Outliers Detection in Networks with Missing Links

Solenne Gaucher, Olga Klopp, Geneviève Robin

Outliers Detection in Networks with Missing Links

Solenne Gaucher [*1], Olga Klopp [†2,3], and Geneviève Robin [‡4]

[1]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay.
[2]ESSEC Business School
[3]CREST, ENSAE
[4]LaMME, CNRS, Université d'Évry Val d'Essonne

November 29, 2020

Abstract

Outliers arise in networks due to different reasons such as fraudulent behaviour of malicious users or default in measurement instruments and can significantly impair network analyses. In addition, real-life networks are likely to be incompletely observed, with missing links due to individual non-response or machine failures. Identifying outliers in the presence of missing links is therefore a crucial problem in network analysis. In this work, we introduce a new algorithm to detect outliers in a network that simultaneously predicts the missing links. The proposed method is statistically sound: we prove that, under fairly general assumptions, our algorithm exactly detects the outliers, and achieves the best known error for the prediction of missing links with polynomial computation cost. It is also computationally efficient: we prove sub-linear convergence of our algorithm. We provide a simulation study which demonstrates the good behaviour of the algorithm in terms of outliers detection and prediction of the missing links. We also illustrate the method with an application in epidemiology, and with the analysis of a political Twitter network. The method is freely available as an R package on the Comprehensive R Archive Network.

Keywords: outlier detection, robust network estimation, missing observations, link prediction

1 Introduction

Networks are powerful tools to analyse complex systems: agents are represented as nodes, and pairwise interactions between agents are recorded as edges between these nodes. Examples of fields of applications include biology, where networks may be used to describe protein-protein interactions; ecology, where they may represent food webs [17] or spatial distributions in crop diversity networks [58]; ethnology, where networks summarise relationships or trades between individuals or communities [48, 44]; sociology, where the recent development of online social networks offers unprecedented possibilities while fostering new challenges [59]. Those real-life networks are often modelled as realisations of random graphs or, equivalently, as noisy versions of more structured networks. In this setting, recovering the "noiseless" version of the graph, i.e. estimating the underlying probabilities of interactions between agents, is a key problem that has recently gained considerable attention (see, e.g., [35, 20, 22, 64]). Most of the proposed methods are based on models describing the connectivity of the majority of nodes. However, in many examples those models fail to describe networks containing a small number of outliers nodes with abnormal behaviour. Following Hawkins [28], we define an outlier as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".

[*]solenne.gaucher@ensae.fr
[†]kloppolga@math.cnrs.fr
[‡]genevieve.robin@inria.fr

Detecting nodes with anomalous behaviour is an important problem in applications. For example, in social networks, malicious nodes corresponding to fake accounts created to spread fake news, to distribute malware, or to spam other users may be hidden among the regular nodes [2]. These outliers often exhibit connection patterns that differ from that of normal nodes: the authors of [53] show, for example, that spam attackers are often connected with numerous nodes in a random fashion, thus forming characteristic hubs. By contrast, the connections between regular nodes are more sparse and more structured: they may, for example, exhibit community structures. Identifying those malicious nodes is crucial to protect users from the threat they represent. In the context of graphs obtained from survey data, anomalous behaviour may indicate that participants are providing false answers to distort the public opinion on a subject [4, 16]. In other cases, defaults of measurement instruments or fraudulent behaviours can lead to abnormal connectivity patterns. Finally, in contact networks, individuals with anomalous connection patterns may play an important role in the propagation of diseases, and their identification finds important applications in epidemiology [61]. These examples illustrate how identifying outlier nodes can provide us with hindsight on the network. Moreover, detecting these nodes allows us to control the bias induced by their anomalous behaviour in the network analysis. For example, it has been shown that the presence of hubs in graphs exhibiting community structure can hinder the estimation of these communities [10, 32].

In addition, many real-life networks are polluted by missing data [25, 26]. Indeed, complete exploration of all pairwise interactions between agents can be expensive, time consuming, and requires significant effort. In social sciences, graphs constructed from survey data are likely to be incomplete, due to non-response or drop-out of participants. Online social network data are often obtained through crawling of users profile; however the gigantic size of these networks may drive analysts to stop prematurely this crawling, and work with a sub-sample of the network [12]. Protein-protein interactions networks provide a blatant example of incompleteness, as the existence of each interaction must be tested experimentally, and most of these interactions have yet to be tested [66]. When dealing with a partially observed network, being able to predict the probability of existence of non-observed edges is of particular interest and finds numerous applications, for example in biology [7], recommender systems [42] and ecology [19].

In this paper, we propose a new algorithm that detects the outliers in networks. In addition, this method robustly estimates the probabilities of connection of the nodes in the network, which allows to predict the missing links. The present paper is mostly related to two lines of work in network analysis: anomaly detection in networks and estimation in networks with missing values. Anomaly detection in networks has indeed been studied under several sets of assumptions on the behaviour of outlier nodes; we refer the interested reader to [2] for a review of these technics. For instance, many algorithms based on trust propagation rely on the assumption that outlier nodes are not well connected with normal nodes [67, 46]. Other algorithms, based on community structure, assume that outliers [60, 43] fail to be well connected to communities of normal nodes. However, it has been shown in [65] that these assumptions do not hold in many situations. In addition, most of these technics focus on outliers detection, and do not study the estimation of underlying structure. Meanwhile, robust estimation of the graph structure in the presence of outlier nodes has been less studied. In [10], the authors aim to recover community structures when the majority of the nodes follow an assortative stochastic block model in the presence of arbitrary outlier nodes. However, their algorithm does not allow to detect these outlier nodes. Note that our problem is different, as we would like to estimate connection probabilities between nodes rather than recover community structures, and our assumptions on the random graph are more general.

On the other hand, estimation in networks with missing observations, and its application to link prediction has known a quite recent development. In [20], the authors study the least squares estimator for the stochastic block model assuming observations are missing uniformly at random, and show that the procedure is minimax optimal. In [22], the authors show that the maximum likelihood estimator is minimax optimal in the same setting, while being adaptive to more general sampling schemes. These two estimators are too costly to compute to be used in practice (computationally efficient approximations exist for the maximum likelihood). In [68], the authors consider the setting where non-existing edges can be erroneously recorded as observed (or existing edges recorded as not observed), both errors occurring at a fixed rate. More recently, [55] and [62] proposed algorithms to estimate the edge probabilities under different missing observations schemes, and [40] proposed a method for consistent community detection also under several missing values scenarios. Both papers present convincing numerical experiments, but lack theoretical guarantees.

Finally, our work is also closely related to recent developments in the field of robust matrix completion. Indeed, in our general model presented in Section 2, we assume that the matrix of connection probabilities can be decomposed as the sum of a low rank component (connectivity pattern of inliers), and that of a column-wise sparse component (non-zero columns corresponding to outliers). Our problem is to estimate the low-rank matrix in order to reconstruct the connectivity of inliers, and to *reconstruct the support of the column-wise sparse component*, in order to detect outliers. The problem of estimating the low-rank matrix is related to that of robust matrix completion, in which one aims at estimating a low-rank matrix from incomplete and corrupted observations of its entries; see, for example, [11, 13, 30, 63, 3, 14, 41, 34]. More recently, the problem of robust matrix completion with binary observations has been studied in [50, 52]. However, to the best of our knowledge, existing work on sparse plus low-rank matrix decomposition in the noisy case do not provide guarantees concerning support recovery of the sparse component. In this paper, we provide such results and prove that our algorithm exactly recovers the support of the sparse matrix. Another shortcoming of existing results on binary robust matrix completion (e.g. [50, 52]) is that applying them to the estimation of connection probabilities in networks yields sub-optimal error rates. Indeed, in our framework, the signal to noise ratio is critically low, as the variances of the variables are of the same order as their expectations. The main difficulty arising in our case, and that we tackle in the present paper, is therefore to obtain the optimal dependence on the sparsity of the network.

In the present work, we present a new algorithm to detect the outliers and to estimate the connection probabilities of the remaining nodes, which is robust to missing observations. For this algorithm, we provide both statistical and computational guarantees. In particular, in Theorem 3, we prove that under fairly general assumptions our algorithm achieves exact detection of the outliers. In Theorem 4, we also prove an upper bound on the estimation error of connection probabilities between inliers. Importantly, the estimation error of our method matches the best known error for tractable algorithms [64]. We also analyse the algorithm convergence complexity in Theorem 1, and show sub-linear convergence. In Section 5, we provide a simulation study with comparisons to state-of-the-art technics, indicating that the proposed method has good empirical properties in terms of outliers detection and link prediction. Finally, we illustrate the performance of our method with two applications in epidemiology and social network analysis.

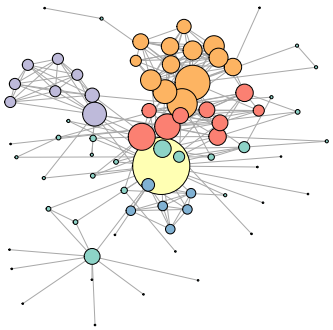
1.1 Example: "Les Misérables" characters network

Before introducing our general model, let us start with an example."Les Misérables" characters network encodes interactions between characters of Victor Hugo's novel; the network was created by Donald Knuth, as part of the Stanford Graph Base [36]. It contains 77 nodes corresponding to characters of the novel, and 254 edges connecting two characters whenever they appear in the same chapter. The book itself spans around two decades of nineteenth century France and numerous characters. It is structured in five volumes, each one focused on a specific period and featuring handful of characters. One expects to observe communities in this network, corresponding roughly to the plots narrated in each volume: such structures are well captured by the classical Stochastic Block Model (SBM). In the SBM (see, e.g., [29]), nodes are classified into k communities (for example corresponding to volumes of the book). Denote by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ the graph, where $\mathcal{V}$ is the set of nodes, and $\mathcal{E}$ the set of edges. For any $i \in \mathcal{V}$, denote by $c(i)$ its community assignment. Then, the probability that an edge connects two nodes only depends on their community assignments:

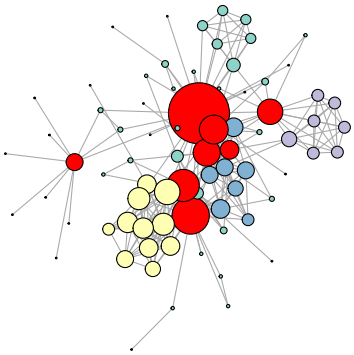
$$\mathbb{P}((i,j) \in \mathcal{E}) = \boldsymbol{Q}_{c(i)c(j)}. \tag{1}$$

In (1), $\boldsymbol{Q}$ denotes a $k \times k$ symmetric matrix of connection probabilities between communities. Usually, in the Stochastic Block Model, the community assignment is unknown and learned from data.

However, some of the characters behave differently, as their stories follow the entire novel. For instance, the main character, Jean Valjean, acts as a *hub* with 36 connections, well above the second most connected character Gavroche, with a degree of 22. Other characters, for instance, Cosette, do not necessarily have a large degree but are connected to characters across all the volumes, and thus also stand out from the communities structure. Nodes such as Cosette correspond to outliers with *mixed membership* profile. In Figure 1a, we display the communities assignment resulting from the classical SBM. Note that the node corresponding to Jean Valjean (large yellow node), is alone in its community. In addition, one of the clusters (in red) contains most of the main characters of the novel (Les Thénardier, Éponine, Javert).



(a) SBM model with 6 communities (the number of communities is chosen to minimise the Integrated Completed Likelihood criterion).



(b) Proposed Stochastic Block Model with outliers. The detected outliers are coloured in red, and classification is performed on the rest of the nodes.

Figure 1: Les Misérables characters network. The nodes are represented with size proportional to their degree, and coloured according to their community assignment. On the left in Figure 1a, classification is performed according to the classical SBM model. On the right in Figure 1b, the detected outliers are indicated in red, and classification is performed on the rest of the nodes (inliers).

To model simultaneously the community structure and the outlier profiles, we propose to decompose $\mathcal{V}$ into two sets of nodes: the inliers $\mathcal{I}$ following the classical Stochastic Block Model structure and the outliers $\mathcal{O}$ for which we do not make any assumption on their connection pattern. As a result, the probability of connection between inliers is given, for any $(i,j) \in \mathcal{I}^2$, by

$$\mathbb{P}((i,j) \in \mathcal{E}) = \boldsymbol{L}^*_{ij},$$

where $\boldsymbol{L}^*$ is a symmetric matrix with entries in $[0,1]$ corresponding to a classical SBM. On the other hand, for any outlier $i \in \mathcal{O}$ and for any node $j \in \mathcal{V}$ we set

$$\mathbb{P}((i,j) \in \mathcal{E}) = \left(\boldsymbol{S}^* + \boldsymbol{S}^{*\top}\right)_{ij},$$

with $\boldsymbol{S}^*$ an arbitrary matrix in $[0,1]^{n \times n}$. Our only assumption regarding the outliers is that their number is small compared to the size of the network, i.e., the matrix $\boldsymbol{S}^*$ is column-wise sparse. Note that the inlier and outlier sets are unknown a priori, and learned from data. In Figure 1b, we display the communities assignment resulting from our model. The outlier nodes – which are selected automatically by our procedure – are indicated in red, and coincide with central characters of the novel. They correspond either to hubs (Jean Valjean, Myriel) or to nodes with mixed memberships (Cosette, Javert, Marius).

1.2 Organisation of the paper

The rest of the paper is organised as follows. First, in Section 1.3, we summarise notation used throughout this paper and, in Section 2, we introduce our model. Then, in Section 3, we present a computationally efficient algorithm for detecting outliers and estimating the connection probabilities between inliers. We also provide theoretical guarantees on the speed of convergence of this algorithm. In Section 4, we provide bounds on the error of the outliers detection and on the error of the estimation of the connection probabilities between inliers. In Section 5, we present numerical experiments which demonstrate the good empirical behaviour of our method, both in terms of outliers detection and in terms of prediction of the missing links. The method is implemented in the R [49] package `GSBM` available on the Comprehensive R Archive Network. The proofs are relegated to the Appendix A.

1.3 Notations

The notation used in the paper is gathered in the following paragraph :

- We use bold notations for matrices and vectors: for any matrix $\boldsymbol{M}$, we denote by $\boldsymbol{M}_{ij}$ its entry on row i and column j. The vector corresponding to its i-th row is denoted by $\boldsymbol{M}_{i,\cdot}$, and the vector corresponding to its j-th column is denoted by $\boldsymbol{M}_{\cdot,j}$. The notation $\mathbf{0}$ denotes either a matrix or a vector with entries all equal to 0.

- We write $\odot$ to denote the entry-wise product for matrices or vectors. For any vector $\mathbf{v} \in \mathbb{R}^n$, we denote by $\|\mathbf{v}\|_2$ its Euclidean norm. For any two matrices $\boldsymbol{M}, \boldsymbol{N} \in \mathbb{R}^{n\times n}$, $\langle \boldsymbol{M} | \boldsymbol{N} \rangle \triangleq \sum_{ij} \boldsymbol{M}_{ij}\boldsymbol{N}_{ij}$ is the Frobenius scalar product between $\boldsymbol{M}$ and $\boldsymbol{N}$. For any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, $\|\boldsymbol{M}\|_F$ is its Frobenius norm, $\|\boldsymbol{M}\|_*$ is its nuclear norm (the sum of its singular values), $\|\boldsymbol{M}\|_{\text{op}}$ is its operator norm (its largest singular value), and $\|\boldsymbol{M}\|_\infty \triangleq \max_{ij} |\boldsymbol{M}_{ij}|$ is the largest absolute value of its entries. Its column-wise 2,1-norm is denoted by $\|\boldsymbol{M}\|_{2,1} \triangleq \sum_j \sqrt{\sum_i \boldsymbol{M}_{ij}^2}$, and the column-wise 2,$\infty$-norm is denoted by $\|\boldsymbol{M}\|_{2,\infty} \triangleq \max_j \sqrt{\sum_i \boldsymbol{M}_{ij}^2}$. The weighed L_2-norm with respect to the sampling probability $\boldsymbol{\Pi}$ is written $\|\boldsymbol{M}\|_{L_2(\boldsymbol{\Pi})}$. Finally, for any matrix $\boldsymbol{M}$ and any vector $\mathbf{v}$, we denote respectively by $(\boldsymbol{M})_+$ and $(\mathbf{v})_+$ the matrix and vector obtained by considering the positive part of their entries.

- For a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, we denote by $\mathcal{P}_{\boldsymbol{M}}$ the projection defined as follows: for any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathcal{P}^\perp_{\boldsymbol{M}}(\boldsymbol{A}) = \boldsymbol{A} - \mathcal{P}^\perp_{\boldsymbol{M}}(\boldsymbol{A})$, where $\mathcal{P}^\perp_{\boldsymbol{M}}(\boldsymbol{A}) = P^\perp_{U(\boldsymbol{M})}\boldsymbol{A}P^\perp_{V(\boldsymbol{M})}$, and $P^\perp_{U(\boldsymbol{M})}$ and $P^\perp_{V(\boldsymbol{M})}$ denote respectively the projection on the spaces orthogonal to the spaces spanned by the right and left singular vectors of $\boldsymbol{M}$.

