interior of the allowed set $\mathcal{A}$ and that $\bar{\Delta} > (1+\epsilon) \, ||\Delta||_F$ for some $\epsilon > 0$.

Since we have seen in Theorem 1 that the error of algorithm (3) tends to $\lambda_{\max}(K_{W^0} L^{\dagger}_{W^R})||\Delta||^2_F$ when $||\Delta||^2_F$ is small, we conclude that *algorithm* (3) *is asymptotically optimal for small disturbances,* including in terms of multiplicative constant. For larger disturbances, one could regret the presence in Theorem 1 and Corollary 1 of (a) the exponential term, (b) and of the inverse $(\underline{\alpha}^R)^{-1}$ of the lowest measured value, and (c) the coefficient $c \geq \max_{(i,j)\in\Omega} A^R_{ij}/A^0_{ij}$. We will show that (a) and (b) cannot be avoided in our conditions.

We say that the revealed entries A^R are an exact subsample of a rank-1 matrix A if $A^R_{ij} = A_{ij}$ for all $(i,j) \in \Omega$, which means that A^R could be obtained by sampling the rank-1 A without any disturbance. We then say that an algorithm is *consistent* if it always returns $\hat{A} = A$ when A^R is an exact subsample of a unique rank-1 matrix A. Algorithms relying only on A^R would be expected to be consistent, for they would otherwise make nonzero errors in the absence of disturbance. Algorithms using additional information do not necessarily need to be consistent: If the algorithm uses bounds on the matrix entries, it could return a matrix different from A in case A does not meet these bounds, as in Section 6. Algorithms may also not be consistent when using some form of regularization, which should then be based on a bound or an implicit estimate on the disturbance, see e.g. [2]. A zero error would then only be guaranteed when the absence of disturbance is known. We now show that some inherent limitations apply to consistent algorithms.

Theorem 3.
(a) There exists a family of $n \times n$ matrices A^0, masks $Omega$ and disturbances Δ with constant norm $||\Delta||_F$ and uniformly bounded A^R_{ij} and A^0_{ij}, for which when n grows

$$\left|\left|\hat{A} - A^0\right|\right|_F^2 \geq E_n \rightarrow_{n\to\infty} \left(\exp\left(||\Delta||_F \sqrt{R_{\max}\left(\frac{1}{2} - O(n^{-1/2})\right)}\right) - 1\right)^2$$

for any consistent algorithm, where $R_{\max}$ is the largest resistance of the unweighted graph G.

(b) For every even n, there exists a family of square matrices A^R, A^0, masks Ω with $||\Delta||_F$, $\max A^0_{ij}$, $c = 1 + \max \Delta_{ij}/A^0_{ij}$ bounded uniformly independently of n such that for any consistent algorithm,

$$\left|\left|\hat{A} - A^0\right|\right|_F^2 \geq \frac{n^2}{9}(\min A^R_{ij})^{-2}.$$

In Claim (b), n is assumed even for simplicity in claim, but a very similar result holds for general n.

6 A stationary distribution-based algorithm

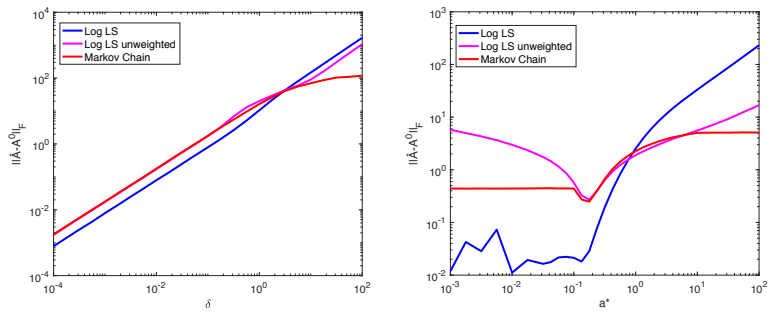
We now assume we have lower and upper bounds on the entries of A^0: $\underline{\alpha} \leq (A^0)_{ij} \leq \bar{\alpha}$ for all $i \in I_x, j \in I_y$, and propose an algorithm based on the stationary distribution of a Markov chain that can leverage this information. For small disturbances, its guaranteed performances are in general weaker than the asymptotically optimal algorithm in Section 4. But it presents important advantages for larger disturbances.

6.1 Algorithm

Pre-processing: Since we know that every $(A^0)_{ij}$ lies in $[\underline{\alpha}, \bar{\alpha}]$ for some known positive $\underline{\alpha}, \bar{\alpha}$, we will project all revealed entries on that interval: $(A^R)'_{ij} = \min\left(\max\left(\underline{\alpha}, (A^R)_{ij}\right), \bar{\alpha}\right)$. This step can only reduce the disturbance. Hence, we assume without loss of generality in the sequel that $(A^R)_{ij} \in [\underline{\alpha}, \bar{\alpha}]$. As part of the algorithms involve rescaling of the entries around 1, we define $\mu = \sqrt{\bar{\alpha}\underline{\alpha}}$ and $\rho = \sqrt{\bar{\alpha}/\underline{\alpha}}$, so that the interval $[\underline{\alpha}, \bar{\alpha}]$ can be re-expressed as $[\mu\rho^{-1}, \mu\rho]$.

The algorithm will consist in computing the stationary distribution of a Markov chain defined on G. We define the matrix $M^R \in \mathbb{R}^{(m+n)\times(m+n)}$ as

$$\begin{aligned}
(M^R)_{ij} &= \frac{\mu}{\mu + (A^R)_{ij}} && (i,j) \in \Omega, \\
(M^R)_{ji} &= \frac{(A^R)_{ij}}{\mu + (A^R)_{ij}} && (i,j) \in \Omega, \\
(M^R)_{ii} &= -\sum_{j \in I_y} (M^R)_{ij} && i \in I_x, \\
(M^R)_{jj} &= -\sum_{i \in I_x} (M^R)_{ji} && j \in I_y,
\end{aligned}$$



(a) (b)

Figure 1: Evolution of the average error $||\hat{A} - A^0||_F$ for our two algorithms and for an unweighted version of the algorithm of Section 4 in (a) a scenario where all revealed entries are perturbed by a random noise of magnitude $\delta/2$ (50×50 matrices with on average 20% of revealed entries), and (b) a targeted scenario where the smallest revealed entry is replaced by $a*$ (10×10 matrices with on average 50% of revealed entries). Initial matrices have entries between 10^{-1} and 10.

and all other entries are 0. We define M^0 in the same way, replacing A^R by A^0. This implies

$$(z^T M^0)_i = \sum_{j \in I_y} \frac{(A^0)_{ij} z_j - \mu z_i}{\mu + (A^0)_{ij}}, \forall i \in I_x, \qquad (z^T M^0)_j = \sum_{i \in I_x} \frac{\mu z_i - (A^0)_{ij} z_i}{\mu + (A^0)_{ij}}, \forall j \in I_y. \quad (4)$$

Moreover, we remind that the matrix A^0 can be written as xy^T for some $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, and we let $z_j = x_i$ for every $i \in I_x$ and $z_j = 1/y_j$ for every $j \in I_y$. Observe that

$$M^0_{ij} x_i / \sqrt{\mu} = \frac{\sqrt{\mu} x_i}{\mu + (A^0)_{ij}} = \frac{(A^0)_{ij} \sqrt{\mu}/(y_i}{\mu + (A^0)_{ij}} = M^0_{ij} (\sqrt{\mu}/y_i)^{-1}. \quad (5)$$

Hence (4) implies that the vector $(x^T/\sqrt{\mu}, \sqrt{\mu}(y^{-1})^T)$ is the principal left eigenvector of M^0, and hence proportional to the stationary distribution of the corresponding Markov chain. So one could expect the principal left eigenvector of M^R to provide a good approximation of that vector. The algorithm is built on this idea, akin to that used in [14, 15] to identify weights of a Bradley-Terry-Luce model, adding a projection step to further exploit the bounds on the matrix entries.

Algorithm 1 Projected Eigenvector Algorithm

1: Compute $\pi^R \in \mathbb{R}^{m+n}$, the principal left-eigenvector of M^R, normalized so that $e^T \pi^R = 1$
2: Let $\hat{\pi}$ be obtained by projecting each entry of π^R onto $[\frac{\rho^{-2}}{m+n}, \frac{\rho^2}{m+n}]$.
3: Return the matrix $\hat{A} \in \mathbb{R}^{m \times n}$ defined as $\hat{A}_{ij} = \mu^2 \hat{\pi}_i / \hat{\pi}_j$.

