Minimax Rank-1 Matrix Factorization

Author 1
Institution 1

Author 2
Institution 2

Author 3
Institution 3

Abstract

We consider the problem of recovering a rank-one matrix when a perturbed subset of its entries is revealed. We propose a method based on least squares in the log-space and show its performance matches the lower bounds that we derive for this problem in the small-perturbation regime, which are related to the spectral gap of a graph representing the revealed entries. Unfortunately, we show that for larger disturbances, potentially exponentially growing errors are unavoidable for any consistent recovery method. We then propose a second algorithm relying on encoding the matrix factorization in the stationary distribution of a certain Markov chain. We show that, under the stronger assumption of known upper and lower bounds on the entries of the true matrix, this second method does not have exponential error growth for large disturbances. Both algorithms can be implemented in nearly linear time.

1 Introduction

We consider the problem of finding a rank-one approximation xy^T of a matrix $A \in \mathbb{R}^{m\times n}$ when only a subset Ω of the entries of A are revealed. We do not impose any stochastic assumptions on the support set Ω (i.e., the entries in Ω do not need to be randomly chosen) nor assume any structure on the underlying matrix A. We are looking for stable algorithms: this means that if what is revealed is not $\{A_{ij} \mid (i,j) \in \Omega\}$ but rather the perturbed entries $\{A_{ij} + \Delta_{ij}\}$, then we need to be able to bound the error in the recovered matrix as a function of the size of the perturbations $||\Delta||_F := \sum_{(i,j)\in\Omega} \Delta_{ij}^2$.

In particular, we are interested in analyzing this question in a minimax framework. We would like to understand, for a given error size $||\Delta||_F$, how large can the error in recovering A can be. Moreover, we would like to know which algorithm(s) can always guarantee this minimax level of performance. Finally, we would like to understand how these quantities depend on the support set Ω.

We will be making only the minimal requirement on the support Ω: the only condition we will impose is that the true matrix A be identifiable, meaning that it is possible in principle to complete the matrix A from the unperturbed entries $\{A_{ij} \mid (i,j) \in \Omega\}$. These conditions were worked out in [1, 2, 3] and we describe them next.

First, it is natural to assume that A should have no zero entries. Indeed, if A has zero entries, then it may be impossible to complete A even from a very large number of unperturbed revealed entries Ω. For example, consider the case where every matrix entry except the $(1,1)$ entry is revealed, and equals zero; it is impossible to know what A_{11} is even though the set of revealed entries is only a single entry away from being complete.

Provided that A has no zero entries, a simple graph-theoretic condition exists for identifiability. Specifically, we associate the support set Ω with an undirected bipartite graph, G, with node set $\mathcal{I}_x \cup \mathcal{I}_y$, where $I_x = \{1,\dots,m\}$ and $I_y = \{1,\dots,n,\}$ (recall that $A \in \mathbb{R}^{m\times n}$), with nodes $i \in I_x$ and $j \in I_y$ connected if the ijth entry is an element of Ω. If A has no zero entries, identifiability is equivalent to the connectivity of this bipartite graph (see [1] as well as discussion in [3]). *Thus, henceforth it will be standing assumptions that A has no zero entries and that G is connected.*

Our motivation stems from several practical applications ranging from worker skill estimation in crowdsourcing [2, 4, 5, 6, 7], inferring latent information from limited observations in collaborative filtering and recommender systems [8], and in other matrix completion applications such as global positioning and system identification [9], all of which can be formulated in terms of rank-1 matrix completion.

Preliminary work. Under review by AISTATS 2020. Do not distribute.

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 [8, 9, 10, 6, 7, 11, 2, 12, 3, 4, 13]. Apart from [3, 4], much of this literature assumes some form of incoherence on the matrices, a probabilistic model for Ω, or other structures on what indices of Ω are revealed. *This separates it from the present work, which does not use any of these assumptions.*

Many of these methods [9, 10, 6, 12, 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 [3, 4] we impose no such structure and so such spectral properties can no longer be leveraged for recovery.

In the unperturbed case, it is easy to complete a matrix A_{ij} from a subset of revealed entries using a "propagation" approach; this consists in fixing one entry, say $x_1 = 1$, and then solving for entries of y from the revealed entries in the first row, and then iterating this scheme as more entries are fixed. In particular, this was discussed in [2]. Unfortunately, [3] points out that this technique performs very poorly in the presence of perturbations even on some very simple examples. Relatedly, [13] 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 [9]. 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 [3]. Ridge-regression based approaches have also been considered [4, 14], and appear natural for our setting, since Tikhonov regularization typically provides stable solutions. Nevertheless, as pointed out in [3], even these approaches are unstable. Moreover, they require solving non-convex optimization problems with a potentially high number of local minima of the form

$$\min_{x,y} \left|\left|(xy^T - A^R)_\Omega\right|\right|_F + \lambda(||x||_2 + ||y||_2), \quad (1)$$

where Ω selects the entries for which data is available, and A^R denotes the revealed entries. More recently alternative minimization (ALM) methods, wherein the two vectors x and y are alternately updated to optimize $\left|\left|(xy^T - A^R)_\Omega\right|\right|_F$, have been proposed to handle the computational bottleneck of optimizing over low-rank matrices. [11], following up on a long line of works establishes recovery guarantees, under strong-coherence assumptions. As these authors point out, ALM methods leverage the key property "that the spectral property only need to hold in an average sense" to guarantee recovery of ALM based methods. In contrast to these methods, we impose no constraints either on the ground-truth matrix or assume random sampling of revealed entries.

Our work is closely related to [3]. Like that work, we require our algorithms be stable. On the other hand, we differ from [3]'s method in several ways. First, [3] 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, [3] 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. As we will discuss later in the paper, our methods actually run in linear time (up to log factors), and are thus considerably faster than methods based on SDP relaxation.

2.1 Our contributions

In this paper, we develop two efficient and stable approximation methods for rank-one estimation of a partially observed matrix.

Our first scheme is based on formulating the problem as weighted least squares in a certain logarithmically transformed space. *Our first contribution is to demonstrate that, for small perturbations, the performance of this scheme matches 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.

Unfortunately, the recovery error of the weighted log-least squares method will scale exponentially in $||\Delta||_F$. While this may be acceptable for small perturbations Δ, it makes the minimax performance quite poor if $||\Delta||_F$ is not small. Unfortunately, *our second contribution is to show that this is unavoidable.* Specifically, we consider the class of consistent algorithms, defined those methods that require correct recovery of A when $\Delta_{ij} = 0$ for all $(i, j) \in \Omega$. We show that any consistent scheme must suffer an estimation error that scales exponentially in $||\Delta||_F$.

This negative result leads us to consider a minor modification of our problem. Specifically, we consider the setting where we additionally know upper and lower bounds on the entries of A. We propose a method, based on the encoding of the rank-one factors x and

y (from the decomposition $A = xy^T$) into a stationary distribution of a suitable Markov chain, and whose parameters leverage these known lower and upper bounds. *Our final contribution is to give an estimate of the recovery error associated with this method and show that it does not scale exponentially in* $||\Delta||_F$.

3 The first algorithm: weighted log-least squares

We begin with a heuristic derivation of our method. Let us first consider the unperturbed case, i.e., when $\Delta_{ij} = 0$ for all $(i, j) \in \Omega$.

We begin with the observation that it suffices to deal with the case where A is positive. Indeed, if $A = xy^T$ is rank-one with no zero entries, then the same holds for $|A| = |x||y^T|$. Any method which works in the positive case can be used to compute $|A|$ by taking the absolute value of the revealed entries $\{|A_{ij}| \mid (i, j) \in \Omega\}$. Having obtained the full-matrix A, we can then easily compute $|x|$ and $|y|$ by a standard rank-1 factorization. Finally, we can use "sign propagation" to figure out the sign of all the entries of x and y: we fix $\text{sign}(x_1) = +$, and repeatedly figure out the signs other elements of x and y by inspecting the revealed entries A_{ij}. It is easy to see that this process will work provided the graph G is connected: specifically, the process will result either in the recovery of x and y, or in the recovery of $-x$ and $-y$, which amounts to the same thing.