- We denote by $[n]$ the set of integers from 1 to n, by $\mathcal{I}$ the set of inlier nodes, and by $\mathcal{O}$ the set of outlier nodes. For a set of indices $\mathcal{S}$ and a matrix $\boldsymbol{M} \in \mathcal{R}^{n\times n}$, we write $\boldsymbol{M}_{|\mathcal{S}} \triangleq \mathbb{1}_{\mathcal{S}} \odot \boldsymbol{M}$ where $\mathbb{1}_{\mathcal{S}}$ is the indicator matrix of the set $\mathcal{S}$. For any set $\mathcal{S}$, we denote by $|\mathcal{S}|$ its cardinality.

2 General model

We consider an undirected, unweighted graph with n nodes indexed from 1 to n. To encode the set of edges, we use the *adjacency matrix* of the graph, which we denote by $\boldsymbol{A}$. This matrix is defined as follows: set $\boldsymbol{A}_{ij} = 1$ if there exists an edge linking node i and node j, and $\boldsymbol{A}_{ij} = 0$ otherwise. Note that since the graph is undirected we have $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji}$. We assume there are no loops in the graph: no edge can connect a node to itself, and thus $\boldsymbol{A}_{ii} = 0$. The nodes can be divided into inliers and outliers. Inliers correspond to the majority of the nodes, and their connection probabilities are given by a low-rank model. Outliers correspond to a small number of nodes with anomalous connections, and connect arbitrarily to inlier and outlier nodes.

Probability of connection between inliers For any pair of inliers $(i,j) \in \mathcal{I}^2$, $i < j$ we assume that $\boldsymbol{A}_{ij} \overset{ind.}{\sim} \text{Bernoulli}(\boldsymbol{L}^*_{ij})$, where $\boldsymbol{L}^*$ is a $n \times n$ symmetric matrix with entries in $[0,1]$. For inliers, we consider a more general model than the classical Stochastic Block Model assuming that $\boldsymbol{L}^*$ is low-rank. This assumption is enough to model some interesting properties of the SBM, such as positive and negative homophily, and stochastic equivalence. Indeed, when $\text{rank}(\boldsymbol{L}^*) = k$, there exist a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times k}$ and a diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{k\times k}$ such that $\boldsymbol{L}^* = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. The model can then be interpreted as follows: each row $\boldsymbol{U}_{i,\cdot}$ corresponds to a vector of k latent attributes describing the node i. If $\boldsymbol{\Lambda}_{aa} > 0$, two nodes sharing attributes of the same sign along the a-th coordinate will have a tendency to be more connected (everything else being equal), modelling positive homophily along this coordinate. If $\boldsymbol{\Lambda}_{aa} < 0$, they will tend to be less connected, modelling negative homophily. Note that two nodes with similar characteristics in the latent space will have similar stochastic behaviour (i.e. their probabilities of connection to other nodes will be given by similar vectors of probabilities). On the other hand, assuming that $\boldsymbol{L}^*$ is low-rank closely relates to the *latent eigenmodel*, described, for example, in [15]. In this model, the probability of connection of nodes i and j is given by

$f(\boldsymbol{L}^*_{ij})$, where $\boldsymbol{L}^*$ is of rank k and f is a link function. Note that our algorithm can be extended to the latent eigenmodel by replacing $\boldsymbol{L}$ by $f(\boldsymbol{L})$ in the objective function (4).

Finally, most graphs encountered by practitioners are *sparse*, with a small average degree compared to the number of nodes. To account for the sparsity, we assume that the entries of $\boldsymbol{L}^*$ are bounded by ρ_n where ρ_n is a sequence of sparsity inducing parameters such that $\rho_n \to 0$. In particular, we have that the average degree of the graph grows as $\rho_n n$. In the rest of the paper we assume that $\rho_n \leq \frac{1}{2}$. This assumption is only intended to clarify the exposition of our results, and can be easily removed.

Probability of connection of outlier nodes In our model we have no assumptions on the connectivity of outliers. In particular, we do not assume a block constant or a low rank structure. We set $\boldsymbol{L}^*_{ij} = 0$ for any pair of nodes (i, j) such that either $i \in \mathcal{O}$ or $j \in \mathcal{O}$, and we use matrix $\boldsymbol{S}^*$ to describe the outliers. For any inlier $j \in \mathcal{I}$, the j-th column of $\boldsymbol{S}^*$ is null. Therefore, the matrix $\boldsymbol{S}^*$ has at most $s = |\mathcal{O}|$ non-zero columns, where the number of outliers s is small compared to the number of nodes n. For any outlier $j \in \mathcal{O}$, the j-th column of $\boldsymbol{S}^*$ describes the connectivity of j: for any $j \in \mathcal{O}$ and $i \in \mathcal{I}$, $\boldsymbol{A}_{ij} \sim \text{Bernoulli}(\boldsymbol{S}^*_{ij})$ and for any $(i, j) \in \mathcal{O} \times \mathcal{O}$, $\boldsymbol{A}_{ij} \sim$Bernoulli$(\boldsymbol{S}^*_{ij} + \boldsymbol{S}^*_{ji})$. We set $\boldsymbol{S}^*_{ii} = 0$ for any $i \in [n]$. With these notations, we have that

$$\mathbb{E}[\boldsymbol{A}] = \boldsymbol{L}^* - \text{diag}(\boldsymbol{L}^*) + \boldsymbol{S}^* + (\boldsymbol{S}^*)^\top. \tag{2}$$

In this model, the outliers may account for different types of behaviour of the nodes, such as hubs or mixed membership profiles. In practice, while most nodes may be assigned to a community and share a similar stochastic behaviour with members of their community, a fraction of the nodes may belong to two or more communities. Our model allows for such a behaviour by considering the nodes with mixed membership as outliers. In these cases, being able to detect nodes with singular behaviour provides valuable information on the network. Note that this setting includes as particular case the Generalised Stochastic Block Model, introduced in [10]. In this model, the n nodes consist of $n - s$ inliers obeying the Stochastic Block Model (SBM), and s outliers, which are connected with other nodes in an arbitrary way.

Missing data pattern We say that we sample the pair (i, j) if we observe the presence or absence of the corresponding edge. We denote by $\boldsymbol{\Omega}$ the sampling matrix such that $\boldsymbol{\Omega}_{ij} = 1$ if the pair (i, j) is sampled, $\boldsymbol{\Omega}_{ij} = 0$ otherwise. The graph is unoriented and the sampling matrix is therefore symmetric; moreover we set $\text{diag}(\boldsymbol{\Omega}) = \boldsymbol{0}$ since an observation of a entry on the diagonal of $\boldsymbol{A}$ does not carry any information. We assume that the entries $\{\boldsymbol{\Omega}_{ij}\}_{i<j}$ are independent random variables and that $\boldsymbol{\Omega}$ and $\boldsymbol{A}$ are independent. We denote by $\boldsymbol{\Pi} \in \mathbb{R}^{n\times n}$ the expectation of the random matrix $\boldsymbol{\Omega}$. Then, for any pair (i, j), $\boldsymbol{\Omega}_{ij} \sim \text{Bernoulli}(\boldsymbol{\Pi}_{ij})$. For any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, we define

$$\|\boldsymbol{M}\|^2_{L_2(\boldsymbol{\Pi})} \triangleq \mathbb{E}\left[\|\Omega \odot \boldsymbol{M}\|^2_F\right].$$

This fairly general sampling scheme covers some of the settings encountered by practitioners. In particular, it covers the case of random dyad sampling (described, e.g., in [56]), where the probability of sampling any pair depends on the matrices $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ (and, if we consider the Stochastic Block Model, on the communities of the adjacent nodes).

Identifiability of the model The matrices $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ appearing in the decomposition (2) may not be unique. Since we estimate $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ from a noisy, incomplete observation of their sum, we cannot achieve exact reconstruction of these matrices, and do not require strong identification conditions. We restrict our attention to pairs of matrices $\left(\boldsymbol{L}^{(1)}, \boldsymbol{S}^{(1)}\right)$ such that

$$\left(\boldsymbol{L}^{(1)}, \boldsymbol{S}^{(1)}\right) \in \arg\min\left\{\text{rank}(\boldsymbol{L}) + \|\boldsymbol{S}\|_{2,0} : \mathbb{E}[\boldsymbol{A}] = \boldsymbol{L} - \text{diag}(\boldsymbol{L}) + \boldsymbol{S} + (\boldsymbol{S})^\top, (\boldsymbol{L}, \boldsymbol{S}) \in \mathcal{M}\right\},$$

where $\|\boldsymbol{S}\|_{2,0}$ is the number of non-zero columns of the matrix $\boldsymbol{S}$, and $\mathcal{M}$ is the set of admissible pairs of matrices :

$$\mathcal{M} = \left\{(\boldsymbol{L}, \boldsymbol{S}) : \boldsymbol{L} \in [0, \rho_n]^{n\times n}_{sym}, \quad \boldsymbol{S} \in [0, 1]^{n\times n}, \ \forall j \in [n], \boldsymbol{S}_{\cdot,j} \neq 0 \ \Leftrightarrow \ \boldsymbol{L}_{\cdot,j} = 0\right\}.$$

Among matrices verifying equation (3), we choose to consider matrices $\boldsymbol{L}$ with minimal rank, as they reflect our belief that inlier nodes should have a low-rank connectivity pattern. Thus, for $c = \text{rank}(\boldsymbol{L}^{(1)}) + \|\boldsymbol{S}^{(1)}\|_{2,0}$, we define

$$(\boldsymbol{L}^*, \boldsymbol{S}^*) \in \arg\min \Big\{ \text{rank}(\boldsymbol{L}) : \mathbb{E}\left[\boldsymbol{A}\right] = \boldsymbol{L} - \text{diag}(\boldsymbol{L}) + \boldsymbol{S} + (\boldsymbol{S})^\top, (\boldsymbol{L}, \boldsymbol{S}) \in \mathcal{M},\ \text{rank}(\boldsymbol{L}) + \|\boldsymbol{S}\|_{2,0} = c \Big\}. \quad (3)$$

Again, the solution of equation (3) may not be unique. We show in Section 4 that under assumption 4, strong identifiability is guaranteed, and we can detect exactly all outliers with large probability.

When assumption 4 does not hold, we can still show that all matrices $\boldsymbol{L}^*$ solution to (3) are close to each other in Frobenius norm. By definition, all solutions $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ of equation (3) are such that $\text{rank}(\boldsymbol{L}^*) = k$ and $\|\boldsymbol{S}^*\|_{2,0} = s$. Moreover, for all solution $(\tilde{\boldsymbol{L}}, \tilde{\boldsymbol{S}}) \neq (\boldsymbol{L}^*, \boldsymbol{S}^*)$, we can show that $\boldsymbol{L}^*$ and $\tilde{\boldsymbol{L}}$ are close in Frobenius norm. Indeed, let $\mathcal{I} = \{j : \boldsymbol{L}^*_{\cdot,j} \neq \boldsymbol{0}\}$ (respectively $\tilde{\mathcal{I}} = \{j : \tilde{\boldsymbol{L}}_{\cdot,j} \neq \boldsymbol{0}\}$) be the support of the columns of $\boldsymbol{L}^*$ (respectively of $\tilde{\boldsymbol{L}}$), and $\mathcal{O} = \{j : \boldsymbol{S}^*_{\cdot,j} \neq \boldsymbol{0}\}$ (respectively $\tilde{\mathcal{O}} = \{j : \tilde{\boldsymbol{S}}_{\cdot,j} \neq \boldsymbol{0}\}$) be the support of the columns of $\boldsymbol{S}^*$ (respectively of $\tilde{\boldsymbol{S}}$). Then,

$$\boldsymbol{L}^* = \mathbb{E}[\boldsymbol{A}]_{|\mathcal{I}\times\mathcal{I}} \text{ and } \tilde{\boldsymbol{L}} = \mathbb{E}[\boldsymbol{A}]_{|\tilde{\mathcal{I}}\times\tilde{\mathcal{I}}}.$$

Thus, $\boldsymbol{L}^* - \tilde{\boldsymbol{L}}$ is has support in the symmetrical difference between $\mathcal{I}\times\mathcal{I}$ and $\tilde{\mathcal{I}}\times\tilde{\mathcal{I}}$. Thus, $\boldsymbol{L}^* - \tilde{\boldsymbol{L}}$ has at most

$$2|(\mathcal{I}\cap\tilde{\mathcal{O}})\times(\mathcal{I}\cap\tilde{\mathcal{I}})| + |(\mathcal{I}\cap\tilde{\mathcal{O}})\times(\mathcal{I}\cap\tilde{\mathcal{O}})| + 2|(\tilde{\mathcal{I}}\cap\mathcal{O})\times(\tilde{\mathcal{I}}\cap\mathcal{I})| + |(\tilde{\mathcal{I}}\cap\mathcal{O})\times(\tilde{\mathcal{I}}\cap\mathcal{O})|$$

non zero entries, and each entry is bounded by ρ_n (because it belongs either to $\mathcal{I}$ or to $\tilde{\mathcal{I}}$). Since $|\tilde{\mathcal{O}}| = |\mathcal{O}| = s$ and $|\tilde{\mathcal{I}}| = |\mathcal{I}| \leq n$, the solution $\tilde{\boldsymbol{L}}$ is therefore in a Frobenius ball of radius $\sqrt{(4ns + 2s^2)}\rho_n \leq \sqrt{6ns}\rho_n$, centered at $\boldsymbol{L}^*$. Now, Corollary 1 ensures that our estimator $\widehat{\boldsymbol{L}}$ is in a ball centered at $\boldsymbol{L}^*$ of radius

$$R = C\mu_n^{-1/2}\left(\frac{\nu_n}{\mu_n}\rho_n kn + (\nu_n\rho_n \vee \tilde{\nu}_n\gamma_n)\rho_n sn\right)^{1/2},$$

where ν_n, $\tilde{\nu}_n$ and μ_n are upper and lower bounds on the sampling probabilities defined in Section 4, γ_n is an upper bound on the entries of $\mathbb{E}[A]$, and C is an absolute constant. Since $R \geq \sqrt{6ns}\rho_n$, the distance between our estimator $\widehat{\boldsymbol{L}}$ and any matrix $\tilde{\boldsymbol{L}}$ solution of (3) is bounded by $2R$.

3 Estimation procedure

In order to estimate the matrices $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$, we consider the following objective function:

$$\mathcal{F}(\mathbf{S}, \boldsymbol{L}) \triangleq \frac{1}{2}\|\Omega \odot (\boldsymbol{A} - \boldsymbol{L} - \mathbf{S} - (\boldsymbol{S})^\top)\|_F^2 + \lambda_1\|\boldsymbol{L}\|_* + \lambda_2\|\boldsymbol{S}\|_{2,1}, \quad (4)$$

defined by a least squares data-fitting term penalised by a hybrid regularisation term. On the one hand, the nuclear norm penalty $\|\boldsymbol{L}\|_*$ is a convex relaxation of the rank constraint, meant to induce low-rank solutions for $\boldsymbol{L}$. On the other hand, the term $\|\boldsymbol{S}\|_{2,1}$ is a relaxation of the constraint on the number of non-zero columns in $\boldsymbol{S}$, meant to induce column-wise sparse solutions for $\boldsymbol{S}$. Our estimators are defined as

$$\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{L}}\right) \in \underset{\boldsymbol{S}\in[0,1]^{n\times n}, \boldsymbol{L}\in[0,\rho_n]^{n\times n}_{sym}}{\arg\min} \mathcal{F}(\boldsymbol{S}, \boldsymbol{L}). \quad (5)$$

When information on the presence or absence of some edges is missing, the objective function may not have a unique minimiser. We propose to approximate our target parameters $(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{L}})$ by minimising the objective (4) with an additional ridge penalisation term, $\frac{\epsilon}{2}(\|\boldsymbol{L}\|_F^2 + \|\boldsymbol{S}\|_F^2)$, which ensures strong convexity of the objective function. This additional penalty is not necessary to obtain convergence in terms of the objective value, and setting $\epsilon = 0$ does not impact the convergence of the algorithm. However, it is required to obtain convergence of the parameters themselves: this additional penalty allows also to ensure approximate matching of the estimation and approximation errors, as detailed in our theoretical results. Note that, by choosing ϵ sufficiently small, $\mathcal{F}_\epsilon$ can be arbitrarily close to $\mathcal{F}$, but the choice of ϵ will impact the speed of convergence of our algorithm.