The eigenvector π^R can again be computed in near-linear time with respect to $|\Omega|$ [16], and the projection step has a cost $O(m+n)$. The reconstruction of $\hat{A}$ requires of course mn multiplications.

6.2 Accuracy Results

The accuracy guarantee of this algorithm are naturally expressed in terms of first-order norms rather than quadratic ones. Hence we use the "first-order Froebenius norm" $||M||_{F:1} = \sum_{i,j} |M_{ij}|$. We also provide a bound in term of (quadratic) Froebenius norm for comparison purposes.

Theorem 4. *The estimate $\hat{A}$ computed by Algorithm 1 satisfies*

$$\left|\left|\hat{A} - A^0\right|\right|_F \leq \left|\left|\hat{A} - A^0\right|\right|_{F:1} \leq 3(m+n)^2 \rho^4 \frac{\log 2\rho\sqrt{m+n}}{\lambda_2(M^0)} \max\left(||\Delta||_\infty, ||\Delta||_1\right)$$
$$\leq 3(m+n)^{2.5} \rho^4 \frac{\log 2\rho\sqrt{m+n}}{\lambda_2(M^0)} ||\Delta||_F,$$

with $\rho = \bar{\alpha}/\underline{\alpha}$ for $\bar{\alpha}, \underline{\alpha} > 0$ known upper and lower bounds on the elements of A^R, and $\lambda_2(M^0)$ real.

The next result result provides a bound independent of A^0.

Corollary 2. *The estimate $\hat{A}$ computed by Algorithm 1 satisfies*

$$\left|\left|\hat{A} - A^0\right|\right|_F \leq \left|\left|\hat{A} - A^0\right|\right|_{F:1} \leq 6(m+n)^{2.5}\rho^7 \frac{\log 2\rho\sqrt{m+n}}{\lambda_2(L)} ||\Delta||_F .$$

7 Discussion and Conclusions

We have presented two very different algorithms, based respectively on a weighted log least square problem and on the comparison of stationary distributions of two Markov chains. Both algorithms are computationally very efficient: Estimating the vectors x, y is done in near-linear time with respect to the number of revealed entries, so the main cost is in many cases driven by the reconstruction of the estimated matrix $\hat{A}$, which requires mn multiplications, i.e. the size of the output. There is thus little room for improvement, except maybe for large numbers of revealed entries.

The log least-square based algorithm of Section 4 was shown to achieve our minimax bound for small disturbances, and in that regime its error satisfies

$$\left|\left|\hat{A}_{LS} - A^0\right|\right|_F / ||\Delta||_F \leq \rho\sqrt{\frac{m+n}{\lambda_2(L)}} + o(1). \tag{6}$$

For larger disturbances, it suffers from three potentially important factors, (a) one exponential in the disturbance and the graph resistance, (b) the inverse of the smallest revealed entry (which can be large even if the actual smallest entry remains bounded away from zero), and (c) the largest ratio c between a revealed entry and its uncorrupted value, which could be large in situations where a large disturbance is concentrated on a single entry whose real value is small. The first two factors were shown to be unavoidable in the absence of additional information on the disturbance or on the matrix. The last one appears to be caused by the use of logarithms. Our second algorithm, which uses the additional knowledge of lower and upper bounds on the matrix entries and does not involve logarithms, avoids these three issues, and should therefore be preferred when disturbances are large. For small disturbances, however, its error guarantee is worse than (6):

$$\left|\left|\hat{A}_{MC} - A^0\right|\right|_F / ||\Delta||_F \leq \tilde{O}\left(\rho^7 \frac{(m+n)^{2.5}}{\lambda_2(L)}\right),$$

where $\tilde{O}(.)$ ignores the logarithmic factors. This difference should be nuanced because the proof Theorem 4 may rely on more conservative steps than that of Theorem 1. Besides, quadratic norms are natural for our log least square algorithm.

Experimental results presented in Figure 1 confirm that for random disturbances the log least square algorithm performs better than Algorithm 1. For large disturbances, the trend is inverted thanks in part to the ability of Algorithm 1 to exploit the knowledge of bounds on A^0. A more significant difference is observed for targeted attacks where the smallest (and thus most sensitive) entry is replaced by an arbitrary value a^*. We see then the efficiency of the weights used in the log least square algorithm, allowing it to tune down the influence of that entry if it becomes too small. These experiments also confirm the efficiency of our method, with an average of less than $3.10^{-3}s$ to recover 50×50 matrices using Matlab code not optimized for sparsity. We have also implemented a ridge-regression, which led to promising results for small matrices, but to important errors, numerous instabilities and convergence issues for larger ones, even for small disturbances. It was also significantly slower. See Appendix E for more details on the numerical experiments, in particular Figure 2.

As a comparison, [2] provides graph-independent bound in $O((m+n)^{3.5})$ on the *relative* error $||\hat{M} - M^0||_F / ||M^0||_F$ for the estimation of a matrix M_0 of moments that contain all first, second, third and fourth order monomials in x_i, y_j. This matrix contains A, but there appears to be no direct way of relating this relative error to an error on A^0. Besides, M has $O((m+n)^2)$ row and columns. As the method in [2] requires solving an SDP in involving M, the cost of which is known to be in the best case quadratic in the number of entries, its total computational cost is $O((m+n)^8)$.

Future works could aim at combining the advantages of our two algorithms, i.e. maintaining the optimal character for small disturbances, while avoiding the potentially large multiplicative factors in other regimes. Another challenge is to embed the possibility of exploiting estimates of the disturbance magnitude while preserving our algorithms low complexity. Finally, the question would also arise to use our tools for the completion of higher-rank matrices.

References

[1] Thomas Bonald and Richard Combes. A Minimax Optimal Algorithm for Crowdsourcing. In *Neural Information Processing Systems Conference NIPS*, Los Angeles, United States, 2017.

[2] Augustin Cosse and Laurent Demanet. Stable rank one matrix completion is solved by two rounds of semidefinite programming relaxation. *arXiv preprint arXiv:1801.00368*, 2017.

[3] Y. Ma, A. Olshevsky, V. Saligrama, and C. Szepesvari. Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[4] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979.

[5] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pages 285–294. ACM, 2013.

[6] Y. Zhang, X. Chen, D. Zhou, and M.I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *NIPS*, pages 1260–1268, 2014.

[7] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 713–719, New York, NY, USA, 2005. ACM.

[8] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010.

[9] Matthaeus Kleindessner and Pranjal Awasthi. Crowdsourcing with arbitrary adversaries. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2708–2717, 2018.

[10] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, June 2010.

[11] R. Ge, J. Lee, and T. Ma. Matrix completion has no spurious local minimum. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2973–2981. Curran Associates, Inc., 2016.

[12] Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1257–1264. Curran Associates, Inc., 2008.

[13] Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

[14] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in neural information processing systems*, pages 2474–2482, 2012.

[15] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2016.

[16] Michael B Cohen, Jonathan Kelner, John Peebles, Richard Peng, Anup B Rao, Aaron Sidford, and Adrian Vladu. Almost-linear-time algorithms for markov chains and new spectral primitives for directed graphs. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 410–419. ACM, 2017.

[17] Bojan Mohar. Eigenvalues, diameter, and mean distance in graphs. *Graphs and combinatorics*, 7(1):53–64, 1991.

[18] Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, pages 70–79, 2018.

[19] Tosio Kato. Perturbation theory for linear operators. *A Series of Comprehensive Studies in Mathematics*, 132, 1980.

[20] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

A Accuracy proofs for the log least square algorithm

A.1 Theorem 1

Notations: We will use the index $_\Omega$ to denote that we take a vectorized version of the elements of a matrix corresponding to the revealed entries, in an order consistent with the incidence matrix B (and thus the diagonal matrix W^R). A^R_Ω is thus a vector containing all the revealed entries, while A^0_Ω contains the real values of the entries that have been revealed. The 2-norm of such vectors is equivalent to their Froebenius norm. In particular $||\Delta||_F = \left|\left|A^R_\Omega - A^0_\Omega\right|\right|_F = \left|\left|A^R_\Omega - A^0_\Omega\right|\right|_2$.