Now in the case the rank-1 matrix $A = xy^T$ is positive, it follows that x and y can be taken to be positive vectors. We define

$$\begin{aligned} z_i &= x_i \text{ for all } i \in I_x \\ z_j &= y_j^{-1} \text{ for all } j \in I_y \end{aligned}$$

In terms of these new variables we have that

$$A_{ij} = z_i / z_j \text{ for all } (i, j) \in \Omega,$$

or

$$\log A_{ij} = \log z_i - \log z_j \text{ for all } (i, j) \in \Omega. \qquad (2)$$

The assumption of positivity of A_{ij} was necessary in order to be able to take logarithm in the last equation. We now observe that these equations are linear in the quantities $\log z_i$. This leads to a natural idea: we can solve the linear system of Eq. (2) for the quantities $\log z_i$, and then find x, y by exponentiating.

We now turn to the discussion of the case where the perturbations Δ_{ij} are not necessarily zero. Provided the perturbations are not so large as to change the sign of the entries, we can proceed as before by taking the absolute values of the revealed matrix and recovering the signs during post-processing using a sign-propagation step.

However, the problem that arises is that simply solving Eq. (2) is not the best thing to do anymore, because different entries of the matrix display different levels of senstivity to perturbations. Indeed, observe that if we solve

$$\begin{aligned} \log z_i - \log z_j &= \log(A_{ij} + \Delta_{ij}) \\ &= \log A_{ij} + (\log(A_{ij} + \Delta_{ij}) - \log A_{ij}) \\ &:= \log A_{ij} + D_{ij} \end{aligned}$$

then we see that the same disturbance Δ_{ij} might create a larger or smaller D_{ij} depending on the matrix entry A_{ij}. More specifically, if A_{ij} is smaller, a disturbance Δ_{ij} of the same magnitude can result in a larger D_{ij}. Informally speaking, an adversary with a fixed budget for the disturbance might choose to perturb smaller entries.

Our approach to deal with this is to re-weigh the equations so that an adversary could not take advantage of this, at least when the perturbations Δ_{ij} are small relative to A_{ij}. Indeed, observe that using first-order approximations

$$\begin{aligned} \frac{D_{ij}}{\Delta_{ij}} &= \frac{\log(A_{ij} + \Delta_{ij}) - \log A_{ij}}{\Delta_{ij}} \\ &\approx \frac{1}{A_{ij}} \\ &\approx \frac{1}{A_{ij} + \Delta_{ij}}, \end{aligned}$$

where the first approximation used that $(\log x)' = 1/x$ while the second approximation used that Δ_{ij} should be small.

This string of equations leads to a natural heuristic: we can simply multiply the equation in Eq. (2) corresponding to (i, j) by the revealed entry $A_{ij} + \Delta_{ij}$. If we do that, small perturbations Δ_{ij} will have the same effect on every equation. Thus adopting the shorthand

$$A^R_{ij} := A_{ij} + \Delta_{ij} \text{ for all } (i, j) \in \Omega,$$

where the superscript "R" stands for "revealed", our algorithm is to solve the system of equations

$$A^R_{ij}(\log z_i - \log z_j) = A^R_{ij} \log A^R_{ij}, \qquad \forall (i, j) \in \Omega \quad (3)$$

in the least square sense. Naturally, this is not the same as solving Eq. (2) in the least-squares sense, since multiplying the (i, j)'th equation by A^R_{ij} effectively "weights" each equation in Eq. (2) differently.

Computational Efficiency. We conclude this section by explaining how this system of equations is a Laplacian linear system, which allows us to leverage existing results to show it can be solved in time nearly

linear in the size of Ω. We begin by introducing a particular weighted version of the bipartite graph G: G_{W^R} has the same node set and edges as G with the weight of the edge (i,j) being $(A^R_{ij})^2$. Let L_{W^R} be the Laplacian of this weighted graph, and let B be its incidence matrix. It is standard that

$$L_{W^R} = BW^RB^T, \tag{4}$$

where $W^R \in \mathbb{R}^{|\Omega|\times|\Omega|}$ a diagonal matrix collecting all the weights $(A^R_{i,j})^2$. The linear system of Eq. (3) can then be expressed as

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

where A^R_Ω denotes the vector collecting the revealed entries A^R_{ij}. Using Eq. (4), least squares solutions of this systems are solutions of the linear system

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

For example, one least-squares solution is

$$\log \hat{z} = L^\dagger_{W^R} BW^R \log A^R_\Omega \tag{5}$$

where † denotes the Monroe-Penrose inverse.

Computational Efficiency. The main advantage of this rewriting is that it now follows from the now-classic results of Spielman and Teng [15] that Eq. (5), being a system of equations with a graph Laplacian, can be solved in near linear time (up to log terms) in the size of Ω. More precisely, a solution with precision ϵ can be obtained in $O(|\Omega| \log^\kappa(n+m)\log(\epsilon^{-1}))$ operations for some constant $\kappa > 0$.

The pseudocode of the weighted log-least squares method is given below.

Algorithm 1 Weighted Log-Least Squares Method

1: **Input:** Positive revealed entries $\{A^R_{ij} \mid (i,j) \in \Omega\}$.
2: Solve Eq. (5) for $\hat{z}$.
3: For $i \in I_x$, set $\hat{x}_i = z_i$ and for $j \in I_y$, set $\hat{y}_j = z_j^{-1}$.
4: Return $\hat{A} = \hat{x}\hat{y}^T$.

3.1 Accuracy Results

We now move to a discussion of our results. Our first theorem gives an error bound for the performance of the weighted log-least squares method.

Theorem 1. *Suppose that A is positive and*

$$A_{ij} + \Delta_{ij} > 0 \text{ for all } (i,j) \in \Omega.$$

Suppose that the disturbances further satisfy

$$\Delta_{ij} \le (c-1)A_{ij} \text{ for } c > 1 \text{ and all } (i,j) \in \Omega.$$

Then the weighted-log least squares method returns an estimate $\hat{A}$ satisfying the error bound

$$\left|\left|\hat{A} - A\right|\right|_F^2 \le c^2 \lambda_{\max}(K_W L^\dagger_{W^R})\, ||\Delta||_F^2\, e^{2c\sqrt{R_{W^R,\max}}||\Delta||_F},$$

where $R_{W^R,\max}$ is the largest pairwise resistance between any pair of nodes in the weighted bipartite graph G_{W^R}, and K_{W^R} is the Laplacian of the complete bipartite graph on $I_x \times I_y$ where the edge of weight (i,j) is the true entry A_{ij}.

To parse this theorem, note that the positivity of $A_{ij} + \Delta_{ij}$ effectively bounds Δ from below, while the condition $\Delta_{ij} < (c-1)A_{ij}$ bounds it from above. The latter condition is just another way of stating that the revealed entry $A_{ij} + \Delta_{ij}$ is not more than c times the actual entry A_{ij}. Finally, as discussed earlier, we can consider the positive case without loss of generality due to the trick of taking the absolute value of revealed entries and using sign propagation.

For a formal definition and discussion of the electrical resistance of graphs, we refer the reader to [16]. Informally, the resistance of a graph is defined as the largest resistance in an electrical circuits where each edge is replaced by a resistor with resistance *inversely* proportional to the weight of that edge.

The assumption that the revealed entry A^R_{ij} has the same sign as A_{ij} is unavoidable. To see why, consider the rank-1 matrix

$$A = \begin{pmatrix} \epsilon & \epsilon \\ \epsilon & \epsilon \end{pmatrix}$$

Supposing that

$$A^R = \begin{pmatrix} \epsilon^2 & -1 \\ -1 & * \end{pmatrix}$$

where the star represents unrevealed entries, we see that the revealed entries of A^R are consistent with the matrix

$$\begin{pmatrix} \epsilon \\ -\epsilon^{-1} \end{pmatrix} \begin{pmatrix} \epsilon & -\epsilon^{-1} \end{pmatrix}$$

Thus even though $||\Delta||_F$ is constant, the recovery error will scale at least as ϵ^{-2}. Choosing a sufficiently small ϵ, we can obtain an arbitrarily large error.

We note that the necessity of the same-sign assumption is not particularly dependent on the choice of method, as this pair of matrices is a counterexample for all methods which return a rank-1 matrix completion of A^R whenever it is available (in our next section, we will prove lower bounds on the performance of such methods, which we call *consistent*).