Furthermore, we assume for simplicity that the box constraints on $\boldsymbol{S}$ and $\boldsymbol{L}$ are always inactive. We make a final simplification by dropping the symmetry constraint on $\boldsymbol{L}$. Indeed, we will see later on that the low-rank matrix $\boldsymbol{L}$ remains symmetric throughout the algorithm, provided that it is initialised by a symmetric matrix. Thus, in the end, we (approximately) solve the following optimisation problem:

$$\text{minimize} \quad \mathcal{F}_\epsilon(\boldsymbol{S}, \boldsymbol{L}) \triangleq \mathcal{F}(\boldsymbol{S}, \boldsymbol{L}) + \tfrac{\epsilon}{2}(\|\boldsymbol{L}\|_F^2 + \|\boldsymbol{S}\|_F^2). \tag{6}$$

Let us now describe the optimisation procedure. First, we consider the augmented objective function:

$$\Phi_\epsilon(\boldsymbol{S}, \boldsymbol{L}, R) \triangleq \frac{1}{2}\|\Omega \odot (\boldsymbol{A} - \boldsymbol{L} - \boldsymbol{S} - (\boldsymbol{S})^\top)\|_F^2 + \lambda_1 R + \lambda_2 \|\boldsymbol{S}\|_{2,1} + \frac{\epsilon}{2}(\|\boldsymbol{L}\|_F^2 + \|\boldsymbol{S}\|_F^2),$$

with $R \in \mathbb{R}_+$. Note that, if an optimal solution to (6) $(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon)$ satisfies $\|\hat{\boldsymbol{L}}_\epsilon\|_* \leq \bar{R}$ for some $\bar{R} \geq 0$, then any optimal solution to the augmented problem

$$\begin{array}{ll} \text{minimise} & \Phi_\epsilon(\boldsymbol{S}, \boldsymbol{L}, R) \\ \text{such that} & \|\boldsymbol{L}\|_* \leq R \leq \bar{R} \end{array} \tag{7}$$

will also be optimal to (6) (we will show in appendix A.2 how the upper bound $\bar{R}$ can be chosen and tightened adaptively inside the algorithm). Thus, solving (7) we directly obtain the solution to our initial problem (6). Finally, our estimators are defined as the minimisers of the following augmented objective function:

$$\begin{array}{ll} (\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \tilde{R}) & \in \operatorname{argmin} \Phi_\epsilon(\boldsymbol{S}, \boldsymbol{L}, R) \\ \text{such that} & \|L\|_* \leq R \leq \bar{R}. \end{array}$$

A natural option to solve problem (7) is the coordinate descent algorithm, where the parameters $(\boldsymbol{S}, \boldsymbol{L}, R)$ are updated alternatively along descent directions. To update $\boldsymbol{S}$, we apply the proximal gradient method. We use the conjugate gradient method (or Frank-Wolfe method [31], which relies on linear approximations of the objective function) to update $(\boldsymbol{L}, R)$. Similar Mixed Coordinate Gradient Descent (MCGD) algorithms were considered in [45, 51, 21] to estimate sparse plus low-rank decomposition with hybrid penalty terms combining an ℓ_1 and a nuclear norm penalties. Here, we extend the procedure to handle the $\ell_{2,1}$ penalty as well. The details of the algorithm are described in Appendix A.2. The entire procedure is sketched in Algorithm 1, where we also define our final estimators $\left(\boldsymbol{L}^{(T)}, \boldsymbol{S}^{(T)}\right)$.

Algorithm 1 Mixed coordinate gradient descent (MCGD)

Initialisation: $(\boldsymbol{L}^{(0)}, \boldsymbol{S}^{(0)}, R^{(0)}, t) \leftarrow (\boldsymbol{0}, \boldsymbol{0}, 0, 0)$
for $t = 1, \ldots, T$ **do**
  $t \leftarrow t + 1$
  Compute the proximal update (8) to obtain $\mathbf{S}^{(t)}$.
  Compute the upper bound $\bar{R}^{(t)} = \lambda_1^{-1}\Phi_\epsilon(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})$.
  Compute the direction $(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)})$ using (11).
  Compute the Conjugate Gradient update (9), with step size β_t defined in (10) to obtain $(\boldsymbol{L}^{(t)}, R^{(t)})$.
end for
return $\left(\boldsymbol{L}^{(T)}, \mathbf{S}^{(T)}\right)$

Denote by $\boldsymbol{G}_L^{(t-1)} = -\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \boldsymbol{S}^{(t)} - (\boldsymbol{S}^{(t)})^\top) + \epsilon \boldsymbol{L}^{(t-1)}$ the gradient with respect to $\boldsymbol{L}$ of the quadratic part of the objective function, evaluated at $(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})$ and by $\boldsymbol{G}_S^{(t-1)} = -2\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \boldsymbol{S}^{(t-1)} - (\boldsymbol{S}^{(t-1)})^\top) + \epsilon \boldsymbol{S}^{(t-1)}$ the gradient with respect to $\boldsymbol{S}$ of the quadratic part of the objective function, evaluated at $(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)})$. In Algorithm 1, the column-wise sparse component $\boldsymbol{S}$ is updated with a proximal gradient step:

$$\begin{aligned} \boldsymbol{S}^{(t)} & \in \operatorname{argmin} \left(\eta\lambda_2 \|\boldsymbol{S}\|_{2,1} + \tfrac{1}{2} \left\| \boldsymbol{S} - \boldsymbol{S}^{(t-1)} + \eta \boldsymbol{G}_S^{(t-1)} \right\|_F^2 \right), \\ & = \mathsf{Tc}_{\eta\lambda_2} \left(\boldsymbol{S}^{(t-1)} - \eta \boldsymbol{G}_S^{(t-1)} \right), \end{aligned} \tag{8}$$

where $\mathsf{Tc}_{\eta\lambda_2}$ is the column-wise soft-thresholding operator such that for any $\mathbf{M} \in \mathbb{R}^{n\times n}$ and for any $\lambda > 0$, the j-th column of $\mathsf{Tc}_{\lambda}(\mathbf{M})$ is given by $(1 - \lambda/\|\mathbf{M}_{.,j}\|_2)_+\mathbf{M}_{.,j}$. The step size η is constant and fixed in advance, and satisfies $\eta \leq 1/(2+\epsilon)$. The low-rank component given by $(\boldsymbol{L}, R)$ is updated using a conjugate gradient step as follows:

$$\left(\boldsymbol{L}^{(t)}, R^{(t)}\right) = \left(\boldsymbol{L}^{(t-1)}, R^{(t-1)}\right) + \beta_t \left(\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}, \tilde{R}^{(t)} - R^{(t-1)}\right), \tag{9}$$

where $\beta_t \in [0, 1]$ is a step size set to:

$$\beta_t = \min\left\{1, \frac{\langle \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}, \boldsymbol{G}_L^{(t-1)}\rangle + \lambda_1(R^{(t-1)} - \tilde{R}^{(t)})}{(1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2}\right\}. \tag{10}$$

The direction $(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)})$ is defined by:

$$\begin{array}{rll}\left(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)}\right) \in & \operatorname{argmin}_{\mathbf{Z},R} & \langle \mathbf{Z}, \boldsymbol{G}_L^{(t-1)}\rangle + \lambda_1 R \\ & \text{such that} & \|\mathbf{Z}\|_* \leq R \leq \bar{R}^{(t)}.\end{array} \tag{11}$$

Note that, if the matrix $\boldsymbol{L}^{(t)}$ is symmetric, then the matrix $\boldsymbol{L}^{(t+1)}$ remains symmetric at iteration $t+1$. Indeed, the gradient $\boldsymbol{G}_L^{(t)}$ is defined in terms of the matrices $\boldsymbol{A}$, $\boldsymbol{\Omega}$, and $\boldsymbol{S}^{(t)} + (\boldsymbol{S}^{(t)})^\top$, all three symmetric matrices. Therefore, to obtain a symmetric estimator of $\boldsymbol{L}$, it suffices to initialise the algorithm with symmetric $\boldsymbol{L}^{(0)}$.

The Mixed Coordinate Gradient Descent algorithm described in Algorithm 1 converges sublinearly to the optimal solution of (7), as shown by the following result:

Theorem 1. *Let $\delta > 0$. After $T_\delta = \mathcal{O}(1/\delta)$ iterations, the iterate satisfies:*

$$\mathcal{F}_\epsilon(\boldsymbol{S}^{(T_\delta)}, \boldsymbol{L}^{(T_\delta)}) - \mathcal{F}_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon) \leq \delta. \tag{12}$$

In addition, by strong convexity of $\mathcal{F}_\epsilon$,

$$\|\boldsymbol{S}^{(T_\delta)} - \hat{\boldsymbol{S}}_\epsilon\|_F^2 + \|\boldsymbol{L}^{(T_\delta)} - \hat{\boldsymbol{L}}_\epsilon\|_F^2 \leq \frac{2\delta}{\epsilon}. \tag{13}$$

In Appendix A.3 we provide a more detailed result, with an estimation of the constant in $\mathcal{O}(1/\delta)$.

4 Theoretical analysis of the estimator

In this section we provide theoretical analysis of our algorithm. First, we provide guarantees on the support recovery of the outliers. Next, we prove a non asymptotic bound on the risk of our estimator. We start by introducing assumptions on the missing values mechanism.

4.1 Assumption on the sampling scheme

Our first assumption on the sampling scheme requires that all the edges between the inliers are observed with a non-vanishing probability. Recall that $I = \mathcal{I} \times \mathcal{I}$ denote the pairs of inlier nodes.

Assumption 1. *There exist a strictly positive sequence μ_n such that for any $(i,j) \in I$, $\mu_n \leq \boldsymbol{\Pi}_{ij}$.*

Bounding the probabilities of observing any entry away from 0 is a usual assumption in the literature dealing with missing observations (different patterns for missing observations are discussed, e.g., in [33, 38, 47]). We denote by ν_n and $\tilde{\nu}_n$ two sequences such that for any $i \in I$, $\sum_{j\in\mathcal{I}} \boldsymbol{\Pi}_{ij} \leq \nu_n n$ and for any $i \in [n]$, $\sum_{j\in\mathcal{O}} \boldsymbol{\Pi}_{ij} \leq \tilde{\nu}_n s$. We always have $\nu_n \leq 1$ and $\tilde{\nu}_n \leq 1$, but when ν_n and $\tilde{\nu}_n$ are decreasing sequences, we obtain better error rates by taking advantage of the fact that observations are distributed over different nodes in the network. Note that our estimators do not require the knowledge of the sequences μ_n and $\tilde{\nu}_n$. On the other hand, for the theoretical analysis we need an upper bound on $\nu_n\rho_n n$ (the average observed connectivity of inlier nodes), which can be estimated robustly (for example by using Median of Means [39]).

Recall that we do not observe any entry on the diagonal of $\boldsymbol{A}$. Combined with Assumption 1, this implies that for any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$

$$\left\|\boldsymbol{M}_{|I}\right\|_F^2 \leq \frac{1}{\mu_n}\left\|\boldsymbol{M}\right\|_{L_2(\boldsymbol{\Pi})}^2 + n\left\|\boldsymbol{M}\right\|_\infty^2. \tag{14}$$

Moreover, since $|O| = 2ns + (s-1)(s-2)/2 \leq 3ns$, we find that

$$\left\|\boldsymbol{M}_{|O}\right\|_F^2 \leq 3ns\left\|\boldsymbol{M}\right\|_\infty^2. \tag{15}$$

Before stating the second assumption, recall that ρ_n is a sparsity inducing sequence such that $\|\boldsymbol{L}^*\|_\infty \leq \rho_n$. Similarly, we define $\gamma_n = \|\mathbb{E}[\boldsymbol{A}]\|_\infty$. Since $\|\boldsymbol{S}^*\|_\infty \leq \gamma_n$, γ_n characterises the sparsity of connections of the outlier nodes. Note that outliers and inliers may have different sparsity levels, i.e., γ_n and ρ_n may be of different orders of magnitude.

Assumption 2. *$\nu_n\rho_n \geq \log(n)/n$ and $\tilde{\nu}_n\gamma_n \geq \log(n)/n$.*

Assumption 2 implies that the *observed* average node degree is not too small. Note that considering very sparse graphs, where the expectation of the probability of observing an edge is of order $\frac{1}{n}$, is of lesser interest since it has been shown in [20] that the trivial null estimator is minimax optimal in this setting. On the other hand, the sparsity threshold $\log(n)/n$ is known to correspond to phase transition phenomenons for recovering structural properties in the SBM [1]. We also need the following assumption on the "signal to noise ratio".

Assumption 3. *$\nu_n\rho_n n \geq \tilde{\nu}_n\gamma_n s$*

Here, edges connecting inliers to inliers can be seen as a "signal term" in the estimation of connection probabilities, while edges connecting outliers to any other nodes can be seen as a "noise term". Now, recall that ρ_n bounds the probability of any inlier to be connected to any inlier, while γ_n bounds the probability of any inlier to be connected to any outlier. Then, Assumption 3 requires that we observe more connection between inliers than between inliers and outliers, or equivalently that the "signal" induced by the connections of the inliers be stronger than the "noise". For example, under a uniform sampling, all entries are observed with the same probability, so $\mu_n = \nu_n = \tilde{\nu}_n = p$. Then, Assumption 3 becomes $\rho_n n \geq \gamma_n s$, and requires that inlier nodes be more connected with other inlier nodes than with outliers. As the number of outliers s is typically much smaller than the number of inlier nodes, $n - s$, this assumption is not restrictive.

4.2 Outlier detection

The $\|\cdot\|_{2,1}$-norm penalisation induces the column-wise sparsity of the estimator $\widehat{\boldsymbol{S}}$ (when appropriately calibrated, it allows only a small number of columns of $\widehat{\boldsymbol{S}}$ to be non-zero). Using this sparsity, we define the set of estimated outliers as

$$\widehat{\mathcal{O}} \triangleq \left\{j \in [n] : \widehat{\boldsymbol{S}}_{\cdot,j} \neq \boldsymbol{0}\right\}. \tag{16}$$

The following lemma, proven in Appendix A.8.1, provides a characterisation of this set:

Lemma 1. *For any $j \in [n]$, $\widehat{\boldsymbol{S}}_{\cdot,j} \neq \boldsymbol{0} \;\Leftrightarrow\; \left\|\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)_+\right\|_2 > \frac{\lambda_2}{2}$.*

Lemma 1 provides a lower bound on λ_2 that will prevent from erroneously reporting inliers as outliers by choosing λ_2 larger than the expected norm of columns corresponding to inliers. Note that for any inlier j, $\mathbb{E}[\|(\boldsymbol{\Omega} \odot (\boldsymbol{A}_{\cdot,j} - \boldsymbol{L}^*_{\cdot,j})_{|I})_+\|_2]$ is of the order $\sqrt{\nu_n\rho_n(n-s) + \tilde{\nu}_n\gamma_n s}$. If λ_2 falls below this threshold, some inliers are likely to be erroneously reported as outliers. Therefore, we choose $\lambda_2 \gtrsim \sqrt{\nu_n\rho_n(n-s) + \tilde{\nu}_n\gamma_n s}$. Under Assumption 3, this condition becomes $\lambda_2 \gtrsim \sqrt{\nu_n\rho_n n}$. With this choice of λ_2 we have the following results proven in Appendix A.4:

Theorem 2. *Let $\lambda_2 = 19\sqrt{\nu_n\rho_n n}$. Then, under Assumptions 1-3, there exists an absolute constant $c > 0$ such that with probability at least $1 - c/n$*

$$\widehat{\mathcal{O}} \cap \mathcal{I} = \emptyset. \tag{17}$$

One cannot hope to further separate outliers from inliers without additional assumptions on how the first group differs from the second one. Here, we provide an intuition about our condition on the connectivity of outliers that is sufficient for outliers detection. According to Lemma 1, any outlier j will be reported as such if $\|(\boldsymbol{\Omega}_{\cdot,j} \odot (\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}))_+\|_2 > \lambda_2/2$. So, in order to detect an outlier j, the threshold λ_2 must be at least smaller than $\mathbb{E}[\|(\boldsymbol{\Omega}_{\cdot,j} \odot (\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}))_+\|_2]$. Recalling that $\widehat{\boldsymbol{L}}$ and $\widehat{\boldsymbol{S}}$ have non-negative entries, we see that

$$\mathbb{E}\left[\left\|\left(\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)\right)_+\right\|_2\right] \leq \mathbb{E}\left[\left\|\left(\boldsymbol{\Omega}_{\cdot,j} \odot (\boldsymbol{A}_{\cdot,j})\right)_+\right\|_F\right] = \sqrt{\sum_{i\in\mathcal{I}} \boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij} + \sum_{i\in\mathcal{O}} \boldsymbol{\Pi}_{ij}(\boldsymbol{S}^*_{ij} + \boldsymbol{S}^*_{ij})}.$$

Thus, the condition $\sqrt{\nu_n\rho_n n} \lesssim \lambda_2 \lesssim \min_{j\in\mathcal{O}} \sqrt{\sum_{i\in\mathcal{I}} \boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}}$ appears naturally when separating the inliers from the outliers. This condition is formalised in the following assumption:

Assumption 4. $\min_{j\in\mathcal{O}} \sum_{i\in\mathcal{I}} \boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij} > C\rho_n\nu_n n$ *where* C *is an absolute constant defined in Section A.5.*

When the outliers represent only a small fraction of the nodes, we have that $|\mathcal{I}| \simeq n$. Then, Assumption 4 is met when outlier nodes have higher expected observed degree than inlier nodes. When the sampling probabilities are uniform, this assumption essentially reads $\gamma_n \geq C\rho_n$. This assumption is compatible with assumption 3, as the number of outliers s is typically much smaller than the number of nodes n. The following Lemma shows that assumption 4 ensures strong identifiability of the set of outliers and inliers.

Lemma 2. *Under assumption 4, the solution* $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ *to equation (3) is unique up to diagonal terms.*

Lemma 2 ensures that under Assumption 4, the set of outliers is well defined. Moreover, all outliers are detected with large probability, as indicated by the following result proven in Appendix A.5:

Theorem 3. *Let* $\lambda_2 = 19\sqrt{\nu_n\rho_n n}$. *Under Assumptions 1-4, there exists an absolute constant* $c > 0$ *such that* $\mathcal{O} = \widehat{\mathcal{O}}$ *with probability at least* $1 - cs/n$.

Theorem 3 provides guarantees on the recovery of the support of the column-sparse component of the decomposition (2). To the best of our knowledge, this is the first result of this sort in the noisy setting, where the exact reconstruction of both components, the low-rank and the sparse one, is impossible. For both Theorem 3 and Theorem 2, we actually show that the results hold with probabilities at least $1 - 8se^{-c_n n}$ and $1 - 6e^{-c_n n}$ respectively, where c_n is a sequence depending on ν_n and ρ_n such that $c_n \geq \log(n)/n$.