We first express a bound on the error $\log \hat{A}_{ij} - \log A^0_{ij}$ on the logarithm of individual entries, which we will need both as intermediate step towards our global bound, and to derive a required bound on the largest relative error on any individual entry of the matrix. We remind that $D = \log A^R_\Omega - \log A^0_\Omega = \log(A^0_\Omega + \Delta) - \log A^0_\Omega$.

Lemma A.1. *The error on the logarithm of the individual estimates satisfies*

$$(\log \hat{A}_{ij} - \log A^0_{ij})^2 = (W^{R\frac{1}{2}}D)^T \left(W^{R\frac{1}{2}} B^T L^\dagger_{W^R}(e_i - e_j)(e_i - e_j)^T L^\dagger_{W^R} B W^{R\frac{1}{2}}\right)(W^{R\frac{1}{2}}D).$$

Proof. We have seen in (1) that $B^T \log z = \log A^0_\Omega$, and thus $L_{W^R} \log z = BW^R \log A^R_\Omega$. Since z is defined up to a multiplicative constant, we assume without loss of generality that $e^T \log z = 0$, and thus that it lies in the image of $L^\dagger_{W^R}$. Hence there holds $z = L^\dagger_{W^R} BW^R \log A^0_\Omega$. By definition of the algorithm (3) we have then

$$\begin{aligned}\log \hat{z} - \log z &= L^\dagger_{W^R} BW^R (\log A^R_\Omega - \log A^0_\Omega) \\ &= L^\dagger_{W^R} BW^R D.\end{aligned}$$

Hence using again $\log A_{ij} = \log z_i - \log z_j$, there holds

$$\begin{aligned}\log \hat{A}_{ij} - \log A^0_{ij} &= (\log \hat{z}_i - \log z_i) - (\log \hat{z}_j - \log z_j) \\ &= (e_i - e_j)^T L^\dagger_{W^R} BW^R D,\end{aligned}$$

from which the result follows. □

In order to use Lemma A.1, we now relate the magnitude of a weighted version $W^{R^{1/2}}D$ of the perturbation $D = \log A^R_\Omega - \log A^0_\Omega$ on the logarithm to that of the initial additive perturbation Δ.

Lemma A.2. *If $\Delta_{ij} \leq (c-1)A^0_{ij}$ for every $(i,j) \in \Omega$ for some $c \geq 1$, then $\left|A^R_{ij} D_{ij}\right| \leq c\left|\Delta_{ij}\right|$, for every $(i,j) \in \Omega$ and*

$$\left|\left|W^{R^{1/2}}D\right|\right|_F \leq c\,||\Delta||_F\,.$$

Proof. By concavity of the logarithm, we have

$$\left|\log A^R_{ij} - \log A^0_{ij}\right| \leq \left|A^R_{ij} - A^0_{ij}\right| \max\left(\frac{1}{A^R_{ij}}, \frac{1}{A^0_{ij}}\right) = |\Delta_{ij}| \max\left(\frac{1}{A^R_{ij}}, \frac{1}{A^0_{ij}}\right)$$

Moreover, the assumption of the Lemma implies $A^R_{ij} \leq cA^0_{ij}$ for $c \geq 1$. Hence we can bound

$$\left|A^R_{ij} D_{(i,j)}\right| = A^R_{ij}\left|\log A^R_{ij} - \log A^0_{ij}\right| \leq |\Delta_{ij}| \max\left(\frac{A^R_{ij}}{A^R_{ij}}, \frac{A^R_{ij}}{A^0_{ij}}\right) \leq c\left|\Delta_{ij}\right|,$$

proving the first claim of the Lemma. The second one follows from the definition of the $|\Omega| \times |\Omega|$ diagonal matrix W^R whose elements are the $(A^R_{ij})^2$. □

We can now provide a first bound on the error on the logarithm of the individual elements of the matrix, and a bound on the estimate $\hat{A}_{ij}$ to be used in the sequel.

Proposition A.1. *If $\Delta_{ij} \leq (c-1)A^0_{ij}$ for every $(i,j) \in \Omega$ for some $c \geq 1$, then*

$$\left|\log \hat{A}_{ij} - \log A^0_{ij}\right| \leq c\sqrt{R_{W^R,ij}}\,||\Delta||_F\,, \tag{7}$$

where $R_{W^R,ij}$ is the resistance distance between i and j on the weighted graph G_{W^R}. As a consequence,

$$\hat{A}_{ij} \leq A^0_{ij}\exp\left(c\sqrt{R_{W^R,ij}}\,||\Delta||_F\right) \tag{8}$$

Proof. It follows from Lemma A.1 and the symmetry of $L^\dagger_{W^R}$ that

$$(\log \hat{A}_{ij} - \log A^0_{ij})^2 \leq \left|\left|W^{R\frac{1}{2}}D\right|\right|_F^2 \lambda_{\max}\left(W^{R\frac{1}{2}}B^T L^\dagger_{W^R}(e_i-e_j)(e_i-e_j)^T L^\dagger_{W^R}BW^{R\frac{1}{2}}\right), \tag{9}$$

Remembering $L_{W^R} = BW^RB^T$, This eigenvalue can be re-expressed as

$$\begin{aligned}\lambda_{\max}\left((e_i-e_j)^T L^\dagger_{W^R}BW^{R\frac{1}{2}}W^{R\frac{1}{2}}B^T L^\dagger_{W^R}(e_i-e_j)\right) &= (e_i-e_j)^T L^\dagger_{W^R}L_{W^R}L^\dagger_{W^R}(e_i-e_j)\\ &= (e_i-e_j)^T L^\dagger_{W^R}(e_i-e_j),\end{aligned}$$

which is by definition the resistance $R_{W^R,ij}$. The claim (7) follows then from (9) and the bound $||W^{R\frac{1}{2}}D||_F^2 \leq c\,||\Delta||_F^2$ of Lemma A.2, and directly implies the bound (8) □

One could directly derive a bound on the global error $\hat{A} - A^0$ based on the individual bounds (7) and (8) but it would be very conservative, as for a given value $||\Delta||_F$, Δ cannot be simultaneously the worst possible for each individual $(i,j) \in \Omega$. The next proposition provides a global bound exploiting Δ being unique.

Proposition A.2. *If $\Delta_{ij} \leq (c-1)A^0_{ij}$ for every $(i,j) \in \Omega$ for some $c \geq 1$ and $\hat{A}_{ij} \leq \gamma A^0_{ij}$ for every $i \in I_x, j \in I_y$ and some $\gamma \geq 1$. Then*

$$\left|\left|\hat{A} - A^0\right|\right|_F^2 \leq \gamma^2 c^2 \lambda_{\max}\left(K_{W^0}L_{W^R}{}^\dagger\right)||\Delta||_F^2\,,$$

where this maximal eigenvalue is real.

Proof. Since $\left|e^b - e^a\right| \leq \max(e^a,e^b)\,|b-a|$ and $\max(\hat{A}_{ij}, A^0_{ij}) \leq \gamma A^0_{ij}$ by assumption, we have

$$\left|\hat{A}_{ij} - A^0_{ij}\right| \leq \gamma A^0_{ij}\left|\log \hat{A}_{ij} - \log A^0_{ij}\right|.$$

If follows then from Lemma A.1 that

$$(\hat{A}_{ij} - A^0_{ij})^2 \leq \gamma^2 (A^0_{ij})^2 (W^{R\frac{1}{2}}D)^T\left(W^{R\frac{1}{2}}B^T L^\dagger_{W^R}(e_i-e_j)(e_i-e_j)^T L^\dagger_{W^R}BW^{R\frac{1}{2}}\right)(W^{R\frac{1}{2}}D).$$