When $||\Delta||_F$ is small, both the exponential factor and the constant c in the above theorem approach one, so

that we have

$$\lim_{||\Delta||_F \to 0} \frac{||\hat{A} - A||_F^2}{||\Delta||_F^2} \leq \lambda_{\max}(K_{W^0} L^{\dagger}_{W^R}). \qquad (6)$$

Theorem 1 thus identifies the key graph-theoretic quantity that governs robustness in the small-perturbation regime. Because it may be difficult to trace how this quantity scales in the number of nodes or other graph-theoretic quantities, we provide the following corollary which gives a bound in terms of the more usual graph characteristics.

Corollary 1. *Let $\bar{\alpha}$ be an upper bound on the entries of the matrix A and let $\underline{\alpha}^R$ be a lower bounds on the revealed entries A^R_{ij}. Under the same assumption as in Theorem 1, the estimate produced by the weighted log-least squares method satisfies*

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

where $R_{\max}$ is the maximal pairwise resistance in the unweighted bipartite graph G, $\lambda_2(L)$ is the second-smallest eigenvalue of the Laplacian L corresponding to this graph, and, as before, $A \in \mathbb{R}^{m \times n}$.

4 Lower bounds

It is natural to wonder to what extent the upper bounds we have derived in the previous section is optimal. The following theorem considers the limiting case when the perturbation is small. We provide a lower bound under the assumption that the algorithm only uses the revealed entries A^R_{ij} to compute an estimate $\hat{A}$. Note that this assumption applies to the weighted log-least squares method, but will be violated by the algorithm we will propose in the next section.

Theorem 2. *Consider any algorithm that computes an estimate $\hat{A}$ of A based solely on $\{A^R_{ij} \mid (i,j) \in \Omega\}$. Then for any entry-wise positive rank-1 matrix A^R and mask Ω, one can find a matrix A such that*

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

with $\Delta_{ij} = A^R_{ij} - A_{ij}$ for $(i,j) \in \Omega$.

Combining this theorem with Eq. (6), *we obtain our first main result: that the weighted log-least squares method is optimal for small disturbances, and that the relevant graph-theoretic quantity governing performance is $\lambda_{\max}(K_W L^{\dagger}_{W^R})$.*

We next turn to the question of what happens when disturbances are not small. Inspecting Theorem 1, we see that the error grows exponentially in the size of the disturbance $||\Delta||_F$. There is also an exponential scaling in terms of the largest resistance in the graph G with weights coming from W^R. The latter is also concerning, since resistances will often scale polynomially in the number of nodes (for example, on a line of n nodes resistance is linear in n). However, because the resistance of a weighted graph scales inversely in the weights, the upper bound will also blow up for any class of problems where the smallest revealed entry goes to zero.

It is natural to ask whether these poor scalings are avoidable. Unfortunately, our next main result answers this negatively under a plausible assumption.

That assumption is *consistency*, which says that when the revealed entries are the unperturbed entries of a rank-1 matrix A, the algorithm should recover A exactly. Consistency is a natural conditions for algorithms that estimate A based purely on revealed entries. Looking forward, however, we note that consistency is *not* a natural assumption for algorithms that are allowed to use additional information. For example, later on in the paper we will consider algorithms that know upper and lower bounds on the entries of A. Indeed, when revealed entries of a rank-1 matrix A with the property that some entries of A lie outside of these bounds, such algorithms should not simply return A.

Our second main result, presented as the following theorem, says that, in the worst-case, any consistent method suffers from exponentially poor scaling in $||\Delta||_F$ and the largest resistance of G.

Theorem 3.
(a) Fix any positive constant c. There exists a family of matrices $A_n \in \mathbb{R}^{n \times n}$, support sets $\Omega_n \subset \{1,\ldots,n\} \times \{1,\ldots,n\}$, and disturbances Δ_n satisfying

$$||\Delta_n|| \leq c,$$

with uniformly bounded A^R_{ij} and A_{ij}, for which we have

$$\left|\left|\hat{A} - A\right|\right|_F^2 \geq \left(\exp\left(c\sqrt{R_{\max}\left(\frac{1}{2} - O(n^{-1/2})\right)}\right) - 1\right)^2$$

for any consistent algorithm. Here $R_{\max}$ is the largest pairwisse resistance of the (unweighted) graph G.

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

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

5 The second algorithm: Markov chain stationary distributions

We now provide an algorithm that avoids the exponential scaling discussed in the previous section. The reason this does not contradict Theorem 3 is that we now assume we have lower and upper bounds on the entries of A: $\underline{\alpha} \leq (A)_{ij} \leq \overline{\alpha}$ for all $i \in I_x, j \in I_y$, and these quantities $\underline{\alpha}, \overline{\alpha}$ are known to the algorithm. For small disturbances, however, the guarantees on performance of this method will, in the worst-case, be weaker than the asymptotically optimal algorithm in Section 3.

In the sequel, we will find it convenient to define the quantities $\mu = \sqrt{\bar{\alpha}\underline{\alpha}}$ and $\rho = \sqrt{\bar{\alpha}/\underline{\alpha}}$, so that the interval $[\underline{\alpha}, \bar{\alpha}]$ can be re-expressed as $[\mu\rho^{-1}, \mu\rho]$.

5.1 Algorithm Description

Since we know that every $(A)_{ij}$ lies in $[\underline{\alpha}, \bar{\alpha}]$, we will begin by projecting all revealed entries on that interval. Note that this step can only reduce the disturbances.

The algorithm consists in computing the stationary distribution of a continuous time Markov chain defined on the graph G. Specifically, we define the matrix $M^R \in \mathbb{R}^{(m+n)\times(m+n)}$ as

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

and set all other entries are 0. The matrix M is defined in the same way, replacing A^R by A. For background on continuous-time Markov chains, we refer the reader to [16].

The motivation for this method is captured by the following simple observation. Recalling that $A = xy^T$, it turns out that the vector z defined as

$$
z_i = \frac{x_i}{\sqrt{\mu}} \text{ for } i \in I_x, \;\; z_j = \frac{\sqrt{\mu}}{y_j} \text{ for } j \in I_y
$$

is proportional to the stationary distribution of the continuous time Markov chain M. This fact can be verified immediately by observing that the "balance equations" $M_{ij} z_i = M_{ji} z_j$ hold.

In other words, in the case where the perturbations are zero, computing the stationary distribution of M immediately lets us recover x and y, and therefore the matrix A. When the perturbations are nonzero, one might hope that the stationary distribution of M^R will depend smoothly on the perturbations Δ, so that it will be possible to bound the recovery error.

This trick is very similar to the approach used in [17, 18] for the problem of estimating an unknown set of weights from a collection of noisy pairwise comparisons. One difference is that we add a projection step; this seemingly minor difference allows us to bypass the lower bounds of Theorem 3. The pseudocode for the method is given below.

Algorithm 2 Projected Eigenvector Algorithm

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

Finally, the stationary distribution π^R is simply the eigenvector of M^R corresponding to the zero eigenvalue. It can be computed in nearly-linear time as a consequence of the recent results of [19].

5.2 Accuracy Results

The accuracy guarantee of this algorithm are naturally expressed in terms of total variation (i.e., l^1) norm rather than a quadratic loss. We thus introduce the "first-order Frobenius norm" defined as

$$
||M||_{F:1} = \sum_{i,j} |M_{ij}| .
$$

Note that $||M||_F \leq ||M||_{F:1}$, so any upper bound on the first-order Frobenius norm is also an upper bound on the ordinary Frobenius norm.

The error guarantee for the projected eigenvector method is given in the following theorem.

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

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

where $\rho = \bar{\alpha}/\underline{\alpha}$ and $\lambda_2(M)$ is the second-smallest eigenvalue of M, which is real.

The previous theorem implicitly depends on the revealed submatrix A^R, since the matrix M is built from

A^R. The following corollary provides a bound which depends only on the graph and not the revealed entries.

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

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

Since it is elementary that $\lambda_2(L)^{-1}$ is at most polynomial in $m+n$ (for example, the bound $\lambda_2(L)^{-1} \leq O\left((m+n)^2\right)$ follows from Theorem 6.2 in [20]) we come to *our third result: the projected eigenvector method avoids the exponential scaling faced by all the consistent methods.*

6 Synthetic Experiments

In this section, we conduct a number of synthetic experiments to highlight similarities and differences between the algorithms proposed here and prior works.