4.3 Estimation of the connections probabilities

In this section, we establish the non-asymptotic upper bound on the risk of our estimator. We denote the noise matrix $\boldsymbol{\Sigma} \triangleq \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$. Let $\boldsymbol{\Gamma}$ be the random matrix defined as follows: for any (i,j), $\boldsymbol{\Gamma}_{ij} \triangleq \epsilon_{ij}\boldsymbol{\Omega}_{ij}$, where $\{\boldsymbol{\epsilon}_{ij}\}_{1\leq i<j\leq n}$ is a Rademacher sequence. To clarify the exposition of our results, we introduce the following error terms

$$\Phi \triangleq n\rho_n^2\left(\frac{\nu_n k}{\mu_n} + \nu_n s\right), \quad \Psi \triangleq 16\tilde{\nu}_n\gamma_n\rho_n sn \quad \text{and} \quad \Xi \triangleq \frac{\sqrt{\nu_n n}\rho_n}{\lambda_1} + 1.$$

The following theorem, proven in Appendix A.6, provides the error bound for the risk of the estimator $\widehat{\boldsymbol{L}}$ that depends on the choice of the regularisation parameter λ_1:

Theorem 4. *Assume that* $\lambda_1 \geq 3\left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{op}$, *and that* $\lambda_2 = 19\sqrt{\nu_n\rho_n n}$. *Then, under Assumptions 1-3, there exists absolute constants* $C > 0$ *and* $c > 0$ *such that with probability at least* $1 - c/n$,

$$\left\|\left(\widehat{\boldsymbol{L}} - \boldsymbol{L}^*\right)_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} \quad \leq \quad C\left(\frac{\lambda_1^2 k}{\mu_n} + \Phi + \Xi\Psi\right). \tag{18}$$

Next, we provide a choice for λ_1 such that the condition $\lambda_1 \geq 3\left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{op}$ holds with high probability. To do so, we must first obtain a high-probability bound on $\left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{op}$. This is done in the following Lemma:

Lemma 3. $\mathbb{P}\left(\left\|\boldsymbol{\Omega}\odot\boldsymbol{\Sigma}_{|I}\right\|_{op}\geq 28\sqrt{\nu_n\rho_n n}\right)\leq e^{-\nu_n\rho_n n}$.

Using Lemma 3, we obtain the following corollary proven in Appendix A.7:

Corollary 1. *Choose $\lambda_1 = 84\sqrt{\nu_n\rho_n n}$ and $\lambda_2 = 19\sqrt{\nu_n\rho_n n}$. Then, under the conditions of Theorem 4, there exists absolute constants $C>0$ and $c>0$ such that with probability at least $1-c/n$,*

$$\left\|\left(\widehat{\boldsymbol{L}}-\boldsymbol{L}^*\right)_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} \quad\leq\quad C\left(\frac{\nu_n}{\mu_n}\rho_n kn+(\nu_n\rho_n\vee\tilde{\nu}_n\gamma_n)\rho_n sn\right) \tag{19}$$

and

$$\left\|\left(\widehat{\boldsymbol{L}}-\boldsymbol{L}^*\right)_{|I}\right\|^2_{F} \quad\leq\quad \frac{C}{\mu_n}\left(\frac{\nu_n}{\mu_n}\rho_n kn+(\nu_n\rho_n\vee\tilde{\nu}_n\gamma_n)\rho_n sn\right). \tag{20}$$

Remark 1. *The estimator $(\widehat{\boldsymbol{L}},\widehat{\boldsymbol{S}})$ returned by the MCGD Algorithm does not have the property $\widehat{\boldsymbol{L}}_{\cdot,j}\neq\mathbf{0}\Leftrightarrow\widehat{\boldsymbol{S}}_{\cdot,j}=\mathbf{0}$ (non-overlapping support). To obtain estimators verifying this property, we may define a new estimator $\widehat{\boldsymbol{L}}'$ for $\boldsymbol{L}^*$ such that*

$$\widehat{\boldsymbol{L}}'_{ij}=\begin{cases}\widehat{\boldsymbol{L}}_{ij} \text{ if } j\notin\widehat{\mathcal{O}}\\ 0 \text{ if } j\in\widehat{\mathcal{O}}\end{cases}$$

Note that $\widehat{\boldsymbol{L}}'=\mathcal{P}_{\widehat{\mathcal{I}}\times\widehat{\mathcal{I}}}(\widehat{\boldsymbol{L}})$, where $\widehat{\mathcal{I}}=[n]\backslash\widehat{O}$ is the set of estimated inliers, and $\mathcal{P}_{\widehat{\mathcal{I}}\times\widehat{\mathcal{I}}}$ is the orthogonal projection onto the set of matrices with support in $\widehat{\mathcal{I}}\times\widehat{\mathcal{I}}$. Using Theorem 2, we find that with high probability, $\boldsymbol{L}^*$ has a support in $\widehat{\mathcal{I}}\times\widehat{\mathcal{I}}$. Then, classical properties of orthogonal projections ensure that $\|\boldsymbol{L}^*-\widehat{\boldsymbol{L}}'\|_F\leq\|\boldsymbol{L}^*-\widehat{\boldsymbol{L}}\|_F$. Thus, the new estimator $(\widehat{\boldsymbol{L}}',\widehat{\boldsymbol{S}})$ achieves the same error rate as the estimator $(\widehat{\boldsymbol{L}}',\widehat{\boldsymbol{S}})$ and detects the same outliers, while having non-overlapping support.

To get a better understanding of the results of Corollary 1, we consider the following simple example. We consider a missing data scheme where all entries of $\boldsymbol{A}$ are observed with the same probability p (that is, $\nu_n=\tilde{\nu}_n=\mu_n=p$). Then, the error of our estimator $\hat{\boldsymbol{L}}$ in Frobenius norm is at most $O(\rho_n kn/p+\rho_n\gamma_n sn)$. Assume now that the number of outliers s is bounded by $k/(p\gamma_n)$ (note that when the network is sparse, $\gamma_n\to 0$ and thus the number of outliers may grow). Then, the error rate is of the order $O(\rho_n kn/p)$, which corresponds to the minimax optimal rate for the low-rank matrix estimation problem without outliers. By comparison, applying methods from the low-rank matrix completion literature, we obtain an error rate of the order $O(kn/p)$, which is sub-optimal since ρ_n is typically of the order of $\log(n)/n$.

To the best of our knowledge, no results on robust estimation of the connection probabilities in the presence of outliers and missing observations have been established before. Previous rates of convergence for the problem of estimating the connection probabilities under the Stochastic Block Model with missing links have been established, for the uniform sampling scheme, in [20], and, for more general sampling schemes, in [22]. To compare our bound with these previous results, we consider the case of the uniform sampling and assume that the condition $(\tilde{\nu}_n\mu_n\vee\nu_n\rho_n)\,s\leq\nu_n k/\mu_n$ is met. In [20] and [22], the authors show that the risk of their estimators in $\|\cdot\|_{L_2(\boldsymbol{\Pi})}$-norm is of the order $\rho_n\left(\log(k)n+k^2\right)$, and that it is minimax optimal. The rate provided by Corollary 1 is of the order $\rho_n kn$. So, for the relevant case $k\leq\sqrt{n}$, our method falls short of the minimax optimal rate for this problem by a factor $k/\log(k)$. Note that, estimators proposed in [20] and [22] have non-polynomial computational cost while our estimator can be used in practice. On the other hand, the authors of [64] propose a polynomial-time algorithm for estimating the probabilities of connections in the Stochastic Block Model under complete observation of the network. They show that the risk of their estimator for the connection probabilities is bounded by $C\rho_n kn$. Thus, our method matches the best known rate established for a polynomial time algorithm for the Stochastic Block Model while being robust to missing observations and outliers.

5 Numerical experiments

5.1 Outliers detection

In this section, we illustrate the performance of our method in terms of outliers detection on two different types of outliers: hubs and mixed membership profiles. We start by generating a graph containing $n = 1000$ inlier nodes according to the Stochastic Block Model with three communities of approximately the same size. In each community, the probability of connection between nodes is equal to $p = 0.05$. The probability of connection between communities is equal to $q = 0.01$. With this choice of parameters, the average node degree is of the order of $\log(n)$. Then, we generate $s = 20$ outlier nodes using the following two methods:

1. **Hub**: outlier j connects to any other node i with probability π_{hub}.
2. **Mixed membership**: for any outlier j, we select at random two communities. For any other node i, if it belongs to one of the two communities, outlier j connects to i with probability π_{mix}. Otherwise, it connects to i with probability $q = 0.01$.

Finally, we introduce 20% of missing values in the adjacency matrix uniformly at random. For each of the two types of outliers, we consider increasing values of the ratio

$$\rho = \frac{\min_{j\in\mathcal{O}} \sum_{i\in\mathcal{I}} \mathbf{\Pi}_{ij} \boldsymbol{S}^*_{ij}}{\tilde{\nu}_n \rho_n n},$$

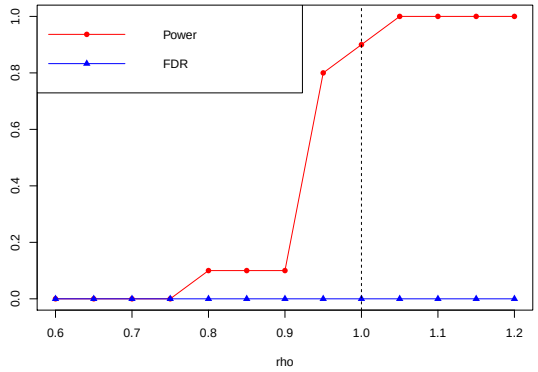
highlighted in Theorem 3 as the crucial quantity to guarantee strong identification of the outliers (see Assumption 4). In our case, it is of the order of $\rho_{\mathsf{hub}} = \frac{3\pi_{\mathsf{hub}}}{p+2q}$ for hubs, and $\rho_{\mathsf{mix}} = \frac{2\pi_{\mathsf{mix}}+q}{p+2q}$ for mixed membership nodes. We fix the size of the network $n = 1000$, the number of outliers $s = 20$ and the connection probability intra and inter communities $p = 0.05$ and $q = 0.02$. Then, we generate outliers with increasing values of π_{hub} and π_{mix} so that the ratios ρ_{hub} and ρ_{mix} spans the range $(0.6, 2)$. For each value of ρ_{hub} and ρ_{mix}, we apply our algorithm to detect outliers, fixing the parameters λ_1 and λ_2 to their theoretical values. The results are presented in Figures 2a and 2b, where we display the power ($\frac{|\hat{\mathcal{O}}\cap\mathcal{O}|}{|\mathcal{O}|}$) and the False Discovery Rate (FDR, $\frac{|\hat{\mathcal{O}}\cap\mathcal{I}|}{|\mathcal{O}|}$) for hubs and mixed membership nodes, respectively. In both cases, the limit $\rho = 1$ is indicated with a dashed black line. Note that, the theoretical detection limit given in Assumption 4 yields $\rho \geq 152 \gg 1$ (see A.5). Thus, our empirical results show that our algorithm is in fact able to detect outliers at much lower signal-to-noise ratio than predicted by theory. In addition we emphasize that, for $\rho = 1$, outliers have approximately the same degree as the inliers, and thus cannot be detected by inspecting the histogram of degrees.

Our numerical results show that for outliers with hubs profiles (Figure 2a), our algorithm successfully detects the outliers, including in "hard" settings where their average degree is the same as inliers. Our simulations also confirm the relevance of our theoretical findings, which highlight the importance of the ratio $\frac{\min_{j\in\mathcal{O}} \sum_{i\in\mathcal{I}} \mathbf{\Pi}_{ij} \boldsymbol{S}^*_{ij}}{\tilde{\nu}_n \rho_n n}$ for outliers detection, even though our theoretical constants may not be optimal. Finally, note that, using the theoretical values of λ_1 and λ_2, our algorithm almost never falsely labels inliers as outliers (FDR is consistently 0). In the case of outliers with mixed membership profiles, we observe a similar behaviour. However, the empirical value of ρ_{mix} required for exact outliers selection is in this case of the order of $\rho_{\mathsf{mix}} \simeq 1.6$, slightly above the observed limit for hubs $\rho_{\mathsf{hub}} \simeq 1$. This seems to indicate that, in practice, mixed membership nodes are "harder" to detect than hubs.

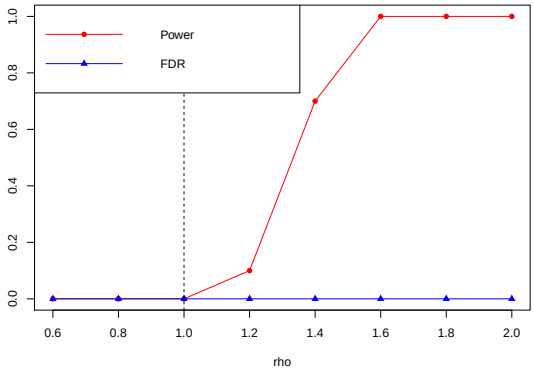
5.2 Estimation of connection probabilities

We now evaluate the performance of our method in terms of estimation of the connection probabilities of inliers. As before, we start by generating a network of size $n = 1000$ using the Stochastic Block Model with three balanced communities. We keep the same parameters for the SBM, with $p = 0.05$ and $q = 0.01$, and introduce 20% of missing values. Then, we study two settings where we introduce s outliers corresponding to hubs and mixed membership nodes, respectively. For each of the two types of outliers, we consider increasing values of the ratio

$$\tau = \frac{\rho_n \nu_n n}{\tilde{\nu}_n \gamma_n s},$$



(a) **Hubs** detection: **Power** (red points) and **FDR** (blue triangles) for increasing $\rho_{\mathsf{hub}} \sim \pi_{\mathsf{hub}}/p$, averaged across 10 replications. $\rho_{\mathsf{hub}} = 1$ indicated with dashed black line.



(b) **Mixed membership** detection: **Power** (red points) and **FDR** (blue triangles) for increasing $\rho_{\mathsf{mix}} \sim \pi_{\mathsf{mix}}/p$, averaged across 10 replications. $\rho_{\mathsf{mix}} = 1$ indicated with dashed black line.

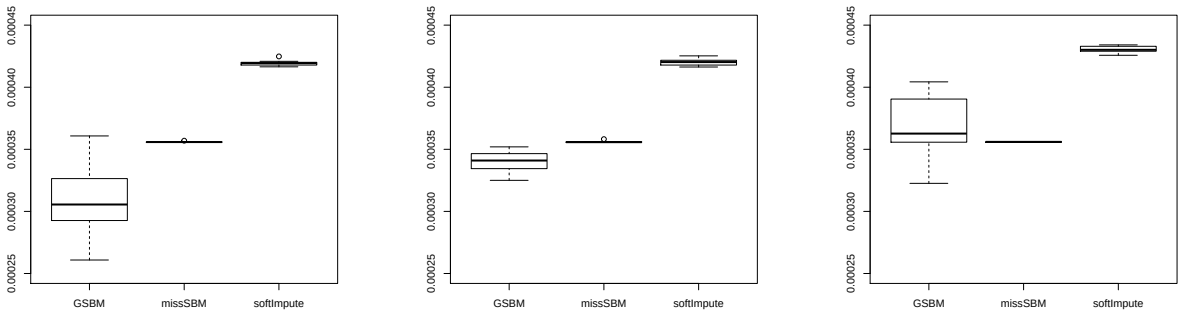
highlighted in Corollary 1 as the signal to noise ratio for the problem of estimation of the connection probabilities of inliers (see Assumption 3). In our case, it is of the order of $\tau_{\mathsf{hub}} = \frac{n(p+2q)}{3s\pi_{\mathsf{hub}}}$ for hubs, and $\tau_{\mathsf{mix}} = \frac{n(p+2q)}{s(2\pi_{\mathsf{mix}}+q)}$ for mixed membership nodes.