Summing over all pairs $(i,j) \in I_x \times I_y$ leads to

$$\left|\left|\hat{A} - A^0\right|\right|_F^2 \leq \gamma^2\left|\left|W^{R\frac{1}{2}}D\right|\right|_F^2 \lambda_{\max}\left(W^{R\frac{1}{2}}B^T L^\dagger_{W^R}K_{W^0}L^\dagger_{W^R}BW^{R\frac{1}{2}}\right), \tag{10}$$

where $K_{W^0} := \sum_{i\in I_x, j\in I_y}(e_i-e_j)(A^0_{ij})^2(e_i-e_j)^T$ is the Laplacian of a complete bipartite graph over $I_x \cup I_y$ where the edge (i,j) has weight $(A^0_{ij})^2$. This maximal eigenvalue is real since it is the eigenvalue of a symmetric matrix. We now compute it using $L_{W^R} = BW^RB^T$,

$$\begin{aligned}\lambda_{\max}\left(W^{R\frac{1}{2}}B^T L^\dagger_{W^R}K_{W^0}L^\dagger_{W^R}BW^{R\frac{1}{2}}\right) &= \lambda_{\max}\left(K_{W^0}L^\dagger_{W^R}BW^{R\frac{1}{2}}W^{R\frac{1}{2}}B^T L^\dagger_{W^R}\right)\\ &= \lambda_{\max}\left(K_{W^0}L^\dagger_{W^R}L_{W^R}L^\dagger_{W^R}\right)\\ &= \lambda_{\max}\left(K_{W^0}L^\dagger_{W^R}\right).\end{aligned}$$

The result follows then from (10), and the bound $||W^{R\frac{1}{2}}D||_F^2 \leq c\,||\Delta||_F^2$ of Lemma A.2. □

Theorem 1 is then obtained by using Proposition A.2 with the bound $\gamma = \exp\left(c\sqrt{R_{W^R,\max}}\,||\Delta||_F\right)$ guaranteed by (8) in Proposition A.1, with $R_{W^R,\max} := \max_{i\in I_x, j\in I_y} R_{W^R,ij}$ the largest resistance in the weighted graph G_{W^R}.

A.2 Corollary 1

In order to relate the bound of Theorem 1 to more usual characteristics of the graph, we now bound the coefficient $\lambda_{\max}(K_{W^0} L^\dagger_{W^R})$.

Proposition A.3. *Let $\bar{\alpha}^0, \underline{\alpha}^R$ be respectively an upper bound on the entries of A^0 and a lower bound on those of A^R.*

$$\lambda_{\max}(K_{W^0} L^\dagger_{W^R}) \leq \left(\frac{\bar{\alpha}^0}{\underline{\alpha}^R}\right)^2 \frac{m+n}{\lambda_2(L)},$$

where we remind that m, n are the number of rows and columns of A^0 and $\lambda_2(L)$ is the algebraic connectivity of the unweighted *bipartite graph G representing the revealed entries.*

Proof. It follows from Lemma D.1 in Appendix D that

$$\lambda_{\max}\left(K_{W^0} L^\dagger_{W^R}\right) \leq \lambda_{\max}(K_{W^0})\lambda_{\max}(L^\dagger_{W^R})$$

Since L_{W^R} is symmetric and has rank $n-1$, we have $\lambda_{\max}(L^\dagger_{W^R}) = \lambda_2(L_{W^R})^{-1}$. Because the absolute values of the off-diagonal elements of L_{W^R} (i.e. the weights) are all at least $(\underline{\alpha}^R)^2$, Lemma D.3 in Appendix D implies then

$$\lambda_2(L_{W^R}) \geq (\underline{\alpha}^R)^2 \lambda_2(L), \tag{11}$$

where we remind that L is the Laplacian of the unweighted bipartite graph G representing the mask Ω. A parallel argument shows that $\lambda_{\max}(K_{W^0}) \leq (\bar{\alpha}^0)^2 \lambda_{\max}(K) = (m+n)(\bar{\alpha}^0)^2$, where K is the Laplacian of the complete bipartite graph on $I_x \cup I_y$, whose maximal eigenvalue is $m+n$, from which the statement of this proposition follows. □

We note that the bound of Proposition A.3 could be conservative in terms of the interplay between the values in A^0, A^R and the graph, but is not very conservative in terms of the graph properties. Indeed, a slightly more complicated argument shows that

$$\lambda_{\max}\left(K_{W^0} L^\dagger_{W^R}\right) \geq \left(\frac{\bar{\alpha}^R}{\underline{\alpha}^0}\right)^2 \frac{\min(m,n)}{\lambda_2(L)},$$

where $\bar{\alpha}^R, \underline{\alpha}^0$ are respectively an upper bound on the entries of A^R and a lower bound on those of A^0.

Since $(\underline{\alpha}^R)^2$ bounds all weight in G_{W^R} from below, the largest resistance $R_{W^R,\max}$ in that graph is at most $(\underline{\alpha}^R)^{-2} R_{\max}$, with $R_{\max}$ the largest resistance of the corresponding unweighted graph G. The first part of Corollary 1 follows from this observation, Proposition A.3 and Theorem 1. Let now $\mathcal{D} \leq m+n$ be the diameter of the graph G. The second part of Corollary 1 follows from the classical bound $R_{\max} \leq \mathcal{D} \leq m+n$ and from $\lambda_2(L) \geq \frac{4}{\mathcal{D}(m+n)} \geq \frac{4}{(m+n)^2}$ [17].

B Lower Bounds

B.1 Small disturbance: proof of Theorem 2

We are given a mask Ω and a rank 1 matrix $A = xy^T$ of which we will be revealed the entries corresponding to Ω (i.e. A^R_Ω). We will prove the result by constructing, for a given value of $||\Delta||_F$ two matrices A^a, A^b that could act as A^0 and are sufficiently distant one from each other.

We take a fixed vector $\zeta \in \mathbb{R}^{m+n}$ and a sufficiently small constant δ, both to be specified later. We let then $A^a = x^a (y^a)^T, A^a = x^b (y^b)^T$ with

$$\begin{aligned} x^a_i &= x_i(1+\delta\zeta_i) \qquad & x^b_i &= x_i(1-\delta\zeta_i) \qquad & \forall i \in I_x \\ y^a_j &= y_j(1-\delta\zeta_j) \qquad & y^b_j &= y_j(1+\delta\zeta_j) \qquad & \forall j \in I_y \end{aligned}$$

We first compute the norm of $\Delta^a := (A^a)_\Omega - A^R_\Omega = (A^a - A)$.

$$\begin{aligned}
||\Delta^a||_F^2 = ||(A^a - A)_\Omega||_F^2 &= \sum_{(i,j)\in\Omega} \left(x_i y_j (1+\delta\zeta_i)(1-\delta\zeta_j) - x_i y_j\right)^2 \\
&= \sum_{(i,j)\in\Omega} x_i^2 y_j^2 \left(\delta(\zeta_i - \zeta_j) - \zeta_i\zeta_j\delta^2\right)^2 \\
&= \delta^2 \sum_{(i,j)\in\Omega} A_{ij}^2 (\zeta_i - \zeta_j)^2 + o(\delta^2) \\
&= \delta^2 \zeta^T L_W \zeta + o(\delta^2), \qquad (12)
\end{aligned}$$

where L_W is the Laplacian of the weighted bipartite graph on $I_x \cup I_y$ corresponding to Ω where the edge (i,j) has weight A_{ij}^2. Parallel arguments show that

$$\left|\left|\Delta^b\right|\right|_F^2 = \left|\left|(A^b - A)_\Omega\right|\right|_F^2 = \delta^2 \zeta^T L_W \zeta + o(\delta^2) \qquad (13)$$

and

$$\left|\left|A^b - A^a\right|\right|_F^2 = 4\delta^2 \zeta^T K_W \zeta + o(\delta^2), \qquad (14)$$

where K_W is the Laplacian of the weighted complete bipartite graph on $I_x \cup I_y$ with weight A_{ij}^2. To select ζ, we let u be the eigenvector of $K_W L_W^\dagger$ corresponding to $\lambda_{\max}(K_W L_W^\dagger)$ and $\zeta := L_W^\dagger u$. It follows from (12) and (13) that

$$\left|\left|\Delta^\ell\right|\right|_F^2 = \delta^2 \zeta^T w + o(\delta^2), \qquad (15)$$

for $\ell = a, b$, and from (14) that

$$\begin{aligned}
\left|\left|A^b - A^a\right|\right|_F^2 &= 4\delta^2 \zeta^T K_W L_W^\dagger w + o(\delta^2) \\
&= 4\delta^2 \lambda_{\max}(K_W L_W^\dagger) \zeta^T w + o(\delta^2). \qquad (16)
\end{aligned}$$