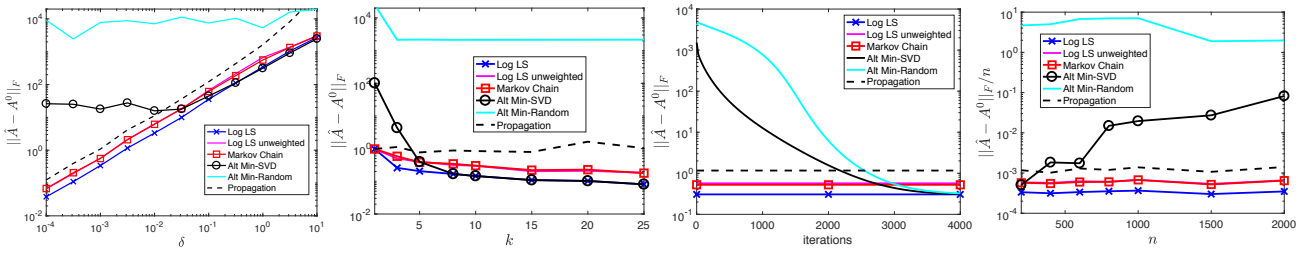
Metrics. Recall that our goal is to approximate the unknown ground-truth rank-one matrix A^0 by a rank-one matrix xy^T, when the observations are perturbed, $A^R = A_0 + \Delta$, $\frac{1}{n}\|\Delta\|_F \leq \delta$, and the revealed entries are sampled components of A^R. We will demonstrate that, under various scenarios, such as different perturbation levels (δ); different samplings of revealed entries: either random or structured; different matrix sizes, either small or large; our algorithm recovers a *stable* solution rapidly with computational time scaling with the number of revealed entries.

Competing Rank-One Approximation Methods. As pointed in our related work, nuclear norm based methods [3] do not guarantee rank-one recovery and so we omit them in our experiments. Ridge-regression Eq. 1 is demonstrably unstable (see Appendix Fig. 4). In summary, we are left with three algorithms, propagation [2], Alternative Minimization with/without clipping and with SVD and with random initialization [11], and our weighted Log-LS method, unweighted Log-LS method, and the Markov chain method, which we report here. Clipping [11] was found to be sub-optimal, and in general Alternative minimization without clipping out-performed clipping for our rank-one setting. For this reason we omitted results with clipping. We report results for Alternative Minimization as Alt-min-SVD and Alt-min-rand.

Datasets.
Ground truth matrices A^0: The matrices were generated by taking random vectors x, y, with $\log x_i, \log y_j$ uniformly distributed in $[-(\log\rho)/2, (\log\rho)/2]$, for $\rho = 10$. As a consequence, all entries lie in $[10^{-1}, 10]$. Unless stated otherwise, matrices were of size 1000×1000.

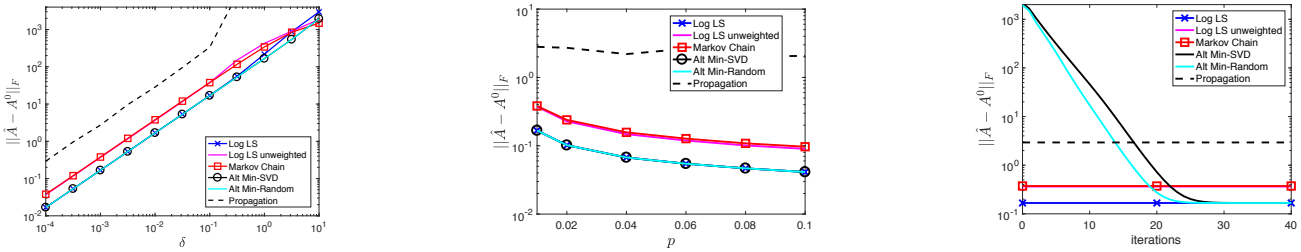
Perturbation on revealed entries (A^R, Δ). The perturbation applied to revealed entries consist of i.i.d. random variable uniformly distributed in $[-\delta/2, \delta/2]$, with $\delta = 10^{-3}$ unless stated otherwise. Other typical perturbations were not observed to lead to significantly different behaviors.
Masks and revealed entries. Two sort of masks were considered. (i) **Random mask:** each entry is revealed or not according to i.i.d. random variable of probability p, equal to 0.01 unless stated otherwise. This corresponds thus to an average of 10 revealed entries per row or column. Masks that did not lead to a connected graph G were discarded, as this is a necessary condition to reconstruct A^0 even in the absence of perturbation. (ii) **Star mask:** the first k rows and columns are revealed. Note that revealing a single row or column corresponds to a star graph, and is the minimal number of elements required to complete a rank-one matrix [4]. Unless otherwise specified, we conduct experiments with 1% revealed entries, and 3-rows/columns for Star masks.

Experiments and Key Findings. All of the reported results are mean values averaged over 50 trials. We found Alt-Min-SVD and Alt-min-rand exhibited high variance over the course of the iterations. We report variances in the appendix (see Fig. 5).
Effect of Perturbation Size. We first determine how the different methods scale with size of perturbation under different structures (random & star), with fixed number of revealed entries ($\sim 1\%$) for a matrix of size 1000×1000, for 1000 Alt-min iterations. Figures 1(a) 2(a) reveals that, unsurprisingly, accuracy degrades with perturbation size. For small perturbation, we notice that proposed weighted Log-LS dominates our other proposals. Nevertheless, for large perturbations, Markov-Chain error saturates and thus performs better than competing methods. Propagation performs poorly against other competing methods even with relatively small perturbation. Alt-min and Log-LS achieve somewhat the same accuracy with 1% randomly revealed entries (see Fig. 2(a)).
Influence of Sampling. Figures 1(a) 2(a) provides a qualitative comparison between random vs. structured star graph sampling. While Alt-min performs well with random sampling, its performance significantly degrades with structured star-masks. This points to the fact that Alt-min performs well only when the spectral properties are preserved [11]. Propagation is somewhat robust to different types of sampling.
Effect of # Revealed Entries. We varied number of revealed entries, keeping all other variables (matrix size, perturbation level, # Alt-min iterations) fixed for random and star-mask sampling (see Fig. 1(b) 2(b)). For star-mask we varied the number of rows/ columns. Both Alt-min-svd and Alt-min-rand initially performed



(a) Perturbation Size (b) # Revealed Columns. (c) Convergence. (d) Matrix Size.

Figure 1: Influence of "star-graph" sampling on proposed and competing methods. On Star graphs our Log-LS dominates other methods with perturbation size, # Revealed Columns, Convergence Speed, and Matrix Size. For each experiment, when one of the variables was varied (for instance perturbation) then other variables were fixed (with matrix size $n = 1000$, 3-column/row star-graph sampling, and 1000 iterations for Alt-min methods).



(a) Perturbation Size (b) # Revealed Fraction. (c) Convergence.

Figure 2: Comparisons for ideal random sampling scenario. Matrix size is omitted as no discernible features between Alt-min and Log-LS were found. Experiments were conducted as in Fig. 1, by varying one parameter and holding other parameters fixed. Qualitatively Log-LS and Alt-Min perform similarly. In contrast to star-graph sampling Alt-min converges rapidly to optimal solution.

worse than proposed methods. Furthermore, Alt-min-rand, could not stably recover the ground-truth even with sufficiently large number of revealed entries. In contrast, Alt-min performed as well as our Log-LS method for random sampling. This is not surprising, since the assumptions for Alt-min are satisfied.

Computational Scaling & Convergence. We refer to Sec. 3 for details on computational scaling of our proposed methods, where we claimed linear scaling with number of revealed entries. Propagation method has a similar scaling. In contrast, Alt-min is iterative, and for each iteration, scales at least linearly with number of entries. For this reason, we also conduct experiments to compare convergence speed for the various algorithms. Fig. 1(c) reveals that Alt-min converges relatively slowly, and exhibits high variance under star-mask sampling. It is indeed surprising that it takes over 30 iterations to converge even under the ideal random-sampling scenario (Fig. 2(c)).