We fix the size of the network $n = 1000$, the intra- and inter-communities connection probabilities $p = 0.05$ and $q = 0.02$, and the values $\pi_{\mathsf{hub}} = 0.2$ and $\pi_{\mathsf{mix}} = 0.3$. Note that, these values of π_{hub} and π_{mix} produce outliers which are much more connected than inliers. This corresponds to a setting where the detection of outliers is "easy" because they have large degrees, but the estimation of the connection probabilities of inliers (parameter $\boldsymbol{L}^*$) is "hard" because outliers have many connections polluting the network. Then, we generate an increasing number of outliers ($s = 20$, $s = 50$, $s = 100$), so that the signal to noise ratios τ_{hub} and τ_{mix} take different values (5, 2, and 1). For each value of τ_{hub} and τ_{mix}, we estimate the connection probabilities of inliers, fixing the parameters λ_1 and λ_2 to their theoretical values. In each case, we compare the estimation results with two competitors: the method implemented in the R [49] package `missSBM` [56, 57] which fits a Stochastic Block Model in the presence of missing links, and matrix completion as implemented in the R package `softImpute` [27]; the methods are compared in terms of the Mean Squared Error (MSE) of estimation, normalized by size of the set of pairs inliers $I = \mathcal{I} \times \mathcal{I}$. The MSE is thus defined for some estimator $\hat{\boldsymbol{L}}$ by:

$$\mathrm{MSE}(\hat{\boldsymbol{L}}) = \frac{\left\| \hat{\boldsymbol{L}}_{|I} - \boldsymbol{L}^*_{|I} \right\|_F^2}{|I|}.$$

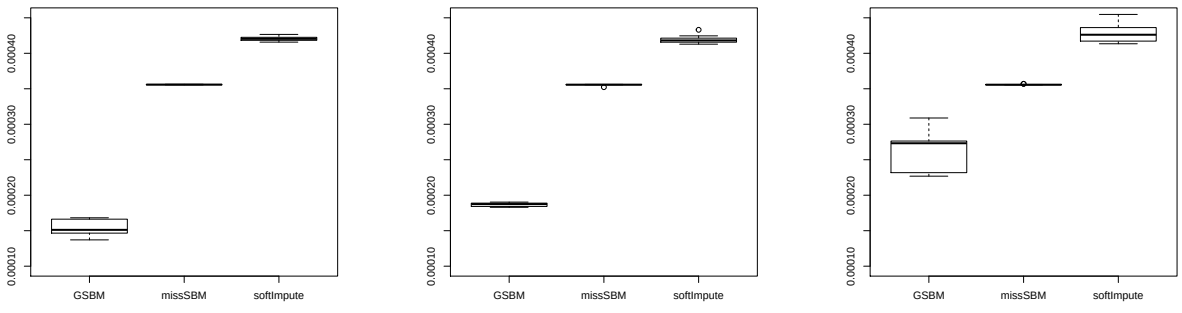
The results are presented in Figure 3 for hubs and Figure 4 for mixed membership nodes, which display (on the same scale) boxplots of the MSE of each method obtained by 10 replications of the experiment, for different values of the signal to noise ratio (from left to right: $\tau = 1$, $\tau = 2$, $\tau = 5$).

For the hubs (Figure 3), we observe that `missSBM` and `softImpute` have similar estimation errors across all settings; overall `missSBM` gives an MSE 20% smaller than `softImpute`. For the large signal to noise ratio $\tau_{\mathsf{hub}} = 5$ where outliers do not impair too much the estimation, our method `gsbm` gives similar results as `missSBM`, but displays a larger variance. However, as the signal to noise ratio τ_{hub} decreases, i.e., in the settings where outliers severely challenge the estimation problem, our method `gsbm` improves over `missSBM` by about 15%. For the mixed membership outliers (Figure 4), we observe that our method `gsbm` consistently improves other methods by 30 to 50%. As in the previous experiment with hubs, we observe that the improvement of `gsbm` over existing methods increases when the signal to noise ratio τ_{mix} decreases i.e., in the settings where outliers are the most challenging for the estimation of connection probabilities.



(a) $\tau_{\mathsf{hub}} = 1$ (100 outliers) (b) $\tau_{\mathsf{hub}} = 2$ (50 outliers). (c) $\tau_{\mathsf{hub}} = 5$ (20 outliers).

Figure 3: **Hubs**: Estimation of connection probabilities of inliers, for different numbers of outliers (left: $s = 100$, middle: $s = 50$, right: $s = 20$) corresponding to three signal to noise ratios (left: $\tau_{\mathsf{hub}} = 1$, middle: $\tau_{\mathsf{hub}} = 2$, right: $\tau_{\mathsf{hub}} = 5$). For each of the three plots, we compare our package `gsbm` to two `missSBM` [56, 57] and `softImpute` in terms of the standardized MSE of estimation $\|\hat{\boldsymbol{L}} - \boldsymbol{L}^*\|_F^2/(n-s)^2$ (10 replications).



(a) $\tau_{\mathsf{mix}} = 1$ (100 outliers) (b) $\tau_{\mathsf{mix}} = 2$ (50 outliers). (c) $\tau_{\mathsf{mix}} = 5$ (20 outliers).

Figure 4: **Mixed membership**: Estimation of connection probabilities of inliers, for different numbers of outliers (left: $s = 100$, middle: $s = 50$, right: $s = 20$) corresponding to three signal to noise ratios (left: $\tau_{\mathsf{mix}} = 1$, middle: $\tau_{\mathsf{mix}} = 2$, right: $\tau_{\mathsf{mix}} = 5$). For each of the three plots, we compare our package `gsbm` to two `missSBM` [56, 57] and `softImpute` in terms of the standardized MSE of estimation $\|\hat{\boldsymbol{L}} - \boldsymbol{L}^*\|_F^2/(n-s)^2$ (10 replications).

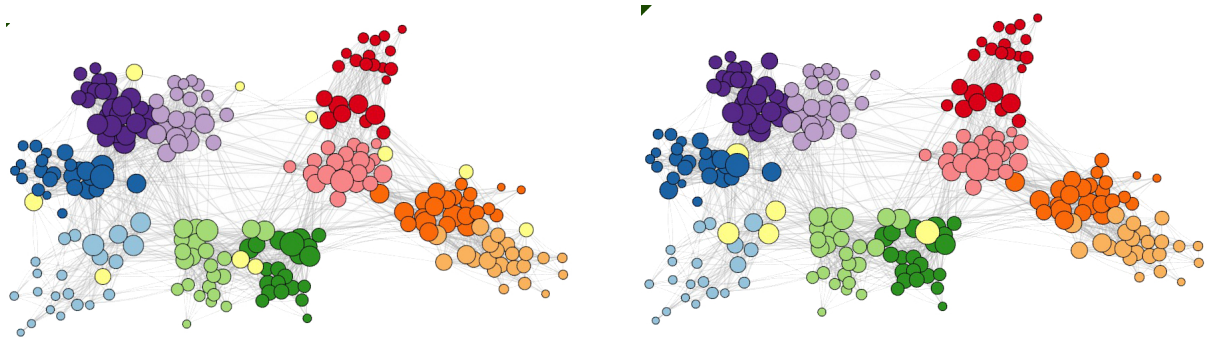


Figure 5: Network of interactions between children and teachers in a primary school over the course of a day. Only interactions lasting at least one minutes are represented. On the left, the colour of each node correspond to its class (if the node represent a child); teachers are indicated by yellow dots. On the right, the nodes are coloured according to estimated labels (the procedure for obtaining these estimated labels is described in Section 5.3). Yellow dots indicate nodes identified as outliers by MCGD.

5.3 Analysis of a contact network in a primary school

Next, we apply our algorithm to analyse a network of contacts within a french elementary school, collected and analysed by the authors of [54], with the objective of better understanding the propagation of respiratory infections. The network records physical interactions occurring within a primary school between 226 children divided into 10 classes and their 10 teachers over the course of a day; it was collected using a system of sensors worn by the participants. This system records the duration of interactions between two individuals facing each other at a maximum distance of one and a half metres. The duration of these interactions varies between 20 seconds and two and a half hours. We consider that a physical interaction has been observed if the corresponding interaction duration is greater than one minute. If an interaction of less than one minute is observed, we consider that this observation may be erroneous, and treat the corresponding data as missing. We thus obtain a 236×236 adjacency matrix with 7054 missing entries (including 236 diagonal entries), and 4980 entries equal to 1 (corresponding to 2490 observed undirected edges). The corresponding network is represented in Figure 5. The analysis of the interactions network provides crucial information from an epidemiological point of view, as it can be used to model the transmission of respiratory-spread pathogens, and design strategies to mitigate the propagation of diseases [23]. Interestingly, the interactions recorded in [54] are strongly structured into communities, as pupils interact mostly with pupils from their class. They also interact with other pupils from the same level, although less frequently. They are scarcely connected with pupils from other age groups. Finally, we observe that pupils are on average connected to one teacher: each one of the ten classes interacts mostly with its teacher. By contrast, the teachers form a smaller group (there are 10 teachers, while there are around 22.6 pupils in each class); yet they do not form a cluster, as they are mostly connected to pupils from their class.

The MCGD algorithm allows us to detect individuals with abnormal connectivities. We run this algorithm for a grid of values of λ_1 and λ_2. We notice that the set of nodes detected as outliers is stable when the parameters (λ_1, λ_2) are chosen around $(8.5, 8)$, and that it contains the nodes $15, 30, 25, 94$ and 180. We also note that those nodes are detected as outliers significantly more frequently than the other nodes when the parameters vary. In the following, we consider estimators $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{S}})$ obtained by running MCGD for this choice of parameters. In order to gain insights on the connectivity of these nodes, we compute the frequencies of their interactions with individuals from each class. Table 1 presents our findings. Nodes 30, 35 and 180 belong to the class "CE1 A", node 180 belongs to the class "CE1 B", and node 15 belongs to the class "CE2 B". We notice from Table 1 the pupil 30 is 4 times more connected to students from class "CE1 B" and 5 times more in contact with students from class "CP B" than the other students from his class. Similarly, we observe that the children identified as outliers are strongly connected with different groups. Thus, the outliers detected by the method correspond to nodes with mixed membership. From a public health perspective, these children can potentially act as super-propagators, and contribute to spreading a

	CE1 A	CE1 B	CE2 A	CE2 B	CM1 A	CM1 B	CM2 A	CM2 B	CP A	CP B	Teachers
CE1 A	0.68	0.07	0.07	0.01	0.00	0.00	0.00	0.00	0.03	0.03	0.10
node 30	0.93	0.30	0.23	0.08	0.00	0.00	0.05	0.00	0.12	0.15	0.11
node 35	0.62	0.24	0.35	0.00	0.06	0.00	0.00	0.05	0.13	0.22	0.11
node 180	1.00	0.43	0.28	0.00	0.00	0.00	0.00	0.00	0.17	0.12	0.12
CE1 B	0.07	0.83	0.02	0.00	0.00	0.01	0.00	0.00	0.05	0.08	0.11
node 94	0.11	0.95	0.06	0.00	0.10	0.00	0.00	0.00	0.22	0.38	0.10
CE2 B	0.01	0.00	0.21	0.94	0.05	0.03	0.00	0.03	0.02	0.00	0.13
node 15	0.00	0.00	0.56	1.00	0.33	0.09	0.00	0.12	0.00	0.00	0.10

Table 1: Frequency of contacts between either a node or an individual from a given class and other individuals from a given class. On average, a pupil from class "CE1 A" has been in contact with a fraction 0.68 of the remaining pupils from his class. By contrast, the node 30, who is in class "CE1 A", is connected with a fraction 0.93 of the remaining pupils from his class.

virus from one group to the others.

Finally, we demonstrate that our estimator for the matrix of connection probabilities $\widehat{\boldsymbol{L}}$ contains significant information on the structure of the network. More precisely, we show that the communities corresponding to the different classes can be recovered from this estimator. To do so, we consider the matrix whose columns contain the 10 left singular vectors of $\widehat{\mathbf{L}}$, and we estimate the classes of the different nodes by running a 10-means algorithm on its rows. This method recovers perfectly the classes of the children considered as "inliers" (up to a permutation of the labels of the classes). While this method is not able to identify teachers, we note that teachers are mapped to the classes in a one-to-one fashion, which indicates that this method succeeds in assigning each class to its teacher. We represent the classes estimated by this method and the nodes identified as outliers in Figure 5.

5.4 Analysis of a political Twitter network

The "#Élysée2017fr" data set, originally introduced in [18] provides data about 22,853 Twitter profiles active during the campaign of the French 2017 presidential election, from November 2016 to May 2017. Among other data, it contains a mentions network, where each node corresponds to a Twitter profile, and a directed edge (from mentioning profile to mentioned one) connects two profiles whenever one of them mentions the other in a Tweet. In total, this amounts to 1,896,262 edges. In the original study, the authors of [18] highlighted a community structure, where communities roughly correspond to affiliations to the 5 main political parties in France: France Insoumise (FI), Parti Socialiste (PS), Les Républicain (LR), La République en Marche (LREM), Rassemblement National (RN), with preferential attachment between nodes of the same political party. Detecting outliers in this network is of interest in order to detect, for example, influential figures. We apply our algorithm to detect potential outliers to a subsample of this network containing the 10,000 most connected nodes; we also make the network undirected by drawing an edge between two nodes whenever one of them mentions the other. After subsampling and symmetrization of the adjacency matrix, the number of edges in the network is 1,562,419. Using the estimated theoretical values of the regularization parameters λ_1 and λ_2, we detect around 600 outliers in the network. Inspecting the corresponding 600 profile annotations, and node degrees we observe that the detected outliers correspond mainly to densely connected hubs or to mixed membership profiles (i.e. profiles affiliated to at least two political parties).

Hubs First of all, we detect large hubs corresponding to main political figures and large media. The first detected outliers are the Twitter profiles of candidates to the election: Emmanuel Macron, Marine Le Pen, François Fillon, Jean-Luc Mélenchon, Benoît Hamon, Nicolas Dupont-Aignan. Other detected private personalities include journalists, deputies and senators (Jean-Jacques Bourdin, Alexis Corbière, Benjamin Griveaux, Yannick Jadot, Richard Ferrand, Éric Ciotti, etc.). Secondly, we detect the Twitter profiles of high-circulation media: BFM TV, Le Figaro, Le Monde, Libération, Mediapart, France Info, Europe 1, France Inter, etc. We also note that some hubs correspond to online, unofficial political groups (@TeamProgressist, @ForceRep_fr, @Presse2Droite, @nomacron, etc.).

Mixed membership We also detect Twitter profiles corresponding to nodes of mixed membership affiliated to multiple parties. Some of these nodes also correspond to smaller hubs, such as Christine Boutin (LR/RN) and La Manif Pour Tous (LR/RN); they have smaller degrees than the main political figures and media (degree around 1000 rather than >5000 for the main hubs). We also find mixed membership profiles corresponding to individual profiles with no public exposition (e.g. @mrericmas: LREM/LR, @erayeye: LR/RN, @Apostillier1: LREM/PS, etc.). After inspecting the Twitter profiles, these seem to be individuals sharing their own political opinions on Twitter, which would not necessarily be detected by checking only the histogram of degrees.

6 Conclusion

In this paper, we have proposed a new, computationally efficient algorithm for detecting nodes with anomalous connection patterns. This algorithm, which is robust against missing observations, allows for simultaneous estimation of the probabilities of connections of the remaining, normal nodes. A convergence analysis of this algorithm is provided, which proves that this algorithm converges at a sub-linear rate. Moreover our simulations studies indicate that its running time remains moderate, even for networks containing a few thousands of nodes. Our theoretical results show that our method detects exactly the outliers under fairly general assumptions. Moreover, our estimator for the probabilities of connections achieves the best known error rate among estimator with polynomial running time. These results are supported by simulation studies which demonstrate the good properties of our estimators in terms of both outliers detection and link prediction. Finally, we have exemplified how our method can be used to detect outlier nodes and recover structural information on the remaining nodes in real world networks, by applying this method to "Les Misérables" characters network, as well as a network of interactions taking place in a primary school, and on a political Twitter network. The results of the present paper pave the way to several extensions, which are of interest in applications. In particular, an important one would be the generalisation of the model to dynamic networks, where the adjacency matrix is observed at multiple time points, and connections, outliers, and possibly underlying communities are allowed to vary across time. This is interesting in applied problems where the outliers have characteristic dynamic behaviour. For instance, to detect fake news in social networks where, contrary to regular users, malicious users tend to have very unstable connectivity patterns across time.

References

[1] Abbe, E. (2018). Community detection and stochastic block models: Recent developments. Journal of Machine Learning Research, 18(177):1–86.

[2] Adewole, K. S., Anuar, N. B., Kamsin, A., Varathan, K. D., and Razak, S. A. (2017). Malicious accounts: Dark of the social networks. Journal of Network and Computer Applications, 79:41 – 67.

[3] Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. Ann. Statist., 40(2):1171–1197.

[4] Akoglu, L., Chandy, R., and Faloutsos, C. (2013). Opinion fraud detection in online reviews by network effects. International AAAI Conference on Web and Social Media.

[5] Bandeira, A. and van Handel, R. (2014). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. The Annals of Probability, 44.

[6] Beck, A. and Tetruashvili, L. (2013). On the convergence of block coordinate descent type methods. SIAM Journal on Optimization, 23(4):2037–2060.

[7] Bleakley, K., Biau, G., and Vert, J.-P. (2007). Supervised reconstruction of biological networks with local models. Bioinformatics, 23(13):i57–i65.

[8] Boucheron, S., Lugosi, G., and Massart, P. (2013). Concentration Inequalities: A Nonasymptotic Theory of Independence. OUP Oxford.

[9] Boyd, S. and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press.

[10] Cai, T. T. and Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. Ann. Statist., 43(3):1027–1059.

[11] Candès, E., Li, X., Ma, Y., and Wright, J. (2009). Robust principal component analysis? Journal of ACM, 58(1):1–37.

[12] Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. (2011). Crawling facebook for social network analysis purposes. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS '11, New York, NY, USA. Association for Computing Machinery.

[13] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. SIAM J. Optim., 21(2):572–596.

[14] Chen, Y., Jalali, A., Sanghavi, S., and Caramanis, C. (2013). Low-rank matrix recovery from errors and erasures. IEEE Transactions on Information Theory, 59(7):4324 –4337.

[15] D. Hoff, P. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference, 20.

[16] Dai, H., Zhu, F., Lim, E., and Pang, H. (2012). Detecting anomalies in bipartite graphs with mutual dependency principles. In 2012 IEEE 12th International Conference on Data Mining, pages 171–180.

[17] Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002). Food-web structure and network theory: The role of connectance and size. Proceedings of the National Academy of Sciences of the United States of America, 99(20):12917–12922.