Since u belongs by construction to the image $\mathrm{span}\{e\}^\perp$ of the Laplacian K_W, it is orthogonal to the kernel $\mathrm{span}\{e\}$ of the Laplacian L_W and of its pseudo-inverse $L_W^\dagger$, so that the definition $\zeta := L_W^\dagger u$ implies $u = L_W \zeta$, and $\zeta^T u = u^T L_W^\dagger u > 0$. Hence (14) and (16) imply that for $\ell = a, b$,

$$\left|\left|A^b - A^a\right|\right|_F^2 = 4\lambda_{\max}(K_W L_W^\dagger) \left|\left|\Delta^\ell\right|\right|_F^2 + o(\left|\left|\Delta^\ell\right|\right|_F^2). \qquad (17)$$

Suppose now that $A_{ij}^R = A_{ij}$ for every $(i,j) \in \Omega$, and that A is in the interior of the set of allowed matrices $\mathcal{A}$. For sufficiently small δ and thus $||\Delta||_F$, we will have $A^a, A^b \in \mathcal{A}$, $||\Delta^a||_F^2 \leq (1+\epsilon)\left|\left|\Delta^b\right|\right|_F^2$ and $\left|\left|\Delta^b\right|\right|_F^2 \leq (1+\epsilon)||\Delta^a||_F^2$, so that both A^a, A^b would be possible values of A^0 even if the algorithm explicitly uses the set $\mathcal{A}$ and a bound $\bar{\Delta} \geq (1+\epsilon)||\Delta||_F^2$. It follows then from the triangular inequality and (17) that for any estimate $\hat{A}$ there would hold

$$\left|\left|\hat{A} - A^0\right|\right|^2 \geq \lambda_{\max}(K_W L_W^\dagger) \left|\left|\Delta^\ell\right|\right|_F^2 + o(\left|\left|\Delta^\ell\right|\right|_F^2). \qquad (18)$$

for at least one choice among $A^0 = A^a$ or $A^0 = A^b$. To conclude the result, we need to relate $\lambda_{\max}(K_W L_W^\dagger)$ to $\lambda_{\max}\left(K_{W^0} L_{W^R}^\dagger\right)$. Observe first that $L_{W^R} = L_W^\dagger$ because $A_{ij}^R = A_{ij}$. We define the function $t \to \tilde{K}_W(t) \in \mathbb{R}^{(n+m)\times(n+m)}$ by

$$\begin{aligned}
(\tilde{K}_W(t))_{ij} = (\tilde{K}_W(t))_{ji} &= x_i(1+t\zeta_i) y_j (1 - t\zeta_j) && \forall i \in I_x, j \in I_y \\
(\tilde{K}_W(t))_{ii} &= -\sum_{j\in I_y} (\tilde{K}_W(t)) && \forall i \in I_x, \\
(\tilde{K}_W(t))_{jj} &= -\sum_{i\in I_x} (\tilde{K}_W(t)) && \forall j \in I_y,
\end{aligned}$$

and the other entries being 0. Observe that $\tilde{K}_W$ is analytic, $K_W = \tilde{K}_W(0)$, $K_{W^0} = \tilde{K}_W(\delta)$ if $A^0 = A^a$ and $\tilde{K}_W(-\delta)$ if $A^0 = A^b$. Besides, $\delta = \Theta(||\Delta||_F)$. Lemma D.2 and $L_W^\dagger = L_{W^R}$ imply then

$$\lambda_{\max}\left(K_W L_W^\dagger\right) = \lambda_{\max}\left(K_{W^0} L_{W^R}^\dagger\right) + o(||\Delta||_F),$$

which implies the result of Theorem 2 together with (18).

B.2 Larger disturbance: proof of Theorem 3

We begin with the claim (a) about the exponential factor. For any given n, we take $A^0 = ee^T$, and the mask $\Omega = \{(i,i), i = 1,\dots,n\} \cup \{(i,i-1), i = 2,\dots,n\}$, that is, the entries on the main diagonal and the first other diagonal. We then take the disturbances

$$\Delta_{ii} = 0 \qquad \Delta_{i(i-1)} = \delta,$$

for all i for which these are defined and for some $\delta > 0$. The revealed entries are then

$$A^R_{ii} = 1 \qquad A^R_{i(i-1)} = 1 + \delta,$$

Clearly, $||\Delta||_F^2 = (n-1)\delta^2$ so $\delta = ||\Delta||_F / \sqrt{n-1}$. We then define the rank-1 matrix A by $A_{ij} = (1+\delta)^i (1+\delta)^{-j}$, and observe that A^R is an exact subsample of A because $A^R_{ij} = A_{ij}$ for every $(i,j) \in \Omega$. Moreover, it is not an exact subsample of any other matrix because the graph corresponding to the Ω is connected. Hence any consistent algorithm returns by definition $\hat{A} = A$, so that $(\hat{A} - A^0)_{ij} = (1+\delta)^{i-j} - 1$. In particular, remembering $\delta = \frac{||\Delta||_F}{\sqrt{n-1}}$, we have

$$\left|\left|\hat{A} - A^0\right|\right|_F^2 \geq (\hat{A}_{1n} - A^0_{1n})^2 = \left(\left(1 + \frac{||\Delta||_F}{\sqrt{n-1}}\right)^{n-1} - 1\right)^2 =: E_n.$$

When n grows for a fixed $||\Delta||_F$, we obtain

$$\lim_{n\to\infty} E_n = \left(e^{||\Delta||_F \sqrt{n-1}} - 1\right)^2.$$

We conclude part (a) of Theorem 3 by observing that the both A^R_{ij} and A^0_{ij} are uniformly bounded, and that the graph G_{W^R} corresponding to the mask Ω is a line graph on $2n$ nodes, with $n-1$ weights $1+\delta$ and n weights 1, so that

$$R_{W^R,\max} = n + (n-1)(1+\delta) = n + (n-1)\left(1 + \frac{||\Delta||_F}{\sqrt{n-1}}\right),$$

so that $\sqrt{n-1} = \sqrt{R_{W^R,\max}(\frac{1}{2} - O(n^{-1/2}))}$.

We now move to part (b). For any fixed even n, we let again $A^0 = ee^T$, and we consider the mask $\Omega = \{(i,j) : i,j \leq \frac{n}{2}\} \cup \{(i,j) : i,j \geq \frac{n}{2}\} \cup \{(1,n)\}$, i.e. we reveal the upper left-hand side quarter of the matrix and the lower right-hand side one, and the most upper right-hand side entry. We take $\Delta_{i,j} = 0$ for every revealed entry except $\Delta_{1,n} = \frac{1}{f} - 1$ for $f > 3$, so that $A^R_{ij)} = 1$ for all $(i,j) \in \Omega$ except $A^R_{1,n} = 1/f$. Clearly, all $||\Delta||_F^2 \leq 1$, and $\max_{(i,j)\in\Omega} \frac{\Delta_{ij}}{A^0_{ij}}$ and $\max_{(i,j)} A^0_{ij}$ are bounded independently of n, f, while $\min A^R_{ij} = f^{-1}$. Observe now that A^R is an exact subsample of the rank-1 matrix

$$A_f = \begin{pmatrix} ee^T & f^{-1}ee^T \\ fee^T & ee^T \end{pmatrix},$$

where the vectors e are of dimension $n/2$, and of no other rank-1 matrix. Hence any consistent algorithm would return $\hat{A} = A$ on the data A^R. Focusing on the error on the lower left-hand side block, and using $f > 3$, we would get

$$\left|\left|\hat{A} - A^0\right|\right|_F^2 \geq \frac{n^2}{4}(f-1)^2 \geq \frac{n^2}{9} f^2 = \frac{n^2}{9}(\min_{ij} A^R_{ij})^{-2}.$$

C Accuracy proofs for the stationary distribution-based algorithm

Observe first that dividing A^R and the lower and upper bounds by a constant c and multiplying the output of Algorithm 1 by the same c does not affect the final estimate $\hat{A}$. Moreover, both sides of the accuracy bound of Theorem 4 scale linearly with c if A^R, A^0, Δ are multiplied by c. Hence we can assume without loss of generality that $\mu = \sqrt{\bar{\alpha}\underline{\alpha}} = 1$, so that $A^0_{ij} \in [\rho^{-1}, \rho]$ for every i,j.

We first characterize the difference between the matrices M^R and M^0.