Matrix Size. We also experimented with matrix size ranging from small values to 4000×4000 size matrices. Results are presented in Fig. 1(d). Surprisingly, both Alt-Min-SVD and Alt-min-Rand degrades with size of the matrix, when all other parameters are kept constant, while proposed method and propagation are robust to matrix size.

7 Conclusions

We have presented two different algorithms for rank-1 approximation based on a set of revealed entries. Both algorithms are computationally very efficient, in that a nearly linear time implementation exists. Our first method, based on weighted log least-squares, was shown to achieve the minimax bound for small disturbances. Unfortunately, it scales exponentially in the size of the disturbance for large disturbances. We show that this is unavoidable because any consistent algorithm has this property. Finally, our last algorithm avoids this exponential scaling by further assuming lower and upper bounds on the entries of the matrix are known. However, its performance guarantees are worse for small disturbances. We then conducted a number of synthetic experiments to highlight salient aspects of our method relative to competing methods. We showed that both in ideal and non-ideal sampling situations, with other varying parameters, such as matrix size and perturbation size, our method is computationally efficient and statistically stable.

References

[1] Franz J Király, Louis Theran, and Ryota Tomioka. The algebraic combinatorial approach for low-rank matrix completion. *The Journal of Machine Learning Research*, 16(1):1391–1436, 2015.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[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] Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 2016.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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

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

[19] 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.

[20] Bojan Mohar. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2(871-898):12, 1991.

[21] Nisheeth K Vishnoi. $Lx = b$. *Foundations and Trends in Theoretical Computer Science*, 8(1–2):1–141, 2013.

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

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

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

Supplementary Material:

A Proof of Theorem 1 and Corollary 1

We now turn to the proof of Theorem 1. 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. For example, A^R_Ω is thus a vector containing all the revealed entries, while A_Ω contains the real values of the entries that have been revealed. The 2-norm of such vectors is equivalent to their Frobenius norm; for example,

$$||\Delta||_F = \left|\left|A^R_\Omega - A_\Omega\right|\right|_F = \left|\left|A^R_\Omega - A_\Omega\right|\right|_2 .$$

We begin by deriving an expression for the logarithmic error $\log \hat{A}_{ij} - \log A_{ij}$, which we will need both as intermediate step towards our final bound. Before stating this equality, we recall our notation

$$D = \log A^R_\Omega - \log A_\Omega = \log(A_\Omega + \Delta) - \log A_\Omega.$$

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

$$\log \hat{A}_{ij} - \log A_{ij} = (e_i - e_j)^T L^\dagger_{W^R} B W^R D.$$

Proof. Eq. (2) may be rewritten as

$$B^T \log z = \log A_\Omega,$$

where, recall, B is the edge-vertex incidence matrix of the graph. Recalling Eq. (4) we obtain as a consequence that

$$L_{W^R} \log z = B W^R \log A^R_\Omega.$$

One solution of this system is $z = L^\dagger_{W^R} B W^R \log A_\Omega$ where $\dagger$ represents the Moore-Penrose pseudoinverse.

Recall that $\hat{z}$ is the solution constructed by our algorithm (see Eq. (5)), and we therefore have

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

Hence using again $\log A_{ij} = \log z_i - \log z_j$, we have that

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

□

Our next step is to bound how much effect the perturbation Δ can have in terms of the resulting perturbation D in the "log space."

Lemma A.2. *If $\Delta_{ij} \le (c-1) A_{ij}$ for every $(i,j) \in \Omega$ for some $c \ge 1$, then*

$$\left|A^R_{ij} D_{ij}\right| \le c \left|\Delta_{ij}\right| ,$$

for every $(i,j) \in \Omega$ and

$$\left|\left|(W^R)^{1/2} D\right|\right|_F \le c \left|\left|\Delta\right|\right|_F .$$

Proof. By concavity of the logarithm, we have

$$\begin{aligned}\left|\log A^R_{ij} - \log A_{ij}\right| &\leq \left|A^R_{ij} - A_{ij}\right| \max\left(\frac{1}{A^R_{ij}}, \frac{1}{A_{ij}}\right) \\ &= |\Delta_{ij}| \max\left(\frac{1}{A^R_{ij}}, \frac{1}{A_{ij}}\right)\end{aligned}$$

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

$$\begin{aligned}\left|A^R_{ij} D_{(i,j)}\right| &= A^R_{ij} \left|\log A^R_{ij} - \log A_{ij}\right| \\ &\leq |\Delta_{ij}| \max\left(\frac{A^R_{ij}}{A^R_{ij}}, \frac{A^R_{ij}}{A_{ij}}\right) \\ &\leq c\,|\Delta_{ij}|\,,\end{aligned}$$

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 $(A^R_{ij})^2$. □

Our next lemma provides a first bound on the "logarithmic error." We also include an estimate on how big the (unrevealed) entries $\hat{A}_{ij}$ can get, which we will use in the sequel.

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

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

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

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

Proof. Let us introduce the notation

$$Q_{ij} = W^{R\frac{1}{2}} B^T L^\dagger_{W^R} (e_i - e_j)(e_i - e_j)^T L^\dagger_{W^R} B W^{R\frac{1}{2}}.$$

Then, using that $\log \hat{A}_{ij} - \log A_{ij}$ is a scalar and Lemma A.1, we have

$$\begin{aligned}(\log \hat{A}_{ij} - \log A_{ij})^2 &= (\log \hat{A}_{ij} - \log A_{ij})^T (\log \hat{A}_{ij} - \log A_{ij}) \\ &= D^T W^{R\frac{1}{2}} Q_{ij} W^{R\frac{1}{2}} D.\end{aligned} \tag{9}$$

where, recall, $L_{W^R} = BW^R B^T$. This implies that

$$(\log \hat{A}_{ij} - \log A_{ij})^2 \leq \left|\left|W^{R\frac{1}{2}} D\right|\right|_F^2 \lambda^{ij}_{\max}, \tag{10}$$

where $\lambda^{ij}_{\max}$ is defined as

$$\begin{aligned}\lambda^{ij}_{\max} &:= \lambda_{\max}(Q_{ij}) \\ &= \lambda_{\max}\left(W^{R\frac{1}{2}} B^T L^\dagger_{W^R} (e_i - e_j)(e_i - e_j)^T L^\dagger_{W^R} B W^{R\frac{1}{2}}\right) \\ &= \lambda_{\max}\left((e_i - e_j)^T L^\dagger_{W^R} B W^{R\frac{1}{2}} W^{R\frac{1}{2}} B^T L^\dagger_{W^R} (e_i - e_j)\right) \\ &= (e_i - e_j)^T L^\dagger_{W^R} L_{W^R} L^\dagger_{W^R} (e_i - e_j) \\ &= (e_i - e_j)^T L^\dagger_{W^R} (e_i - e_j),\end{aligned} \tag{11}$$

This quantity equals the resistance $R_{W^R,ij}$ [21]. We now have that Eq. (7) follows then from (10) and the bound $||W^{R\frac{1}{2}} D||_F^2 \leq c\,||\Delta||_F^2$ of Lemma A.2. Finally, Eq. (7) immediately implies the bound of Eq. (8). □

Our next step is to define some additional notation. The quantity K_{W_0} will be the weighted Laplacian corresponding to the bipartite graph with weights A_{ij}^2, formally defined as

$$K_W := \sum_{i \in I_x, j \in I_y} (e_i - e_j)(A_{ij})^2 (e_i - e_j)^T.$$

Observe that if, in this definition, we replaced A_{ij}^2 by the squares of the revealed entries, and also replaced the sum with only the sum over the revealed entries, then the resulting quantity would be exactly L_{W^R}. We may thus intuitively view the quantity K_W as the Laplacian corresponding to the hypothetical scenario that all entries are revealed without perturbations.

Inspired by the proof above, we will also define

$$Q = W^{R\frac{1}{2}} B^T L_{W^R}^\dagger K_W L_{W^R}^\dagger B W^{R\frac{1}{2}},$$

and, finally, we will use the shorthand

$$\lambda_{\max} := \lambda_{\max}(Q).$$

The symmetry and nonnegative definiteness of Q implies $\lambda_{\max}$ is real and nonnegative.