[18] Fraisier, O., Cabanac, G., Pitarch, Y., Besançon, R., and Boughanem, M. (2018). #Élysée2017fr: The 2017 french presidential campaign on twitter.

[19] Fu, X., Seo, E., Clarke, J., and Hutchinson, R. A. (2019). Link prediction under imperfect detection: Collaborative filtering for ecological networks.

[20] Gao, C., Lu, Y., Ma, Z., and Zhou, H. H. (2016). Optimal estimation and completion of matrices with biclustering structures. J. Mach. Learn. Res., 17(1):5602–5630.

[21] Garber, D., Sabach, S., and Kaplan, A. (2018). Fast generalized conditional gradient method with applications to matrix recovery problems. arXiv e-prints, page arXiv:1802.05581.

[22] Gaucher, S. and Klopp, O. (2019). Maximum likelihood estimation of sparse networks with missing observations. Arxiv preprint, page arXiv:1902.10605.

[23] Gemmetto, V., Barrat, A., and Cattuto, C. (2014). Mitigation of infectious disease at school: targeted class closure vs school closure. BMC Infectious Diseases, 14(1):695.

[24] Giné, E. and Nickl, R. (2016). Mathematical foundations of infinite-dimensional statistical models. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

[25] Guimerà, R. and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 106(52):22073–22078.

[26] Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. Ann. Appl. Stat., 4(1):5–25.

[27] Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. J. Mach. Learn. Res., 16(1):3367–3402.

[28] Hawkins, D. (1980). Identification of Outliers. Chapman and Hall.

[29] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social Networks, 5(2):109 – 137.

[30] Hsu, D., Kakade, S. M., and Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. IEEE Transactions on Information Theory, 57(11):7221–7234.

[31] Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Dasgupta, S. and McAllester, D., editors, Proceedings of the 30th International Conference on Machine Learning, volume 28 of Proceedings of Machine Learning Research, pages 427–435, Atlanta, Georgia, USA. PMLR.

[32] Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. Phys. Rev. E, 83:016107.

[33] Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. Bernoulli, 20(1):282–303.

[34] Klopp, O., Lounici, K., and Tsybakov, A. B. (2017a). Robust matrix completion. Probability Theory and Related Fields, 169(1):523–564.

[35] Klopp, O., Tsybakov, A. B., and Verzelen, N. (2017b). Oracle inequalities for network models and sparse graphon estimation. Ann. Statist., 45(1):316–354.

[36] Knuth, D. E. (1993). The Stanford GraphBase: A Platform for Combinatorial Computing. ACM, New York, NY, USA.

[37] Koltchinskii, V. (2011). Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer Berlin Heidelberg.

[38] Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist., 39(5):2302–2329.

[39] Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means : theory and practice. arXiv e-prints, page arXiv:1711.10306.

[40] Li, T., Levina, E., and Zhu, J. (2020). Community models for partially observed networks from surveys.

[41] Li, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. Constructive Approximation, 37(1):73–99.

[42] Li, X. and Chen, H. (2013). Recommendation as link prediction in bipartite graphs. Decis. Support Syst., 54(2):880–890.

[43] Liu, D., Mei, B., Chen, J., Lu, Z., and Du, X. (2015). Community based spammer detection in social networks. In Dong, X. L., Yu, X., Li, J., and Sun, Y., editors, Web-Age Information Management, pages 554–558, Cham. Springer International Publishing.

[44] Lomnitz, L. A. (1977). Networks of reciprocal exchange. In Lomnitz, L. A., editor, Networks and Marginality, pages 131 – 158. Academic Press.

[45] Mu, C., Zhang, Y., Wright, J., and Goldfarb, D. (2016). Scalable robust matrix recovery: Frank-wolfe meets proximal methods. ArXiv, abs/1403.7588.

[46] Mulamba, D., Ray, I., and Ray, I. (2016). Sybilradar: A graph-structure based framework for sybil detection in on-line social networks. In Hoepman, J.-H. and Katzenbeisser, S., editors, ICT Systems Security and Privacy Protection, pages 179–193, Cham. Springer International Publishing.

[47] Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. J. Mach. Learn. Res., 13:1665–1697.

[48] Nolin, D. A. (2010). Food-sharing networks in lamalera, indonesia. Human Nature, 21(3):243–268.

[49] R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[50] Robin, G., Klopp, O., Josse, J., Éric Moulines, and Tibshirani, R. (2019). Main effects and interactions in mixed and incomplete data frames. Journal of the American Statistical Association, 0(0):1–12.

[51] Robin, G., Wai, H.-T., Josse, J., Klopp, O., and Moulines, E. (2018). Low-rank interaction with sparse additive effects model for large data frames. In Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS'18, pages 5501–5511, USA. Curran Associates Inc.

[52] Shen, J., Awasthi, P., and Li, P. (2019). Robust matrix completion from quantized observations. In Chaudhuri, K. and Sugiyama, M., editors, Proceedings of Machine Learning Research, volume 89 of Proceedings of Machine Learning Research, pages 397–407. PMLR.

[53] Shrivastava, N., Majumder, A., and Rastogi, R. (2008). Mining (social) network graphs to detect random link attacks. 2008 IEEE 24th International Conference on Data Engineering, pages 486–495.

[54] Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., and Vanhems, P. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. PLOS ONE, 6(8):1–13.

[55] Sundar Mukherjee, S. and Chakrabarti, S. (2019). Graphon Estimation from Partially Observed Network Data. arXiv e-prints, page arXiv:1906.00494.

[56] Tabouy, T., Barbillon, P., and Chiquet, J. (2017). Variational Inference for Stochastic Block Models from Sampled Data. ArXiv e-prints.

[57] Tabouy, T., Barbillon, P., and Chiquet, J. (2019). misssbm: An r package for handling missing values in the stochastic block model.

[58] Thomas, M., Verzelen, N., Barbillon, P., Coomes, O., Caillon, S., Mckey, D., Elias, M., Garine, E., Raimond, C., Dounias, E., Jarvis, D., Wencelius, J., Leclerc, C., Labeyrie, V., pham hung, C., Hue, N., Sthapit, B., Rana, R., Barnaud, A., and Massol, F. (2015). A network-based method to detect patterns of local crop biodiversity: Validation at the species and infra-species levels. In Woodward, G. and Bohan, D. A., editors, Ecosystem Services, volume 53 of Advances in Ecological Research, pages 259 – 320. Academic Press.

[59] Tsai, C.-W., Lai, C.-F., Chao, H.-C., and Vasilakos, A. (2015). Big data analytics: A survey. Journal of Big Data, 2.

[60] Viswanath, B., Post, A., Gummadi, K. P., and Mislove, A. (2010). An analysis of social network-based sybil defenses. SIGCOMM Comput. Commun. Rev., 40(4):363–374.

[61] Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. (2003). Epidemic spreading in real networks: an eigenvalue viewpoint. 22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings., pages 25–34.

[62] Wu, Y.-J., Levina, E., and Zhu, J. (2018). Link prediction for egocentrically sampled networks. ArXiv, abs/1803.04084.

[63] Xu, H., Caramanis, C., and Sanghavi, S. (2012). Robust PCA via outlier pursuit. IEEE Trans. Inform. Theory, 58(5):3047–3064.

[64] Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In Dy, J. and Krause, A., editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 5433–5442, Stockholmsmässan, Stockholm Sweden. PMLR.

[65] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. (2014). Uncovering social network sybils in the wild. ACM Trans. Knowl. Discov. Data, 8(1).

[66] Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. Science, 322(5898):104–110.

[67] Yu, H., Kaminsky, M., Gibbons, P. B., and Flaxman, A. D. (2008). Sybilguard: Defending against sybil attacks via social networks. IEEE/ACM Transactions on Networking, 16(3):576–589.

[68] Zhao, Y., Wu, Y.-J., Levina, E., and Zhu, J. (2017). Link prediction for partially observed networks. Journal of Computational and Graphical Statistics, 26(3):725–733.

A Proofs

The proofs are presented as follows. First, we recall in Section A.1 some results that will be used in our proofs. In Section A.2 we provide the details of the Algorithm 1. Section A.3 is devoted to the study of the convergence of our algorithm. Theorem 2 is proved in Section A.4, Theorem 3 is proved in Section A.5, while in Section A.6 we prove Theorem 4. Corollary 1 is proved in Sections A.7. Auxiliary Lemmas used throughout these sections are proved in Section A.8.

To ease notations, we denote henceforth by $\Delta \boldsymbol{S} = \boldsymbol{S}^* - \widehat{\boldsymbol{S}}$ and $\Delta \boldsymbol{L} = \boldsymbol{L}^* - \widehat{\boldsymbol{L}}$ the estimation errors of $\boldsymbol{S}^*$ and $\boldsymbol{L}^*$.

A.1 Tools

In our proofs, we will use Bernstein's inequality on different occasions. We state it here for the reader's convenience.

Theorem 5 (Bernstein's inequality)**.** *Let $X_1, ..., X_n$ be independent centered random variables. Assume that for any $i \in [n]$, $|X_i| \leq M$ almost surely, then*

$$\mathbb{P}\left(\left|\sum_{1\leq i\leq n} X_i\right| \geq \sqrt{2t\sum_{1\leq i\leq n}\mathbb{E}[X_i^2]} + \frac{2M}{3}t\right) \leq 2e^{-t} \tag{21}$$

We will also use Bousquet's theorem, as stated in [24], Theorem 3.3.16.

Theorem 6 (Bousquet)**.** *Let X_i, $i \in \mathbb{N}$ be independent $\mathcal{S}$-valued random variables, and let $\mathcal{F}$ be a countable class of functions $f = (f_1, ..., f_n) : \mathcal{S} \to [-1, 1]^n$ such that $\mathbb{E}[f_i(X_i)] = 0$ for any $f \in \mathcal{F}$ and $i \in [n]$. Set $Z = \sup_{f\in\mathcal{F}}\left|\sum_{1\leq i\leq n} f_i(X_i)\right|$ and $v = \sup_{f\in\mathcal{F}}\sum_{1\leq i\leq n}\mathbb{E}\left[f_i(X_i)^2\right]$. Then, for any $x > 0$,*

$$\mathbb{P}\left(Z > \mathbb{E}[Z] + \frac{x}{3} + \sqrt{2x(2\mathbb{E}[Z] + v)}\right) \leq \exp(-x).$$

To bound the operator norm of random matrices with high probability, we use Corollary 3.6 in [5].

Proposition 1 (Bandeira, Van Handel, 2016)**.** *Let $\boldsymbol{X}$ be a $n\times n$ symmetric random matrix with $\boldsymbol{X}_{ij} = \xi_{ij}b_{ij}$, where $\{\xi_{ij}\}_{i\leq j}$ are independant symmetric random variables with unit variance, and $\{b_{ij}\}_{i\leq j}$ are fixed scalars. Let $\sigma \triangleq \max_i \sqrt{\sum_j b_{ij}^2}$, then for any $\alpha \geq 3$*

$$\mathbb{E}\left[\|\boldsymbol{X}\|_{op}\right] \leq e^{\frac{2}{3}}\left(2\sigma + 14\alpha\max_{ij}\left(\mathbb{E}\left[(\xi_{ij}b_{ij})^{2\alpha}\right]\right)^{\frac{1}{2\alpha}}\sqrt{\log(n)}\right).$$

The following high-probability bound on the spectral norm of a random matrix is based on Remark 3.13 in [5]. This remark provides a bound up to an unspecified absolute constant. In order to make this constant explicit, we follow the lines of the proof of this remark, and we combine Theorem 6.10 in [8], Proposition 1, and a symetrization argument (see, e.g., Corollary 3.3 in [5]) to obtain the following proposition.

Proposition 2. *Let $\boldsymbol{X}$ be an $n \times n$ symmetric matrix with $\boldsymbol{X}_{ij} = \xi_{ij} b_{ij}$, where $\{\xi_{ij}\}_{i \leq j}$ are independent centered random variables with unit variance, and $\{b_{ij}\}_{i \leq j}$ are fixed scalars. Then for every $t \geq 0$ and every $\alpha \geq 3$,*

$$\mathbb{P}\left(\|\boldsymbol{X}\|_{op} \geq 2e^{\frac{2}{3}} \left(2\sigma + 14\alpha \max_{ij} \left(\mathbb{E}\left[(\xi_{ij} b_{ij})^{2\alpha} \right] \right)^{\frac{1}{2\alpha}} \sqrt{\log(n)} \right) + t \right) \leq e^{-t^2 / 2\tilde{\sigma^*}^2}$$

where we have defined $\tilde{\sigma^*} \triangleq \max_{ij} |\boldsymbol{X}_{ij}|$ and $\sigma \triangleq \max_i \sqrt{\sum_j b_{ij}^2}$.

Proof. To prove the desired high-probability bound, we first bound the expectation of the spectral norm, using the same symmetrization trick as in Corollary 3.3 in [5]. Let $\boldsymbol{X}^{'}$ be an independent copy of the random matrix $\boldsymbol{X}$, and let $\boldsymbol{Y}$ be the symmetric matrix with random entries defined as $\boldsymbol{Y}_{ij} \triangleq \boldsymbol{X}_{ij} - \boldsymbol{X}^{'}_{ij}$ for any $(i,j) \in [n] \times [n]$. Note that, for any $(i,j) \in [n] \times [n]$, $i < j$, $\boldsymbol{Y}_{ij} = \sqrt{2} b_{ij} \times \left(\xi_{ij} - \xi^{'}_{ij} \right) / \sqrt{2}$, where $\boldsymbol{\xi}_{ij}$ are independent copies of ξ_{ij}, and $\left(\xi_{ij} - \xi^{'}_{ij} \right) / \sqrt{2}$ are symmetric random variable with unit variance. Applying Proposition 1, we find that

$$\mathbb{E}\left[\|\boldsymbol{Y}\|_{op} \right] \leq e^{\frac{2}{3}} \left(2\sigma_Y + 14\alpha \max_{ij} \left(\mathbb{E}\left[\left(\left(\xi_{ij} - \xi^{'}_{ij} \right) b_{ij} \right)^{2\alpha} \right] \right)^{\frac{1}{2\alpha}} \sqrt{\log(n)} \right)$$

with $\sigma_Y \triangleq \max_i \sqrt{\sum_j 2 b_{ij}^2} = \sqrt{2}\sigma$. Moreover for any $(i,j) \in [n] \times [n]$, $\left(\mathbb{E}\left[\left(\left(\xi_{ij} - \xi^{'}_{ij} \right) b_{ij} \right)^{2\alpha} \right] \right)^{\frac{1}{2\alpha}} \leq 2 \left(\mathbb{E}\left[(\xi_{ij} b_{ij})^{2\alpha} \right] \right)^{\frac{1}{2\alpha}}$. Recall that $\boldsymbol{X}$ is centered. Then, by Jensen inequality, $\mathbb{E}\left[\|\boldsymbol{X}\|_{op} \right] = \mathbb{E}\left[\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|_{op} \right] \leq \mathbb{E}\left[\left\| \boldsymbol{X} - \boldsymbol{X}^{'} \right\|_{op} \right] = \mathbb{E}\left[\|\boldsymbol{Y}\|_{op} \right]$. Hence,

$$\mathbb{E}\left[\|\boldsymbol{X}\|_{op} \right] \leq 2e^{\frac{2}{3}} \left(2\sigma + 14\alpha \max_{ij} \left(\mathbb{E}\left[(\xi_{ij} b_{ij})^{2\alpha} \right] \right)^{\frac{1}{2\alpha}} \sqrt{\log(n)} \right). \tag{22}$$

Then, we use Talagrand's concentration inequality (see [8], Theorem 6.10) and find that for any $t > 0$,

$$\mathbb{P}\left[\|\boldsymbol{X}\|_{op} \geq \mathbb{E}\|\boldsymbol{X}\|_{op} + t \right] \leq e^{\frac{-t^2}{2\tilde{\sigma^*}}} \tag{23}$$

Combining equations (22) and (23) yields the desired result. □

A.2 Mixed coordinate gradient descent algorithm

Below, we describe the details of our algorithm. At iteration $t = 0$, we initialize the parameters $(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)})$; then, at iteration $t \geq 1$, we start by updating $\boldsymbol{S}$. Denote by $\boldsymbol{G}_S^{(t-1)} = -2\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \boldsymbol{S}^{(t-1)} - (\boldsymbol{S}^{(t-1)})^\top) + \epsilon \boldsymbol{S}^{(t-1)}$ the gradient with respect to $\boldsymbol{S}$ of the quadratic part of the objective function, evaluated at $(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)})$. The column-wise sparse component $\boldsymbol{S}$ is updated with a proximal gradient step:

$$\begin{aligned} \boldsymbol{S}^{(t)} &\in \operatorname{argmin} \left(\eta\lambda_2 \|\boldsymbol{S}\|_{2,1} + \tfrac{1}{2} \left\| \boldsymbol{S} - \boldsymbol{S}^{(t-1)} + \eta \boldsymbol{G}_S^{(t-1)} \right\|_F^2 \right), \\ &= \mathsf{Tc}_{\eta\lambda_2} \left(\boldsymbol{S}^{(t-1)} - \eta \boldsymbol{G}_S^{(t-1)} \right), \end{aligned} \tag{24}$$

where $\mathsf{Tc}_{\eta\lambda_2}$ is the column-wise soft-thresholding operator such that for any $\mathbf{M} \in \mathbb{R}^{n \times n}$ and for any $\lambda > 0$, the j-th column of $\mathsf{Tc}_{\lambda}(\mathbf{M})$ is given by $(1 - \lambda / \|\mathbf{M}_{.,j}\|_2)_+ \mathbf{M}_{.,j}$. The step size η is constant and fixed in advance, and satisfies $\eta \leq 1/(2+\epsilon)$. Then, we compute the adaptive upper bound $\bar{R}^{(t)}$ as follows:

$$\bar{R}^{(t)} = \lambda_1^{-1} \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}). \tag{25}$$