Lemma C.1.

$$\left|\left|M^R - M^0\right|\right|_\infty \leq 2\max\left(||\Delta||_\infty, ||\Delta||_1\right),$$

where the norms are the induced matrix norms, with $\Delta_{ij} = 0$ *for all* $(i,j) \notin \Omega$.

Proof. For any $x, y > 0$, there holds

$$\left|\frac{x}{1+x} - \frac{y}{1+y}\right| = \left|\frac{1}{1+x} - \frac{1}{1+y}\right| = \frac{|y-x|}{(1+x)(1+y)} \leq |y-x|.$$

Hence we have for every $(i,j) \in \Omega$

$$\left|M^R_{ij} - M^0_{ij}\right| = \left|M^R_{ji} - M^0_{ji}\right| \leq |\Delta_{ij}|. \tag{19}$$

Observe now that for a given matrix N whose rows sum to 0, we have

$$\begin{aligned}
||N||_\infty = \max_\ell \sum_k |N_{\ell k}| &= \max_\ell \left(|N_{\ell\ell}| + \sum_{k\neq\ell} |N_{\ell k}| \right) \\
&= \max_\ell \left(\left| -\sum_{k\neq\ell} N_{\ell k} \right| + \sum_{k\neq\ell} |N_{\ell k}| \right) \\
&\leq 2\max_\ell \sum_{k\neq\ell} |N_{\ell k}|.
\end{aligned}$$

Since the rows of $M^R - M^0$ sum to zero, we have then

$$\left|\left|M^R - M^0\right|\right|_\infty \leq 2 \max_{\ell \in I_x \cup I_y} \sum_{k \in I_x \cup I_y} \left|M^R_{\ell k} - M^0_{\ell k}\right|. \tag{20}$$

Consider first a $\ell = i \in I_x$. Then by the bipartite structure of M^R, M^0, the only off-diagonal nonzero $\left|M^R_{ik} - M^0_{ik}\right|$ are those for which $k \in I_y$. Hence, using (19), we have

$$\sum_{k \in I_x \cup I_y} \left|M^R_{ik} - M^0_{ik}\right| = \sum_{j \in I_y} \left|M^R_{ij} - M^0_{ij}\right| \leq \sum_{j \in I_y} |\Delta_{ij}| \leq ||\Delta||_\infty.$$

On the other hand, if $\ell = j \in I_y$, then

$$\sum_{k \in I_x \cup I_y} \left|M^R_{jk} - M^0_{jk}\right| = \sum_{i \in I_x} \left|M^R_{ji} - M^0_{ji}\right| \leq \sum_{i \in I_x} |\Delta_{ij}| \leq ||\Delta||_1.$$

The result follows then from (20). □

The sequel of the proof exploits results on the stationary distributions of (discrete-time) Markov chains, hence we need to introduce a reference stationary distribution. Remember we have assumed all entries of A^0 to be in $[\rho^{-1}, \rho]$ so that $\mu = 1$. Lemma D.4 implies that $A^0 = xy^T$ for some vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ with all x_i and y_j^{-1} in $[\rho^{-1}, \rho]$, and we have seen in (5) in Section 6.1 that $(x^T, (y^{-1})^T)$ is a left-eigenvector of M^0 corresponding to its eigenvalue 0. We defined its normalized version π^0 for which $\left|\left|\pi^0\right|\right|_1 = 1$. Due to the bounds on the entries of $(x^T, (y^T)^{-1})$, we see that the elements of π^0 all lie in $[\frac{\rho^{-2}}{m+n}, \frac{\rho^2}{m+n}]$, as those do those of $\hat{\pi}$ by construction, see Algorithm 1. Moreover, π^0 is still a left-eigenvector of M^0, and $\pi^0_i/\pi^0_j = x_i/(y_j^{-1}) = A^0_{ij}$ for every $i \in I_x, j \in I_y$.

Proposition C.1.

$$\left|\left|\hat{\pi} - \pi^0\right|\right|_1 \leq \frac{\log 2\rho\sqrt{m+n}}{2\lambda_2(M^0)} \left|\left|M^R - M^0\right|\right|_\infty$$

Proof. Since we will leverage results for discrete-time Markov-chains, we introduce the auxiliary matrices $P^R = I - \frac{1}{2d_{\max}} M^R$ and $P^0 = I - \frac{1}{2d_{\max}} M^0$, where $d_{\max}$ is the largest degree in G, i.e. the largest number of revealed elements on any row or column. Observe that the off-diagonal elements of M^R, M^0 are non-negative and bounded by 1, and that each row or column contains at most $d_{\max}$ of them. Moreover, $M^R e = M^0 e = 0$. Hence P^R, P^0 are row-stochastic matrices with positive diagonals.

The left-eigenvectors π^R and π^0 of M^R and M^0 corresponding to the eigenvalue 0 are also the principal left-eigenvectors of P^R, P^0, and thus the stationary distributions of the corresponding Markov chains since we have assumed them to be stochastic vectors. It follows then from [18] (Theorems 2, 3 and the discussion immediately after the statement of Theorem 3 in the supplementary materials) that

$$\begin{aligned} \left|\left|\pi^R - \pi^0\right|\right|_1 &\leq \frac{1}{2} \left|\left|P^R - P^0\right|\right|_\infty \left(\frac{\log R}{-\log \lambda_2(P^0)} + \frac{1}{1-\lambda_2(P^0)} \right) \\ &\leq \frac{1}{2} \left|\left|P^R - P^0\right|\right|_\infty \frac{\log R + 1}{1-\lambda_2(P^0)}. \end{aligned} \tag{21}$$

with

$$R = \max_{\ell \in I_x \cup I_y} \sqrt{\frac{1-\pi^0_\ell}{4\pi^0_\ell}}.$$

Moreover $P^0 = I - \frac{1}{2d_{\max}} M^0$ implies $1 - \lambda_2(P^0) = \frac{1}{2d_{\max}} \lambda_2(M^0)$. Since $\pi^0_\ell \geq \frac{\rho^{-2}}{m+n}$ for every ℓ and $\frac{1-x}{4x}$ is decreasing, we have then

$$\max_{\ell \in I_x \cup I_y} \sqrt{\frac{1-\pi^0_\ell}{4\pi^0_\ell}} \leq \sqrt{\frac{1-\rho^{-2}/(m+n)}{4\rho^{-2}/(m+n)}} \leq \frac{1}{2}\rho\sqrt{m+n}.$$

Reintroducing this and the expression $1 - \lambda_2(P^0) = \frac{1}{2d_{\max}} \lambda_2(M^0)$ in (21) leads to

$$\left|\left|\pi^R - \pi^0\right|\right|_1 \leq \left(\frac{1}{2}\right) \frac{1 + \log \rho\sqrt{m+n}/2}{\frac{1}{2d_{\max}}\lambda_2(M^0)} \left|\left|P^R - P^0\right|\right|_\infty \leq \frac{d_{\max} \log 2\rho\sqrt{m+n}}{\lambda_2(M^0)} \left|\left|P^R - P^0\right|\right|_\infty, \tag{22}$$

and the result follows from $\left|\left|P^R - P^0\right|\right|_\infty = \frac{1}{2d_{\max}} ||M^R - M^0||_\infty$, and from $||\hat{\pi} - \pi^0||_1 \leq ||\pi^R - \pi^0||_1$, since each $\hat{\pi}_\ell$ is the projection of π^R_ℓ on a set to which π^0_ℓ belongs. □

The last ingredient in the proof is a relation between the error $||\hat{\pi} - \pi^0||_1$ on the stationary distribution and the error $||\hat{A} - A^0||_F$ on the matrix.