Observe that the existing bounds from Eq. (7) and Eq. (8) allow us to derive a bound on the error $\hat{A} - A$ via a few straightforward manipulations. This can then be turned into a bound on $||A - \hat{A}||_F$. However, this approach would be extremely conservative because the bounds of Eq. (7) and Eq. (8) are all worst-case, and a single Δ will not be worst for all (i, j). Our next proposition shows how to exploit this fact.

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

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

Before proceeding to the proof, we remark that $K_W L_{W^R}^\dagger$ is the product of two symmetric matrices, so its eigenvalues are real; consequently, writing $\lambda_{\max}\left(K_W L_{W^R}{}^\dagger\right)$ makes sense.

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

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

If follows then from Eq. (9) that

$$(\hat{A}_{ij} - A_{ij})^2 \leq \gamma^2 (A_{ij})^2 (W^{R\frac{1}{2}} D)^T Q_{ij} (W^{R\frac{1}{2}} D).$$

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

$$\left|\left|\hat{A} - A\right|\right|_F^2 \leq \gamma^2 \left|\left|W^{R\frac{1}{2}} D\right|\right|_F^2 \lambda_{\max}, \tag{12}$$

Finally using $L_{W^R} = B W^R B^T$,

$$\begin{aligned}
\lambda_{\max} &= \lambda_{\max}(W^{R\frac{1}{2}} B^T L_{W^R}^\dagger K_W L_{W^R}^\dagger B W^{R\frac{1}{2}}) \\
&= \lambda_{\max}\left(K_W L_{W^R}^\dagger B W^{R\frac{1}{2}} W^{R\frac{1}{2}} B^T L_{W^R}^\dagger\right) \\
&= \lambda_{\max}\left(K_W L_{W^R}^\dagger L_{W^R} L_{W^R}^\dagger\right) \\
&= \lambda_{\max}\left(K_W L_{W^R}^\dagger\right). \qquad (13)
\end{aligned}$$

The result now follows immediately from Eq. (12) and 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.

A.1 Corollary 1

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

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

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

where we recall that $A \in \mathbb{R}^{m\times n}$ and $\lambda_2(L)$ is the algebraic connectivity (i.e., second-smallest eigenvalue) of the unweighted *bipartite graph G.*

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

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

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 C.3 in Appendix C implies then

$$\lambda_2(L_{W^R}) \geq (\underline{\alpha}^R)^2 \lambda_2(L), \tag{14}$$

where we remind that L is the Laplacian of the unweighted bipartite graph G representing the mask Ω. A parallel argument shows that $\lambda_{\max}(K_W) \leq (\bar{\alpha}^0)^2\lambda_{\max}(K) = (m+n)(\bar{\alpha}^0)^2$, 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, A^R and the graph, but is not very conservative in terms of the graph properties. Indeed, a slightly more complicated argument shows that

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

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.

Having established proposition A.3, we now have that Corollary 1 follows almost immediately. Indeed, 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 $\lambda_2(L) \geq \frac{4}{\mathcal{D}(m+n)} \geq \frac{4}{(m+n)^2}$ [22].

A Proofs of the Lower Bounds

A.1 Small disturbances: proof of Theorem 2

Let us recall the basic setup. 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). Our approach is to construct, for a given value of $||\Delta||_F$ two matrices A^a, A^b whose entries in Ω are both within $||\Delta||_F$ of the revealed matrix A^R. Lower bounding $||A^a - A^b||$ will then produce a lower bound on the error achievable by any algorithm which only takes into account the set of revealed entries.

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 $\Delta^a := (A^a)_\Omega - A^R_\Omega = (A^a - A)$.

$$\begin{aligned} ||\Delta^a||^2_F &= ||(A^a - A)_\Omega||^2_F \qquad (15) \\ &= \sum_{(i,j)\in\Omega} \left(x_i y_j(1+\delta\zeta_i)(1-\delta\zeta_j) - x_i y_j\right)^2 \\ &= \sum_{(i,j)\in\Omega} x_i^2 y_j^2 \left(\delta(\zeta_i - \zeta_j) - \zeta_i\zeta_j\delta^2\right)^2 \\ &= \delta^2 \sum_{(i,j)\in\Omega} A^2_{ij}(\zeta_i - \zeta_j)^2 + o(\delta^2) \\ &= \delta^2 \zeta^T L_W \zeta + o(\delta^2), \qquad (16) \end{aligned}$$

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^2_{ij}. Parallel arguments show that

$$\left|\left|\Delta^b\right|\right|^2_F = \left|\left|(A^b - A)_\Omega\right|\right|^2_F = \delta^2\zeta^T L_W \zeta + o(\delta^2) \qquad (17)$$

and

$$\left|\left|A^b - A^a\right|\right|^2_F = 4\delta^2\zeta^T K_W \zeta + o(\delta^2), \qquad (18)$$

where K_W is the Laplacian of the weighted complete bipartite graph on $I_x \cup I_y$ with weight A^2_{ij}. 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 (16) and (17) that

$$\left|\left|\Delta^\ell\right|\right|^2_F = \delta^2\zeta^T u + o(\delta^2), \qquad (19)$$

for $\ell = a, b$, and from (18) that

$$\begin{aligned} \left|\left|A^b - A^a\right|\right|^2_F &= 4\delta^2\zeta^T K_W L_W^\dagger u + o(\delta^2) \\ &= 4\delta^2\lambda_{\max}(K_W L_W^\dagger)\zeta^T u + o(\delta^2). \qquad (20) \end{aligned}$$

Since u is an eigenvector corresponding to the largest eigenvalue of $K_W L_W^\dagger$, it is not a multiple of the all-ones vector (that would make it an eigenvector corresponding to the smallest eigenvalue of $K_W L_W^\dagger$, which has nonnegative eigenvalues since it is the product of two nonnegative definite matrices). Thus the definition $\zeta := L_W^\dagger u$ implies $u = L_W\zeta$, and $\zeta^T u = u^T L_W^\dagger u > 0$, since u is not proportional to the all-ones vector. Hence (18) and (20) imply that for $\ell = a, b$,

$$\left|\left|A^b - A^a\right|\right|^2_F = 4\lambda_{\max}(K_W L_W^\dagger)\left|\left|\Delta^\ell\right|\right|^2_F + o(\left|\left|\Delta^\ell\right|\right|^2_F). \qquad (21)$$

Suppose now that $A^R_{ij} = A_{ij}$ 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 $A^a, A^b \in \mathcal{A}$, $||\Delta^a||^2_F \le (1+\epsilon)\left|\left|\Delta^b\right|\right|^2_F$ and $\left|\left|\Delta^b\right|\right|^2_F \le (1+\epsilon)||\Delta^a||^2_F$,

so that both A^a, A^b would be possible values of A 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 (21) that for any estimate $\hat{A}$ there would hold

$$\left|\left|\hat{A} - A\right|\right|^2 \geq \lambda_{\max}(K_W L_W^\dagger)\left|\left|\Delta^\ell\right|\right|_F^2 + o(\left|\left|\Delta^\ell\right|\right|_F^2). \tag{22}$$

for at least one choice among $A = A^a$ or $A = A^b$. To conclude the result, we need to relate $\lambda_{\max}(K_W L_W^\dagger)$ to $\lambda_{\max}\left(K_W L_{W^R}^\dagger\right)$.

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

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

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

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

which implies the result of Theorem 2 together with (22).

A.2 Larger disturbances: proof of Theorem 3

We begin with the claim (a) about the exponential factor. For any given n, we take $A = ee^T$, 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

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

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 $A_{ij}^R = A_{ij}$ 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 $(\hat{A} - A)_{ij} = (1+\delta)^{i-j} - 1$. In particular, remembering $\delta = \frac{||\Delta||_F}{\sqrt{n-1}}$, we have

$$\begin{aligned}
\left|\left|\hat{A} - A\right|\right|_F^2 &\geq (\hat{A}_{1n} - A_{1n})^2 \\
&= \left((1+\delta)^{n-1} - 1\right)^2 \\
&= \left(\left(1 + \frac{||\Delta||_F}{\sqrt{n-1}}\right)^{n-1} - 1\right)^2 =: E_n.
\end{aligned}$$

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_{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 $A = ee^T$, 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 $||\Delta||^2_F \leq 1$, and $\max_{(i,j)\in\Omega} \frac{\Delta_{ij}}{A_{ij}}$ and $\max_{(i,j)} A_{ij}$ 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

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

B Proof of Theorem 4 and Corollary 2

High-level idea of proof: We have already seen in Section 5.1 that $(x, (y^{-1})^T$ (where the inverse is taken element-wise) is an eigenvector of the unperturbed matrix M. We therefore need to to argue that when looking at the eigenvector of the perturbed matrix M^R, we can recover a good approximation to $(x, (y^{-1})^T)$.