Note that, by definition:

$$\begin{aligned}\Phi_\epsilon(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) &\geq \Phi_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \hat{R}) \\ &= \frac{1}{2}\|\Omega \odot (\boldsymbol{A} - \hat{\boldsymbol{L}}_\epsilon - \hat{\boldsymbol{S}}_\epsilon - (\hat{\boldsymbol{S}}_\epsilon)^\top)\|_F^2 + \lambda_1\|\hat{\boldsymbol{L}}_\epsilon\|_* + \lambda_2\|\hat{\boldsymbol{S}}_\epsilon\|_{2,1} \\ &\quad + \frac{\epsilon}{2}(\|\hat{\boldsymbol{L}}_\epsilon\|_F^2 + \|\hat{\boldsymbol{S}}_\epsilon\|_F^2) \\ &\geq \lambda_1\|\hat{\boldsymbol{L}}_\epsilon\|_*,\end{aligned}$$

since every term in the objective function is non-negative. As a result, we obtain that

$$\|\hat{\boldsymbol{L}}_\epsilon\|_* \leq \lambda_1^{-1}\Phi_\epsilon(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}),$$

and we get the upper bound (25). Finally, the low-rank component given by $(\boldsymbol{L}, R)$ is updated using a conjugate gradient step as follows:

$$\left(\boldsymbol{L}^{(t)}, R^{(t)}\right) = \left(\boldsymbol{L}^{(t-1)}, R^{(t-1)}\right) + \beta_t \left(\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}, \tilde{R}^{(t)} - R^{(t-1)}\right), \tag{26}$$

where $\beta_t \in [0, 1]$ is a step size defined later on. Denote by $\boldsymbol{G}_L^{(t-1)} = -\Omega\odot(\boldsymbol{A}-\boldsymbol{L}^{(t-1)}-\boldsymbol{S}^{(t)}-(\boldsymbol{S}^{(t)})^\top)+\epsilon\boldsymbol{L}^{(t-1)}$ the gradient with respect to $\boldsymbol{L}$ of the quadratic part of the objective function, evaluated at $(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})$. The direction $(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)})$ is defined by:

$$\begin{aligned}\left(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)}\right) \in \quad & \operatorname{argmin}_{\mathbf{Z}, R} \quad \langle \mathbf{Z}, \boldsymbol{G}_L^{(t-1)}\rangle + \lambda_1 R \\ & \text{such that} \quad \|\mathbf{Z}\|_* \leq R \leq \bar{R}^{(t)}.\end{aligned} \tag{27}$$

Let σ_1 be the largest singular value of the gradient matrix $\boldsymbol{G}_L^{(t-1)}$, and let u_1 and v_1 be the corresponding left and right singular vectors. Then, (27) admits the following closed-form solution:

$$\left(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)}\right) = \begin{cases} (\mathbf{0}, 0) & \text{if } \lambda_1 \geq \sigma_1 \\ (-\bar{R}^{(t)} u_1 v_1^\top, \bar{R}^{(t)}) & \text{if } \lambda_1 < \sigma_1. \end{cases} \tag{28}$$

The step size β_t is set to:

$$\beta_t = \min\left\{1, \frac{\langle \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}, \boldsymbol{G}_L^{(t-1)}\rangle + \lambda_1(R^{(t-1)} - \tilde{R}^{(t)})}{(1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2}\right\}. \tag{29}$$

We show in appendix A.3 that this choice of step size ensures that the objective function decreases at every iteration. The above steps are repeated iteratively until convergence, or for a predefined number of iterations. In practice, we stop the algorithm when the relative decrease of the objective falls below a predefined threshold (e.g., 10e-6).

A.3 Proof of Theorem 1

To prove Theorem 1, we proceed in three steps. First, we demonstrate that the objective function decreases after every update of $\boldsymbol{S}$ or $\boldsymbol{L}$. In a second step, we compute a lower bound on the amount by which the objective function decreases at each iteration. In a third step, we use this lower bound to demonstrate that the distance to the optimal solution at iteration $t \geq 1$, $\Delta^t = \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\hat{\boldsymbol{S}}, \hat{\boldsymbol{L}}, \hat{R})$, decreases at a rate of the order of $1/t$.

Decrease of the objective between successive iterations: We start by showing that the proximal update for the $\mathbf{S}$ block yields a decrease of the objective. For $t \geq 1$, denote $Q^{(t-1)} = \lambda_2^{-1}\Phi_\epsilon(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})$, and

$$g_S(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}) = \langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \mathbf{S}^{(t-1)} - \tilde{\mathbf{S}}^{(t-1)}\rangle + \lambda_2(\|\mathbf{S}^{(t-1)}\|_{2,1} - \|\tilde{\mathbf{S}}^{(t-1)}\|_{2,1}). \tag{30}$$

In (30), $\boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}) = -2\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \mathbf{S}^{(t-1)} - (\mathbf{S}^{(t-1)})^\top) + \epsilon \mathbf{S}^{(t-1)}$ is the gradient matrix with respect to $\mathbf{S}$ of the quadratic part of the objective function, evaluated at $(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)})$, and

$$\tilde{\mathbf{S}}^{(t-1)} = \arg\min_{\mathbf{S}} \quad \langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \mathbf{S} \rangle + \lambda_2 \|\mathbf{S}\|_{2,1} \quad \text{s.t. } \|\mathbf{S}\|_{2,1} \leq Q^{(t-1)}.$$

Lemma 4. *For $t \geq 1$, the proximal update for the $\mathbf{S}$ block defined in (24) satisfies:*

$$\Phi_\epsilon(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \leq \Phi_\epsilon(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \frac{\eta}{2} \frac{g_S^2(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)})^2}{(2Q^{(t-1)})}.$$

Proof. See Section A.8.4. □

We now prove a similar result, this time concerning the $(\boldsymbol{L}, R)$ block update. Recall that, for $t \geq 1$, $\bar{R}^{(t)} = \lambda_1^{-1} \Phi_\epsilon(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})$.

$$g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) = \langle \boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}), \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t-1)} \rangle + \lambda_1 (R^{(t-1)} - \tilde{R}^{(t-1)}). \tag{31}$$

In (31), $\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) = -\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \mathbf{S}^{(t)} - (\mathbf{S}^{(t)})^\top) + \epsilon \boldsymbol{L}^{(t-1)}$ is the gradient matrix with respect to $\boldsymbol{L}$ of the quadratic part of the objective function, evaluated at $(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)})$. Recall that $M^{(t)} = \|\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})\|_F$. We prove the following result, which ensures a decrease of the objective function after the conditional gradient update.

Lemma 5. *For $t \geq 1$, the conditional gradient update for the $(\boldsymbol{L}, R)$ block defined in (28) satisfies:*

$$\Phi_\epsilon(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) - \Phi_\epsilon(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \leq -\frac{g_L^2(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})}{\max\{2\bar{R}^{(t)}(\lambda_1 + M^{(t)}), 8(1+\epsilon)(\bar{R}^{(t)})^2\}}.$$

Moreover,

$$\Phi_\epsilon(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) - \Phi_\epsilon(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \leq -\frac{(1+\epsilon)}{2} \|\boldsymbol{L}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2. \tag{32}$$

Proof. See Section A.8.5. □

Lower bound on the decrement $\Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\boldsymbol{S}^{(t+1)}, \boldsymbol{L}^{(t)}, R^{(t)})$: Consider the function

$$g^t(Q^{(t)}, \bar{R}^{(t)}) \triangleq g_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) + g_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, \bar{R}^{(t-1)}).$$

In what follows, we compute upper and lower bounds on $g^t(Q^{(t)}, \bar{R}^{(t)})$. Note that $g^t(Q^{(t)}, \bar{R}^{(t)})$ depends on $(Q^{(t)}, \bar{R}^{(t)})$, because computing g_S and g_L involve solving constrained optimization problems, which depend on $Q^{(t)}$ and $\bar{R}^{(t)}$, respectively. By convexity of the quadratic term $\|\Omega \odot (\boldsymbol{A} - \boldsymbol{L} - \boldsymbol{S} - \boldsymbol{S}^\top)\|_F^2/2 + \epsilon/2(\|\boldsymbol{L}\|_F^2 + \|\boldsymbol{S}\|_F^2)$, we obtain that:

$$g^t(Q^{(t)}, \bar{R}^{(t)}) \geq \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\tilde{\boldsymbol{S}}^{(t)}, \tilde{\boldsymbol{L}}^{(t-1)}, \tilde{R}^{(t-1)}).$$

Then, by definition of the minimizer $(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \hat{R})$:

$$g^t(Q^{(t)}, \bar{R}^{(t)}) \geq \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \hat{R}), \tag{33}$$

which gives the lower bound on $g^t(Q^{(t)}, \bar{R}^{(t)})$.

Let us now compute an upper bound for $g^t(Q^{(t)}, \bar{R}^{(t)})$. To do so, we start by upper bounding $g_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})$ defined in (30). By definition,

$$\begin{aligned}
g_S(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}) &= \max_{\|\boldsymbol{S}\|_{2,1} \leq Q^{(t)}} \{\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}), \boldsymbol{S}^{(t)} - \boldsymbol{S}\rangle + \lambda_2(\|\mathbf{S}^{(t)}\|_{2,1} - \|\boldsymbol{S}\|_{2,1})\} \\
&= \max_{\|\boldsymbol{S}\|_{2,1} \leq Q^{(t)}} \{\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}), \boldsymbol{S}^{(t)} - \boldsymbol{S}\rangle \\
&\quad + \langle \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) - \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}), \boldsymbol{S}^{(t)} - \boldsymbol{S}\rangle + \lambda_2(\|\mathbf{S}^{(t)}\|_{2,1} - \|\boldsymbol{S}\|_{2,1})\} \\
&\leq \max_{\|\boldsymbol{S}\|_{2,1} \leq Q^{(t)}} \Big\{\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}), \boldsymbol{S}^{(t)} - \boldsymbol{S}\rangle + \lambda_2(\|\mathbf{S}^{(t)}\|_{2,1} - \|\boldsymbol{S}\|_{2,1}) \\
&\quad + \|\boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) - \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)})\|_F \|\boldsymbol{S}^{(t)} - \boldsymbol{S}\|_F\Big\}. \\
&\leq \underbrace{\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}), \boldsymbol{S}^{(t)}\rangle + \lambda_2\|\mathbf{S}^{(t)}\|_{2,1} - \min_{\|\boldsymbol{S}\|_{2,1} \leq Q^{(t)}} \Big\{\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}), \boldsymbol{S}\rangle + \lambda_2\|\boldsymbol{S}\|_{2,1}\Big\}}_{I} \\
&\quad + \underbrace{\max_{\|\boldsymbol{S}\|_{2,1} \leq Q^{(t)}} \Big\{\|\boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) - \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)})\|_F \|\boldsymbol{S}^{(t)} - \boldsymbol{S}\|_F\Big\}}_{II}
\end{aligned}$$

On the one hand, by definition of $\tilde{\boldsymbol{S}}^{(t)}$ and $g_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)})$ (see (30) and (A.3)), we have:

$$I \leq g_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}). \tag{34}$$

On the other hand, by definition of $Q^{(t)}$, $\|\boldsymbol{S}^{(t)}\|_{2,1} \leq Q^{(t)}$, which implies $\|\boldsymbol{S}^{(t)}\|_F \leq Q^{(t)}$; combined with $\|\boldsymbol{S}\|_F \leq Q^{(t)}$, we obtain that that $\|\boldsymbol{S}^{(t)} - \boldsymbol{S}\|_F \leq 2Q^{(t)}$. Note also that, as the gradient G_S is $(1+\epsilon)$-Lipschitz, we have $\|\boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) - \boldsymbol{G}_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)})\|_F \leq (1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \boldsymbol{L}^{(t)}\|_F$. Finally we obtain:

$$II \leq 2Q^{(t)}(1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \boldsymbol{L}^{(t)}\|_F. \tag{35}$$

Combining (34) and (35), we finally obtain:

$$g_S(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}) \leq g_S(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}) + 2Q^{(t)}(1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \boldsymbol{L}^{(t)}\|_F. \tag{36}$$

We now use (36) to derive our upper bound on $g^t(Q^{(t)}, \bar{R}^{(t)})$ as follows. Using Lemma 4 and Lemma 5, we obtain that:

$$\begin{aligned}
(g^{(t)}(Q^{(t)}, \bar{R}^{(t)}))^2 &\leq 2\Big\{g_L^2(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) + g_S^2(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}) + 4(Q^{(t)})^2(1+\epsilon)^2\|\boldsymbol{L}^{(t-1)} - \boldsymbol{L}^{(t)}\|_F^2\Big\} \\
&\leq 2\Big\{(C_1^{(t)} + C_3^{(t)})(\Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)})) \\
&\quad + C_2^{(t)}(\Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) - \Phi_\epsilon(\boldsymbol{S}^{(t+1)}, \boldsymbol{L}^{(t)}, R^{(t)}))\Big\},
\end{aligned}$$

where

$$C_1^{(t)} = \max\{2\bar{R}^{(t)}(\lambda_1 + M^{(t)}), 8(1+\epsilon)(\bar{R}^{(t)})^2\}, \quad C_2^{(t)} = \frac{8(Q^{(t)})^2}{\eta}, \quad C_3^{(t)} = 8(1+\epsilon)(Q^{(t)})^2.$$

Define:

$$C^{(t)} = 2\max\{C_1^{(t)} + C_3^{(t)}, C_2^{(t)}\}. \tag{37}$$

We finally have the following lower bound:

$$(g^{(t)}(Q^{(t)}, \bar{R}^{(t)}))^2 \leq C^{(t)}(\Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\boldsymbol{S}^{(t+1)}, \boldsymbol{L}^{(t)}, R^{(t)})).$$

Convergence rate of order $1/t$: Recall that $\Delta^t := \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \hat{R})$. Using the fact that

$$(g^{(t)}(Q^{(t)}, \bar{R}^{(t)}))^2 \geq (\Delta^t)^2,$$

proven in (33), we obtain that

$$\Delta^{t+1} \leq \Delta^t - \frac{1}{C^{(t)}}(\Delta^t)^2.$$

We use the following Lemma (see, e.g. [6, Lemma 3.5], [51, Lemma 8]).

Lemma 6. *Let $\{A_k\}_{k\geq 1}$ be a non-negative sequence satisfying:*

$$A_{k+1} \leq A_k - \gamma_k A_k^2, k \geq 1,$$

where $\gamma_k > 0$ for any $k \geq 1$. Then,

$$A_{k+1} \leq \frac{1}{\frac{1}{A_1} + \sum_{i=1}^k \gamma_i}.$$

Proof. See Section A.9 □

Lemma 6 yields that:

$$\Delta^{t+1} \leq \frac{1}{(\Delta^1)^{-1} + \sum_{i=1}^t \frac{1}{C^{(i)}}}.$$

noting that $\Delta^1 \leq \tilde{\Delta}^0 := \Phi_\epsilon(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)}) - \Phi_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}, \hat{R})$, we have:

$$\Delta^{t+1} \leq \frac{1}{(\tilde{\Delta}^0)^{-1} + \sum_{i=1}^t \frac{1}{C^{(i)}}}. \tag{38}$$

Let us derive an upper bound on the time-varying constants $C^{(t)}$ defined in (37). We only need to bound $\bar{R}^{(t)}$, $M^{(t)}$ and $Q^{(t)}$. First note that, by Lemmas 4 and 5, $\bar{R}^{(t)} \leq \lambda_1^{-1}\Phi_\epsilon(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)})$, and $Q^{(t)} \leq \lambda_2^{-1}\Phi_\epsilon(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)})$. To bound $M^{(t)} = \|\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})\|_F$, we start by noticing that the gradient $\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{t-1})$ of the quadratic part of the objective with respect to $\boldsymbol{L}$ is bounded whenever $\boldsymbol{S}^{(t)}$ and $\boldsymbol{L}^{(t-1)}$ are bounded themselves. Since $\lambda_1\|\boldsymbol{L}^{(t-1)}\|_* + \lambda_2\|\boldsymbol{S}^{(t)}\|_{2,1} \leq \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \leq \Phi_\epsilon(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)})$, the parameters $\boldsymbol{S}$ and $\boldsymbol{L}$ are indeed bounded, and we obtain that there exists $\bar{M} \geq 0$ such that $M^{(t)} \leq \bar{M}$ for any t. Define $\mathcal{F}_0 \triangleq \Phi_\epsilon(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)})$,

$$\bar{C}_1 = \max\{8\lambda_1^{-1}(1+\epsilon)\mathcal{F}_0^2, 2\lambda_1^{-1}\mathcal{F}_0(\lambda_1 + \bar{M})\}, \quad \bar{C}_2 = \frac{8\mathcal{F}_0^2}{\eta\lambda_2^2}, \quad \bar{C}_3 = 8\lambda_2^{-1}(1+\epsilon)\mathcal{F}_0^2,$$

and

$$\bar{C} \triangleq \max\left\{\bar{C}_1 + \bar{C}_3, \bar{C}_2\right\}.$$

Then, we obtain the following rate of convergence:

$$\Delta^{t+1} \leq \frac{1}{(\tilde{\Delta}^0)^{-1} + \sum_{i=1}^t \frac{1}{C^{(i)}}} \leq \frac{1}{(\tilde{\Delta}^0)^{-1} + t\bar{C}}. \tag{39}$$