Proposition C.2.

$$\left|\left|\hat{A} - A^0\right|\right|_F \leq 3(m+n)^2 \rho^4 \left|\left|\hat{\pi} - \pi\right|\right|_1.$$

Proof. Remember that $A^0_{ij} = \pi^0_i / \pi^0_j$ and that by construction $\hat{A}_{ij} = \hat{\pi}_i / \hat{\pi}_j$ for every i, j. We can decompose the error on an individual entry as

$$\left|\hat{A}_{ij} - A^0_{ij}\right| = \left|\frac{\hat{\pi}_i}{\hat{\pi}_j} - \frac{\pi^0_i}{\pi^0_j}\right| \leq \left|\frac{\hat{\pi}_i}{\hat{\pi}_j} - \frac{\hat{\pi}_i}{\pi^0_j}\right| + \left|\frac{\hat{\pi}_i}{\pi^0_j} - \frac{\pi^0_i}{\pi^0_j}\right| = \hat{\pi}_i \left|\frac{1}{\hat{\pi}_j} - \frac{1}{\pi^0_j}\right| + \frac{1}{\pi^0_j}\left|\hat{\pi}_i - \pi^0_i\right|,$$

so that

$$\sum_{i,j} \left|\hat{A}_{ij} - A^0_{ij}\right| \leq \sum_i \hat{\pi}_i \left|\left|(\hat{\pi})^{-1} - (\pi^0)^{-1}\right|\right|_1 + \sum_j \frac{1}{\pi^0_j} \left|\left|\hat{\pi} - \pi^0\right|\right|_1. \tag{23}$$

We first bound $\sum_i \hat{\pi}_i$. Observe that

$$\sum_i \hat{\pi}_i = \sum_i \pi^R_i + \sum_i (\hat{\pi}_i - \pi^R_i) = 1 + \sum_i (\hat{\pi}_i - \pi^R_i).$$

Moreover, $(\hat{\pi}_i - \pi_i^R)$ is by construction positive only when $\pi_i^R < \rho^{-2}/(m+n)$, in which case $\hat{\pi}_i = \rho^{-2}/(m+n)$. Hence

$$\sum_i \hat{\pi}_i \leq 1 + \sum_i \frac{\rho^{-2}}{m+n} \leq 2. \tag{24}$$

Second, since $\pi_j^0 \geq \rho^{-2}/(m+n)$, we have $\sum_j \frac{1}{\pi_j^0} \leq \rho^2 m(m+n)$. Finally

$$\left|(\hat{\pi}_i)^{-1} - (\pi_i^0)^{-1}\right| = \frac{\left|\hat{\pi}_i - \pi_i^0\right|}{\hat{\pi}_i \pi_i^0} \leq \left|\hat{\pi}_i - \pi_i^0\right| (m+n)^2 \rho^4. \tag{25}$$

Re-introducing these three bounds in (24) leads to

$$\sum_{i,j} \left|\hat{A}_{ij} - A_{ij}^0\right| \leq 2(m+n)^2\rho^4 \left|\left|\hat{\pi} - \pi\right|\right|_1 + m(m+n)\rho^2 \left|\left|\hat{\pi} - \pi\right|\right|_1 \leq 3(m+n)^2\rho^4 \left|\left|\hat{\pi} - \pi\right|\right|_1,$$

and the result follows then from

$$\left|\left|\hat{A} - A^0\right|\right|_F = \left|\left|\mathrm{vec}(\hat{\mathrm{A}} - \mathrm{A}^0)\right|\right|_2 \leq \left|\left|\mathrm{vec}(\hat{\mathrm{A}} - \mathrm{A}^0)\right|\right|_1 = \sum_{i,j} \left|\hat{A}_{ij} - A_{ij}^0\right| =: \left|\left|\hat{A} - A^0\right|\right|_{F:1}.$$

□

Theorem 4 follows from the combination of Lemma C.1, Propositions C.1 and C.2, together with the bound

$$\max\left(\left|\left|\Delta\right|\right|_\infty, \left|\left|\Delta\right|\right|_1\right) \leq \max\left(\sqrt{n}\left|\left|\Delta\right|\right|_2, \sqrt{m}\left|\left|\Delta\right|\right|_2\right) \leq \sqrt{\max(m,n)}\left|\left|\Delta\right|\right|_F.$$

To prove Corollary 2, observe first that all off-diagonal entries in M^0 are at least $\frac{\rho^{-1}}{1+\rho^{-1}} \leq \rho^{-1}/2$ in absolute values. Moreover, we have seen in (5) in Section 6.1 that $M_{k\ell}^0 \pi_k = M_{\ell k}^0 \pi_\ell$ for every $k, \ell \in I_x \cup I_y$, i.e. the corresponding Markov Chain is reversible, so that $\mathrm{diag}(\pi^0)M^0$ is symmetric. Lemma D.3 implies then

$$\lambda_2(M^0) \geq \frac{\min_k \pi_k^0}{\max_k \pi_k^0} \frac{\rho^{-1}}{2} \lambda_2(L).$$

Remember now that $(\pi^0)^T = K(x^T, (y^{-1})^T)$ for some constant K, and it follows from Lemma D.4 that x, y can be chosen so that $x_i, y_j \in [\rho^{-1}, \rho]$. As a consequence, $\frac{\min_k \pi_k^0}{\max_k \pi_k^0} \geq \rho^{-2}$, and thus $\lambda_2(M^0) \geq \frac{\rho^{-3}}{2}\lambda_2(L)$. Corollary 2 follows from the combination of this bound with Theorem 4.

D Technical Lemmas

Lemma D.1. *Let A, B be two PSD matrices. Then every eigenvalue of AB is real and non-negative, and*

$$\lambda_{\max}(AB) \leq \lambda_{\max}(A)\lambda_{\max}(B).$$

Proof. Since A is PSD, its singular value decomposition is of the form $A = U\Sigma U^T$. The diagonal matrix Σ only contains non-negative values, so $\Sigma^{\frac{1}{2}}$ is well defined. Hence the eigenvalues of $AB = U\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}U^T B$ are exactly the eigenvalues of $M := \Sigma^{\frac{1}{2}}U^T B U \Sigma^{\frac{1}{2}}$, and are thus real since M is symmetric. Moreover, M is positive semi-definite because for any x,

$$x^T M x = x^T \Sigma^{\frac{1}{2}} U^T B U \Sigma^{\frac{1}{2}} x = (U\Sigma^{\frac{1}{2}}x)^T B (U\Sigma^{\frac{1}{2}}x) \geq 0,$$

due to B being positive semi-definite. Since the spectral radius is lower bounded-by the induced 2-norm, with equality for symmetric matrices, there holds

$$\lambda_{\max}(AB) \leq \left|\left|AB\right|\right|_2 \leq \left|\left|A\right|\right|_2 \left|\left|B\right|\right|_2 = \lambda_{\max}(A)\lambda_{\max}(B).$$

□

Lemma D.2. *Let $A : \delta \in I \to A(\delta)$ be an analytical function of the real variable δ for some interval I and whose values are symmetric PSD matrices, and B a PSD matrix. Then $\lambda_{max}(A(\delta)B)$ is Lipschitz continuous with respect to δ on I.*

Proof. Because $A(\delta)$ is analytic and symmetric, we can rewrite it as $A(\delta) = U(\delta)\Sigma(\delta)U(\delta)^T$, where Σ is diagonal and contain the non-negative eigenvalues of $A(\delta)$, U is orthonormal, and both Σ and U are analytic functions of δ, see Chapter 2.2 in [19]. As in Lemma D.1, we see that $\lambda_{\max}(A(\delta)B) = \lambda M(\delta)$, with

$$M(\delta) := \Sigma(\delta)^{\frac{1}{2}}U(\delta)^T BU(\delta)\Sigma(\delta)^{\frac{1}{2}}$$

an analytical function of δ that is always positive semi-definite, so its eigenvalues are real. It follows then from Theorem 6.8 in [20] that the eigenvalues of M can be expressed as analytical functions of δ, and hence that $\max_i \lambda_i(M(\delta))$ is a Lipschitz-continuous function of δ on the interval I. □

Lemma D.3. *Let L be a Laplacian, i.e. $Le = 0$ and $L_{ij} = -A_{ij} \leq 0$ if $i \neq j$. We let $a_{\min} \leq A_{ij}, \forall i \neq j, A_{ij} > 0$ be a lower bound on the absolute values of the nonzero off-diagonal elements, and $\bar{L}$ the corresponding unweighted Laplacian, i.e. $\bar{L}e = 0$ and for all $i, j, i \neq j$, $\bar{L}_{ij} = -1$ if $A_{ij} > 00$ and 0 else.*