In general, this is tricky: it might involve expressions depending on the eigenvectors of the matrices M or M^R, which would be difficult to bound explicitly. However, two things make it possible in our case. The first is that the matrices M and M^R correspond to reversible Markov chains, which allows us to make use of a number of bounds appearing in the literature. The second is that the projection step is key: the assumption that the entries of A lie in $[\underline{\alpha}, \overline{\alpha}]$ allows us to project the resulting stationary distributions, which will therefore never be too small or too big; this allows us to go from an error bound on y^{-1} to an error-bound on y. As we have discussed in the main text of the paper, this seemingly minor difference is crucial: without a-priori bounds on entries of A, exponential growth is unavoidable due to our lower bounds.

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

Our first step is to upper bound the difference between the matrices M^R and M.

Lemma B.1.

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

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

Proof. For any scalars $x, y > 0$, we have the inequality

$$\begin{aligned}\left|\frac{x}{1+x} - \frac{y}{1+y}\right| &= \left|\tfrac{1}{1+x} - \tfrac{1}{1+y}\right| \\ &= \tfrac{|y-x|}{(1+x)(1+y)} \\ &\leq |y-x|.\end{aligned}$$

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

$$\left|M^R_{ij} - M_{ij}\right| = \left|M^R_{ji} - M_{ji}\right| \leq |\Delta_{ij}|\,. \tag{23}$$

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 $M^R - M$ sum to zero, we have then

$$\left|\left|M^R - M\right|\right|_\infty \leq 2 \max_{\ell \in I_x \cup I_y} \sum_{k \in I_x \cup I_y} \left|M^R_{\ell k} - M_{\ell k}\right|. \tag{24}$$

Consider first a $\ell = i \in I_x$. Then by the bipartite structure of M^R, M, the only off-diagonal nonzero $\left|M^R_{ik} - M_{ik}\right|$ are those for which $k \in I_y$. Hence, using (23), we have

$$\begin{aligned}
\sum_{k \in I_x \cup I_y} \left|M^R_{ik} - M_{ik}\right| &= \sum_{j \in I_y} \left|M^R_{ij} - M_{ij}\right| \\
&\leq \sum_{j \in I_y} |\Delta_{ij}| \\
&\leq ||\Delta||_\infty\,.
\end{aligned}$$

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

$$\begin{aligned}
\sum_{k \in I_x \cup I_y} \left|M^R_{jk} - M_{jk}\right| &= \sum_{i \in I_x} \left|M^R_{ji} - M_{ji}\right| \\
&\leq \sum_{i \in I_x} |\Delta_{ij}| \\
&\leq ||\Delta||_1\,.
\end{aligned}$$

The result follows then from Eq. (24). □

The next part of the proof exploits results on the perturbations of stationary distributions of (discrete-time) Markov chains. Our first step is to introduce a reference stationary distribution associated with the true matrix. Recall that we have assumed all entries of A to be in $[\rho^{-1}, \rho]$ so that $\mu = 1$; Lemma C.4, proved in a subsequent appendix, implies that $A = xy^T$ for some vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ with all x_i and y_j^{-1} in $[\rho^{-1}, \rho]$.

Furthermore, we have seen in Section 5.1 that $(x^T, (y^{-1})^T)$ is a left-eigenvector of M corresponding to its eigenvalue 0. We next define normalized version π^0 for which $\left|\left|\pi^0\right|\right|_1 = 1$. Due to the bounds on the entries of $(x^T, (y^T)^{-1})$ discussed in the previous paragraph, we see that the elements of π^0 all lie in $[\frac{\rho^{-2}}{m+n}, \frac{\rho^2}{m+n}]$. Moreover, the values of $\hat{\pi}$ (see Algorithm 1 for a definition of $\hat{\pi}$) lie in the same interval (see Algorithm 1: those values of $\hat{\pi}$ that are outside this interval are projected onto it).

Proposition B.1.

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

Proof. We will leverage results for perturbations of stationary distributions of discrete-time Markov-chains; to that end, we introduce the auxiliary matrices $P^R = I - \frac{1}{2d_{\max}} M^R$ and $P^0 = I - \frac{1}{2d_{\max}} M$, 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 are non-negative and bounded by 1, and that each row or column contains at most $d_{\max}$ of them. Moreover, using e to denote the all-ones vector, $M^R e = Me = 0$, which implies P^R, P^0 are row-stochastic matrices with positive diagonals.

The left-eigenvectors π^R and π^0 of M^R and M 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 [23] (Theorems 2, 3 and the discussion immediately after the statement of Theorem 3 in the supplementary materials) that

$$\begin{aligned}\left|\left|\pi^R - \pi^0\right|\right|_1 &\leq \frac{1}{2}\left|\left|P^R - P^0\right|\right|_\infty \left(\frac{\log R}{-\log \lambda_2(P^0)} + \frac{1}{1-\lambda_2(P^0)}\right) \\ &\leq \frac{1}{2}\left|\left|P^R - P^0\right|\right|_\infty \frac{\log R + 1}{1-\lambda_2(P^0)}. \end{aligned} \tag{25}$$

where the second line was obtained via the standard inequality $\log x \leq x - 1$; here

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

Our first step is to bound R. Indeed, since $P^0 = I - \frac{1}{2d_{\max}} M$ we have

$$1 - \lambda_2(P^0) = \frac{1}{2d_{\max}} \lambda_2(M),$$

where we remark on the difference in the (standard) convention: while $\lambda_2(M)$ is the second-smallest eigenvalue of M, $\lambda_2(P^0)$ refers to the second-largest eigenvalue of the latter.

Since $\pi^0_\ell \geq \frac{\rho^{-2}}{m+n}$ for every ℓ and $\frac{1-x}{4x}$ is decreasing, we have then

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

Reintroducing this and the expression $1 - \lambda_2(P^0) = \frac{1}{2d_{\max}} \lambda_2(M)$ into Eq. (25) leads to

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

and the result now follows from

$$\left|\left|P^R - P^0\right|\right|_\infty = \frac{1}{2d_{\max}} ||M^R - M||_\infty,$$

and from

$$||\hat{\pi} - \pi^0||_1 \leq ||\pi^R - \pi^0||_1,$$

which holds since each entry $\hat{\pi}_\ell$ is the projection of π^R_ℓ on an interval 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 $||\hat{A} - A||_F$ on the matrix.

Proposition B.2.

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

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

$$\begin{aligned}
\left|\hat{A}_{ij} - A_{ij}\right| &= \left|\frac{\hat{\pi}_i}{\hat{\pi}_j} - \frac{\pi_i^0}{\pi_j^0}\right| \\
&\leq \left|\frac{\hat{\pi}_i}{\hat{\pi}_j} - \frac{\hat{\pi}_i}{\pi_j^0}\right| + \left|\frac{\hat{\pi}_i}{\pi_j^0} - \frac{\pi_i^0}{\pi_j^0}\right| \\
&= \hat{\pi}_i \left|\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j^0}\right| + \frac{1}{\pi_j^0}\left|\hat{\pi}_i - \pi_i^0\right|,
\end{aligned}$$

so that

$$\begin{aligned}
\sum_{i,j}\left|\hat{A}_{ij} - A_{ij}\right| \leq \sum_i \hat{\pi}_i \left|\left|(\hat{\pi})^{-1} - (\pi^0)^{-1}\right|\right|_1 \\
+ \sum_j \frac{1}{\pi_j^0}\left|\left|\hat{\pi} - \pi^0\right|\right|_1 .
\end{aligned} \tag{26}$$

We first bound $\sum_i \hat{\pi}_i$. Observe[1] that

$$\begin{aligned}
\sum_i \hat{\pi}_i &= \sum_i \pi_i^R + \sum_i (\hat{\pi}_i - \pi_i^R) \\
&\leq 1 + \sum_i (\hat{\pi}_i - \pi_i^R).
\end{aligned}$$