If L is symmetric, then

$$\lambda_2(L) \geq a_{\min}\lambda_2(\bar{L})$$

If DL is symmetric for some positive diagonal D whose smallest and largest diagonal entries are $d_{\min}$ and $d_{\max}$, then $\lambda_2(L)$ is real and

$$\lambda_2(L) \geq \frac{d_{\min}}{d_{\max}} a_{\min}\lambda_2(\bar{L})$$

Proof. We assume first that L is symmetric. In that case its eigenvectors are orthogonal, and since the vector e corresponds to the its eigenvalue 0, we have

$$\lambda_2(L) = \min_{e^T x = 0} \frac{x^T L x}{x^T x}$$

Using the classical expression of $x^T Lx$ for symmetric Laplacian, we see that for any x, we have

$$\begin{aligned} x^T Lx = \sum_{i<j} A_{ij}(x_i - x_j)^2 &\geq \sum_{i<j, A_{ij}>0} a_{\min}(x_i - x_j)^2 \\ = a_{\min} \sum_{i<j, \bar{L}_{ij} \neq 0} (x_i - x_j)^2 &= a_{\min} x^T \bar{L} x. \end{aligned}$$

Hence we have

$$\lambda_2(L) \geq a_{\min} \min_{e^T x = 0} \frac{x^T \bar{L} x}{x^T x} = a_{\min}\lambda_2(L).$$

We now move to the second claim. Observe that

$$\lambda_2(L) = \lambda_2(D^{-1}DL) = \lambda_2(D^{-1/2}(DL)D^{-1/2}).$$

Hence if DL and thus $D^{-1/2}(DL)D^{-1/2}$ is symmetric, $\lambda_2(L)$ is real. Moreorver, DL is a Laplacian, so e belongs to its kernel, which means that $D^{1/2}e$ is an eigenvector of $D^{-1/2}(DL)D^{-1/2}$ with eigenvalue 0. Hence

$$\begin{aligned} \lambda_2(L) = \lambda_2(D^{-1/2}(DL)D^{-1/2}) &= \min_{x: e^T D^{1/2} x = 0} \frac{x^T D^{-1/2}(DL)D^{-1/2}x^T}{x^T x} \\ = \min_{y: e^T y = 0} \frac{y^T DLt}{x^T D^{-1} x} \geq \frac{1}{d_{\max}} \min_{y: e^T y = 0} \frac{y^T DLy}{y^T y} &= \frac{\lambda_2(DL)}{d_{\max}}. \end{aligned}$$

Observe now that all nonzero off-diagonal elements of DL have an absolute value at least $d_{\min}a_{\min}$. The first claim of this lemma implies then $\lambda_2(DL) \geq d_{\min}a_{\min}$, from which the second claim follows. □

Lemma D.4. *Let $A \in \mathbb{R}^{m\times n}$ be a positive rank-1 matrix such that $A_{ij} \in [\rho^{-1}, \rho]$. Then A can be written as $A = xy^T$ for vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ such that $x_i, y_j \in [\rho^{-1}, \rho]$ for every $i \in I_x, j \in I_y$.*

Proof. Since A is rank-1 and positive it can be written as $\hat{x}\hat{y}^T$ for positive vectors $\hat{x}, \hat{y}$. We use the indices $\min, \max$ to denote the indices of the smallest and largest values of the vectors. Observe first that for an arbitrary index $j \in I_y$, we have

$$\frac{\hat{x}_{\max}}{\hat{x}_{\min}} = \frac{\hat{y}_j\hat{x}_{\max}}{\hat{y}_j\hat{x}_{\min}} = \frac{\max_{i\in I_x} A_{ij}}{\min_{i\in I_x} A_{ij}} \leq \rho^2.$$

The same argument shows $\frac{\hat{y}_{\max}}{\hat{y}_{\min}} \leq \rho^2$. We define

$$x = \frac{\rho}{\hat{x}_{\max}}\hat{x}, \qquad y = \frac{\hat{x}_{\max}}{\rho}\hat{y}.$$

There holds again $A = xy^T$. By construction $x_{\max} = \rho$, which implies $x_{\min} \geq \rho^{-1}$ as we have seen above. Moreover, $y_{\max} \leq 1$, for otherwise we would have $\max_{i,j} A_{ij} = x_{\max}y_{\max} > \rho.1$. So, if $y_{\min} \geq \rho^{-1}$, then these x, y satisfy the required condition. Otherwise, we have $\frac{\rho^{-1}}{y_{\min}} > 1$, and we define

$$x' = \frac{y_{\min}}{\rho^{-1}}\hat{x}, \qquad y' = \frac{\rho^{-1}}{y_{\min}}y,$$

satisfying again $x'(y')^T = A$. By construction $y'_{\min} = \rho^{-1}$, so that $y'_{\max} \leq \rho$. Moreover, since $\frac{\rho^{-1}}{y_{\min}} > 1$, we have $x'_{\max} \leq x_{\max} \leq \rho$. Finally,

$$x'_{\min} = \frac{x_{\min}y_{\min}}{\rho^{-1}} \geq \frac{\rho^{-1}}{\rho^{-1}} = 1,$$

so these x', y' satisfy the conditions. □

E Details of Numerical Experiments

In this section we describe the precise conditions of the numerical experiments. The matrices were generated by taking random vectors x, y, with $\log x_i, \log y_j$ uniformly distributed in $[-(\log\rho)/2, (\log\rho)/2]$. The masks Ω were generated by selecting independently each entry with an probability p. Masks that did not lead to a connected graph G were discarded.

In Figure 1(a) we used $p = .2$, and we average the results over 500 tests for each of the 25 data points. The total simulation time was 127 sec for 12500 tests with the three algorithms on 50×50 matrices, using Matlab code that is not optimized for sparsity and on a regular laptop. The random noise was obtained by adding i.i.d. random values between $-\delta/2$ and $\delta/2$. In Figure 1(b), we use $p = .5$, and average the results over 25000 tests on 10×10 matrices for each of the 31 data-points, this higher number being selected because of a larger variability of the results for small values of a^* for the (weighted) log least square algorithm. The total computation time was 337 sec for 775000 matrices. Further tests not reported here showed that 200×200 matrices with 10% of revealed entries could be recovered in an average of .6 sec, still without optimizing for sparsity. By comparison the method in [2] would have required solving an SDP on a matrix 160000×160000.

The third algorithm, denoted "Log LS unweighted" consists in directly solving the system

$$\log\hat{x}_i - \log\hat{y}_j = \log A^R_{ij}$$

in the least square sense. Its analysis can be conducted in a way similar to our weighted algorithm, but it does not yield asymptotically optimal guarantee. Moreover, as explained in Section 4, the different entries have different sensitivities to perturbations, so this method is more exposed to targeted perturbations.

Our implementation of the ridge-regression consisted in minimizing

$$\left|\left|(A^R - xy^T)_\Omega\right|\right|_F^2 + \lambda(||x||_F^2 + ||y||_F^2)$$

using a gradient descent, projecting x and y at each step on the set $[\mu\rho^{-1}, \mu\rho]$ to which we know the real values belong. Different values of λ were tried. The gradient iterations were interrupted

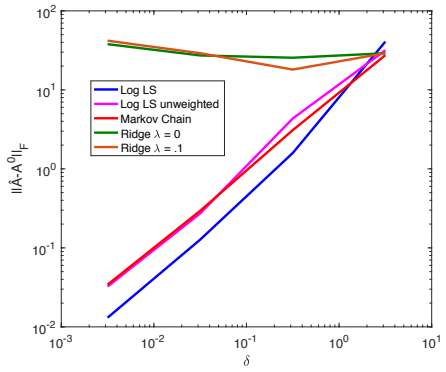


Figure 2: Evolution with δ of the average error $||\hat{A} - A^0||_F$ for our two algorithms, an unweighted version of the algorithm of Section 4, and our implementation of the ridge regression with $\lambda = 0$ (no regularization) and $\lambda = .1$, in a scenario where all revealed entries are perturbed by a random noise of magnitude $\delta/2$ (50×50 matrices with on average 50% of revealed entries). Initial matrices have entries between 10^{-1} and 10. Large errors are observed for the ridge regression methods, even for very small values of δ.

when $||x(k+1) - x(k)||_1 + ||y(k+1) - y(k)||_1 \leq 10^{-12}$ or after 200000 steps. Each problem was solved using 10 different initial $x(0), y(0)$, with values randomly selected in $[\mu\rho^{-1}, \mu\rho]$, and the best final iterate (in term of the objective function) was kept. Examples of results are presented in Figure 2, for the same experimental conditions as in Figure 1(a), except that results are averaged over 5 tests for each data point. We further note that large errors for small values of δ were consistently obtained on every single one of the realizations.