Moreover, by construction $(\hat{\pi}_i - \pi_i^R)$ is positive only when

$$\pi_i^R < \rho^{-2}/(m+n),$$

in which case

$$\hat{\pi}_i = \rho^{-2}/(m+n).$$

Hence

$$\sum_i \hat{\pi}_i \leq 1 + \sum_i \frac{\rho^{-2}}{m+n} \leq 1 + \rho^{-2} \leq 2. \tag{27}$$

Secondly, since

$$\pi_j^0 \geq \frac{\rho^{-2}}{m+n},$$

we have

$$\sum_j \frac{1}{\pi_j^0} \leq \rho^2 m(m+n).$$

Plugging this into Eq. (26), we obtain

$$\sum_{i,j} |\hat{A}_{ij} - A_{ij}| \leq 2||\hat{\pi}^{-1} - (\pi^0)^{-1}||_1 + \rho^2 m(m+n)||\hat{\pi} - \pi^0||_1. \tag{28}$$

Moreover, since

$$\left|(\hat{\pi}_i)^{-1} - (\pi_i^0)^{-1}\right| = \frac{\left|\hat{\pi}_i - \pi_i^0\right|}{\hat{\pi}_i \pi_i^0}. \tag{29}$$

[1]It might be tempting to say that $\sum_i \hat{\pi}_i \leq 1$, but this may not be the cause because $\hat{\pi}_i$ is the projection of the statationary distribution, that that projection could increase the 1-norm.

we have that

$$||\hat{\pi}^{-1} - (\pi^0)^{-1}||_1 \leq ||\hat{\pi} - \pi^0||_1 \rho^4 (m+n)^2.$$

Plugging this into Eq. (28) leads to

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

and the result follows then from

$$\begin{aligned}\left|\left|\hat{A} - A\right|\right|_F &= \left|\left|\mathrm{vec}(\hat{\mathrm{A}} - \mathrm{A})\right|\right|_2 \\ &\leq \left|\left|\mathrm{vec}(\hat{\mathrm{A}} - \mathrm{A})\right|\right|_1 \\ &= \sum_{i,j} \left|\hat{A}_{ij} - A_{ij}\right|.\end{aligned}$$

□

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

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

Finally, to prove Corollary 2, observe first that all off-diagonal entries in M^R are at least $\frac{\rho^{-1}}{1+\rho^{-1}} \geq \rho^{-1}/2$ in absolute values. Moreover, as we have discussed in Section 5.1, $M_{k\ell}\pi_k = M_{\ell k}\pi_\ell$ for every $k, \ell \in I_x \cup I_y$; another way to say this is that $\mathrm{diag}(\pi^0)M$ is symmetric. Lemma C.3 implies then

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

Finally, recall that $(\pi^0)^T = K(x^T, (y^{-1})^T)$ for some constant K, and it follows from Lemma C.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 $\lambda_2(M) \geq \frac{\rho^{-3}}{2}\lambda_2(L)$. Corollary 2 follows from the combination of this bound with Theorem 4. □

C Technical Lemmas

Lemma C.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

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

□

Lemma C.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 [24]. As in Lemma C.1, we see that $\lambda_{\max}(A(\delta)B) = \lambda M(\delta)$, with

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

an analytical function of δ that is always positive semi-definite, so its eigenvalues are real. It follows then from Theorem 6.8 in [] 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 C.3. *Let L be a directed Laplacian: this means that L is a matrix whose rows sum to zero and whose off-diagonal elements are positive (but L may not be symmetric). Let A_{ij} denote these offdiagonal weights, let $a_{\min}$ be a lowr bound on the smallest positive A_{ij}. Finally, let $\bar{L}$ the corresponding Laplacian when all positive weights ar replced by one.*

We make the assumption that $\bar{L}$ is symmetric. If further L is symmetric, then

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

If instead 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 $x^T L x$ for symmetric Laplacian, we see that for any x, we have

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

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

$$L = D^{-1/2} D^{-1/2} DL$$

which implies that L and $D^{-1/2} DL D^{-1/2}$ are similar. IF DL is symmetric, the latter matrix is also symmetric, and we obtain that all the eigenvalues of L are real. Thus it makes sense to talk about

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

which is the second-smallest eigenvalue of L after the smallest eigenvalue of zero.

Observe that that $D^{1/2}e$ is an eigenvector of $D^{-1/2}(DL)D^{-1/2}$ with eigenvalue 0. Hence

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

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 C.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 $A = xy^T$.

For the vector x we have just constructed we have that

$$x_{\max} = \frac{\rho}{\hat{x}_{\max}}\hat{x}_{\max} = \rho.$$

This implies $x_{\min} \geq \rho^{-1}$ by the same argument as above.

Moreover, $y_{\max} \leq 1$, for otherwise we would have $\max_{i,j} A_{ij} = x_{\max}y_{\max} > \rho$. So, if $y_{\min} \geq \rho^{-1}$, then we are done. Otherwise, we have $\frac{\rho^{-1}}{y_{\min}} > 1$, and we can define

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

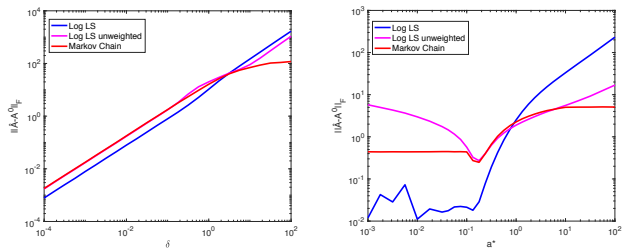
satisfying again $x'(y')^T = A$. By construction $y'_{\min} = \rho^{-1}$, so that $y'_{\max} \leq \rho$. Moreover, since $\frac{\rho^{-1}}{y_{\min}} > 1$, we have that

$$x'_{\max} \leq x_{\max} \leq \rho.$$

Finally,

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

so that x', y' satisfy the conditions we need. □



(a) (b)

Figure 3: Evolution of the average error $||\hat{A} - A||_F$ for the two proposed algorithms (Markov Chain, Log-LS) and for an unweighted version of the algorithm of Section 3 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.

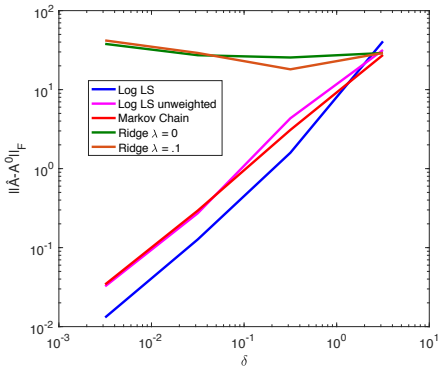
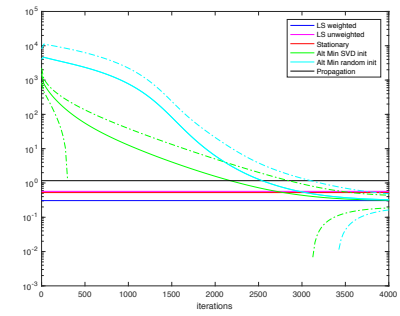


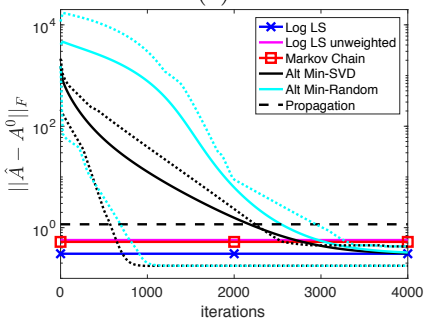
Figure 4: Evolution with δ of the average error $||\hat{A} - A||_F$ for our two algorithms, an unweighted version of the algorithm of Section 3, and our implementation of the ridge regression with $\lambda = 0$ (no regularization) and $\lambda = .1$, in a scenario where all revealed entries are perturbed 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.

D Additional Experiments

Ridge regression: Our implementation of the ridge-regression (Eq. 1) uses 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 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 4, for the same experimental conditions as in Figure 3(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.



(a)



(b)

Figure 5: Plots of Variances and 80% Error Margins with "Star-Graph" Sampling (3 columns and 3 rows)