Recall that $\Phi_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \hat{R}) = \mathcal{F}(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon)$ by equivalence of the two optimization problems (6) and (7). In addition, by definition, $\|L^{(t-1)}\|_* \leq R^{(t-1)}$, which gives $\mathcal{F}_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) \leq \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})$. Thus, we obtain that $\mathcal{F}_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}) - \mathcal{F}_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon) \leq \Phi_\epsilon(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \Phi_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon, \hat{R}) \leq \Delta^{t+1}$.
For $\delta > 0$, let T_δ be the integer number defined by:

$$T_\delta \triangleq \left\lfloor \bar{C}\left(\frac{1}{\delta} - \frac{1}{\mathcal{F}_0 - \mathcal{F}_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon)}\right)\right\rfloor + 1.$$

Then, the T_δ-th iterate of the MCGD sequence satisfies:

$$\mathcal{F}_\epsilon(\boldsymbol{S}^{(T_\delta)}, \boldsymbol{L}^{(T_\delta)}) - \mathcal{F}_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon) \leq \delta,$$

which proves sub-linear convergence of the MCGD iterates. Note that, by definition, $\mathcal{F}_0 - \mathcal{F}_\epsilon(\hat{\boldsymbol{S}}_\epsilon, \hat{\boldsymbol{L}}_\epsilon) \geq 0$, which implies that $T_\delta \leq \left\lfloor \bar{C}/\delta \right\rfloor + 1$. In addition, in the particular case where the initial point is set to $(\boldsymbol{S}^{(0)}, \boldsymbol{L}^{(0)}, R^{(0)}) = (\boldsymbol{0}, \boldsymbol{0}, 0)$, we can compute an upper bound on the constant $\bar{C}$, dependent on the dimensions of the problem. First, note that in this case, $\mathcal{F}_0 = \frac{1}{2}\|\Omega \odot \boldsymbol{A}\|_F^2$ is equal to the number of observed edges in the graph, denoted by E. Furthermore, by definition,

$$M^{(t)} = \|\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})\|_F \leq \|\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \boldsymbol{S}^{(t)} - (\boldsymbol{S}^{(t)})^\top)\|_F + \|\epsilon \boldsymbol{L}^{(t-1)}\|_F.$$

Since, by Lemmas 4 and 5, the objective value decreases at every update of $\boldsymbol{L}$ and $\boldsymbol{S}$. As all the terms of the objective are positive, we have that $\|\Omega \odot (\boldsymbol{A} - \boldsymbol{L}^{(t-1)} - \boldsymbol{S}^{(t)} - (\boldsymbol{S}^{(t)})^\top)\|_F^2 \leq \mathcal{F}_0 = E$, and $\|\epsilon \boldsymbol{L}^{(t-1)}\|_F^2 \leq E$ as well. Thus, we obtain that, for any t, $M^{(t)} \leq 2\sqrt{E}$, which yields $\bar{M} \leq 2\sqrt{E}$. We then obtain that the constant $\bar{C}$ satisfies

$$\bar{C} \leq \bar{C}_0 \triangleq \max\left\{ \frac{2E^2}{\eta \lambda_2^2}, 8(1+\epsilon)E^2 \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} \right) + \frac{2E^{3/2}}{\lambda_1} + 2E \right\}, \tag{40}$$

meaning that the number of iterations increases at most quadratically with the density of the graph. Note that, in practice, the convergence is much faster, and we observe that the algorithm converges after a few iterations.

A.4 Proof of Theorem 2

Recall that, by Lemma 1,

$$j \in \widehat{\mathcal{O}} \;\Leftrightarrow\; \left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 > \frac{\lambda_2}{4}.$$

In a first time, we show that with high probability, no inlier belongs to the set of estimated outliers. Consider $j \in \mathcal{I}$, then

$$\begin{aligned}
\left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 &\leq \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{L}}_{ij} - \widehat{\boldsymbol{S}}_{ji} \right)_+ \right)^2} + \sqrt{\sum_{i \in \mathcal{O}} \left(\boldsymbol{\Omega}_{ij} \left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{L}}_{ij} - \widehat{\boldsymbol{S}}_{ji} \right)_+ \right)^2} \\
&\leq \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \left(\boldsymbol{\Sigma}_{ij} + \Delta \boldsymbol{L}_{ij} - \widehat{\boldsymbol{S}}_{ji} \right)_+ \right)^2} + \sqrt{\sum_{i \in \mathcal{O}} \left(\boldsymbol{\Omega}_{ij} \boldsymbol{A}_{ij} \right)^2}
\end{aligned}$$

where we have used that for $(i,j) \in \mathcal{I} \times \mathcal{I}$, $\boldsymbol{A}_{ij} = \boldsymbol{\Sigma}_{ij} + \boldsymbol{L}^*_{ij}$ and that $\widehat{\boldsymbol{L}}_{ij} \geq 0$ and $\widehat{\boldsymbol{S}}_{ij} \geq 0$. Therefore, we find that

$$\left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 \leq \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \boldsymbol{\Sigma}_{ij} \right)_+^2} + \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \Delta \boldsymbol{L}_{ij} \right)_+^2} + \sqrt{\sum_{i \in \mathcal{O}} \left(\boldsymbol{\Omega}_{ij} \boldsymbol{A}_{ij} \right)^2}.$$

Recalling that $\|\Delta \boldsymbol{L}\|_\infty \leq \rho_n$, we obtain

$$\max_{j \in \mathcal{I}} \left\{ \left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 \right\} \leq \left\| \boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I} \right\|_{2,\infty} + \rho_n \left\| \boldsymbol{\Omega}_{|I} \right\|_{2,\infty} + \left\| \boldsymbol{\Omega} \odot \boldsymbol{A}_{|\mathcal{O} \times \mathcal{I}} \right\|_{2,\infty}. \tag{41}$$

We bound $\left\| \boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I} \right\|_{2,\infty}$, $\rho_n \left\| \boldsymbol{\Omega}_{|I} \right\|_{2,\infty}$ and $\left\| \boldsymbol{\Omega} \odot \boldsymbol{A}_{|\mathcal{O} \times \mathcal{I}} \right\|_{2,\infty}$ using the following Lemma.

Lemma 7. *Under assumptions 1-3,*

$$\mathbb{P}\left(\left\| \boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I} \right\|_{2,\infty} \geq \sqrt{6 \nu_n \rho_n n} \right) \leq 2e^{-\nu_n \rho_n n} \tag{42}$$

$$\mathbb{P}\left(\left\| \boldsymbol{\Omega}_{|I} \right\|_{2,\infty} \geq 4\sqrt{\nu_n n} \right) \leq 2e^{-\nu_n n} \tag{43}$$

$$\mathbb{P}\left(\left\| \boldsymbol{\Omega} \odot \boldsymbol{A}_{|\mathcal{O} \times \mathcal{I}} \right\|_{2,\infty} \geq \sqrt{6 \nu_n \rho_n n} \right) \leq 2e^{-\nu_n \rho_n n}. \tag{44}$$

Proof. See Section A.8.6 □

Recall that $\lambda_2 = 19\sqrt{\nu_n \rho_n n}$. Combining Lemma 7, Lemma 3 and equation (41) yields that with probability larger than $1 - 6e^{-\nu_n \rho_n n}$,

$$\max_{j \in \mathcal{I}} \left\{ \left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 \right\} \quad \leq \quad 9\sqrt{\nu_n \rho_n n} < \frac{\lambda_2}{2}.$$

Using Lemma 1, we conclude that with probability at least $1 - 6e^{-\nu_n \rho_n n}$, $\widehat{\mathcal{O}} \cap \mathcal{I} = \emptyset$.

A.5 Proof of Theorem 3

Here, we prove that with high probability, all outliers are detected when $\min_{j \in \mathcal{O}} \sum_{i \in \mathcal{I}} \boldsymbol{\Pi}_{ij} \boldsymbol{S}^*_{ij} > C \rho_n \nu_n n$ for some absolute constant $C > 0$. For any $j \in [n]$, note that

$$\left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 \quad \geq \quad \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{L}}_{ij} - \widehat{\boldsymbol{S}}_{ji} \right)_+ \right)^2}.$$

We have shown in Theorem 2 that with probability at least $1 - 6e^{-\nu_n \rho_n n}$, $\widehat{\boldsymbol{S}}_{ji} = 0$ for any $i \in \mathcal{I}$ and any $j \in [n]$. When this equation holds, using the bound $\left\| \widehat{\boldsymbol{L}} \right\|_\infty \leq \rho_n$, we find that

$$\left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 \quad \geq \quad \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \left(\boldsymbol{A}_{ij} - \rho_n \right)_+ \right)^2}. \tag{45}$$

We use the following Lemma to obtain a lower bound on the right hand side of equation (45) when $j \in \mathcal{O}$.

Lemma 8. *Assume that* $\min_{j \in \mathcal{O}} \sum_{i \in \mathcal{I}} \boldsymbol{\Pi}_{ij} \boldsymbol{S}^*_{ij} \geq \nu_n \rho_n n$, *then*

$$\mathbb{P}\left(\min_{j \in \mathcal{O}} \sqrt{\sum_{i \in \mathcal{I}} \left(\boldsymbol{\Omega}_{ij} \left(\boldsymbol{A}_{ij} - \rho_n \right)_+ \right)^2} \leq \frac{1}{4} \min_{j \in \mathcal{O}} \sqrt{\sum_{i \in \mathcal{I}} \boldsymbol{\Pi}_{ij} \boldsymbol{S}^*_{ij}} \right) \leq 2se^{-\frac{-\nu_n \rho_n n}{80}}.$$

Proof. See Section A.8.7. □

Combining this Lemma with equation (45), we see that with probability at least $1 - 2se^{-\frac{-\nu_n \rho_n n}{80}} - 6e^{-\nu_n \rho_n n}$,

$$\left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 \quad \geq \quad \frac{1}{4} \min_{j \in \mathcal{O}} \sqrt{\sum_{i \in \mathcal{I}} \boldsymbol{\Pi}_{ij} \boldsymbol{S}^*_{ij}}. \tag{46}$$

Recall that $\lambda_2 = 19\sqrt{\nu_n \rho_n n}$. When $\min_{j \in \mathcal{O}} \sum_{i \in \mathcal{I}} \boldsymbol{\Pi}_{ij} \boldsymbol{S}^*_{ij} > 8 \times 19 \nu_n \rho_n n$, Lemma 8 and equation (46) imply that with probability larger than $1 - 2se^{-\frac{-\nu_n \rho_n n}{80}} - 6e^{-\nu_n \rho_n n}$,

$$\left\| \boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot} \right)_+ \right\|_2 > \frac{\lambda_2}{2}.$$

Combining this result with Lemma 1, we find that with probability at least $1 - 2se^{-\frac{-\nu_n \rho_n n}{80}} - 6e^{-\nu_n \gamma_n n} \geq 1 - 8se^{-\frac{-\nu_n \rho_n n}{80}}$, $\mathcal{O} \subset \widehat{\mathcal{O}}$. This concludes the proof of Theorem 3.

A.6 Proof of Theorem 4

To prove Theorem 4, we use the definition of $\widehat{\boldsymbol{L}}$, the separability of the $\|\cdot\|_*$-norm on orthogonal subspaces, and results on $\widehat{\boldsymbol{S}}$ proved in Theorem 3. Recall that $\Psi \triangleq 16\tilde{\nu}_n\gamma_n\rho_n sn$.

Lemma 9. *Assume that* $\lambda_1 \geq 3\left\|\boldsymbol{\Omega}\odot\boldsymbol{\Sigma}_{|I}\right\|_{op}$, *and that* $\lambda_2 = 19\sqrt{\nu_n\rho_n n}$. *Then,*

$$\|\boldsymbol{\Omega}\odot\Delta\boldsymbol{L}\|_F^2 \leq \frac{\lambda_1}{3}\left(5\left\|\mathcal{P}_{\boldsymbol{L}^*}(\Delta\boldsymbol{L})\right\|_* - \left\|\mathcal{P}^{\perp}_{\boldsymbol{L}^*}(\Delta\boldsymbol{L})\right\|_*\right) + \Psi \tag{47}$$

$$\text{and} \quad \|\Delta\boldsymbol{L}\|_* \leq 6\sqrt{k}\left\|\Delta\boldsymbol{L}_{|I}\right\|_F + 6\sqrt{3ksn}\rho_n + \frac{3\Psi}{\lambda_1}. \tag{48}$$

hold simultaneously with equation (17) with probability at least $1 - 6e^{-\nu_n\rho_n n} - 2e^{-\tilde{\nu}_n\gamma_n sn}$.

Proof. See Section A.8.8. □

Bounding the $\|\cdot\|_{L_2(\boldsymbol{\Pi})}$-norm of the error $\Delta\boldsymbol{L}$ by $\|\boldsymbol{\Omega}\odot\Delta\boldsymbol{L}\|_F^2$ is rather involved, and we use a peeling argument, combined with the bound on $\|\Delta\boldsymbol{L}\|_*$ obtained in equation (48) in Lemma 9. We recall that $\boldsymbol{\Gamma}$ is the random matrix defined as $\boldsymbol{\Gamma}_{ij} = \epsilon_{ij}\boldsymbol{\Omega}_{ij}$ for all $(i,j)\in[n]\times[n]$, where $\{\epsilon\}_{i\leq j}$ is a Rademacher sequence. Moreover, we introduce the following notation :

$$\beta \triangleq \mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right]\left(\frac{48^2\rho_n^2 k}{\mu_n}\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] + 60\rho_n^2\sqrt{ksn} + \frac{32\Psi\rho_n}{\lambda_1}\right). \tag{49}$$

Lemma 10. *Assume that* $\lambda_1 \geq 3\left\|\boldsymbol{\Omega}\odot\boldsymbol{\Sigma}_{|I}\right\|_{op}$, *and that* $\lambda_2 = 19\sqrt{\nu_n\rho_n n}$. *Then, there exists an absolute constant* $C>0$ *such that*

$$\left\|\Delta\boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 \leq C\left(\frac{\lambda_1^2 k}{\mu_n} + \nu_n\rho_n^2 sn + \frac{\nu_n\rho_n^2 kn}{\mu_n} + \Psi + \beta\right) \tag{50}$$

holds simultaneously with equations (17), (47) and (48) with probability at least $1 - 7e^{-\nu_n\rho_n n} - 2e^{-\tilde{\nu}_n\gamma_n sn}$.

Proof. See Section A.8.9. □

Finally, we bound β using the following lemma.

Lemma 11. $\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] \leq 84\sqrt{\nu_n n}$.

Lemma 11 implies that there exists some absolute constant $C>0$ such that

$$\beta \leq C\sqrt{\nu_n n}\left(\frac{\rho_n^2 k}{\mu_n}\sqrt{\nu_n n} + \rho_n^2\sqrt{skn} + \frac{\Psi\rho_n}{\lambda_1}\right).$$

Proof. See Section A.8.10. □

Thus, there exists an absolute constant $C>0$ such that when equation (50) holds,

$$\beta \leq C\left(\frac{\nu_n\rho_n^2 kn}{\mu_n} + \rho_n^2 n\sqrt{\nu_n sk} + \frac{\Psi\sqrt{\nu_n n}\rho_n}{\lambda_1}\right).$$

Combining Lemma 4 and Lemma 9-10, and noticing that $\sqrt{\nu_n sk} \leq \nu_n s + k$ and that $\frac{\nu_n}{\mu_n} \geq 1$, we find that there exists an absolute constant $C>0$ such that with probability at least $1 - 7e^{-\nu_n\rho_n n} - 2e^{-\tilde{\nu}_n\gamma_n sn}$,

$$\begin{aligned}\left\|\Delta\boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 &\leq C\left(\frac{\lambda_1^2 k}{\mu_n} + \nu_n\rho_n^2 sn + \frac{\nu_n\rho_n^2 kn}{\mu_n} + \Psi + \frac{\nu_n\rho_n^2 kn}{\mu_n} + \rho_n^2 n\sqrt{\nu_n sk} + \frac{\Psi\sqrt{\nu_n n}\rho_n}{\lambda_1}\right)\\ &\leq C\left(\frac{\lambda_1^2 k}{\mu_n} + n\rho_n^2\left(\nu_n s + \frac{\nu_n k}{\mu_n}\right) + \Psi\left(\frac{\sqrt{\nu_n n}\rho_n}{\lambda_1} + 1\right)\right).\end{aligned}$$

Recall that $\Phi \triangleq n\rho_n^2\left(\frac{\nu_n k}{\mu_n} + \nu_n s\right)$, and that $\Xi \triangleq \frac{\sqrt{\nu_n n}\rho_n}{\lambda_1} + 1$. With these notations, we find that

$$\left\|\Delta\boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 \leq C\left(\frac{\lambda_1^2 k}{\mu_n} + \Phi + \Psi\Xi\right)$$

with probability at least $1 - 7e^{-\nu_n\rho_n n} - 2e^{-\tilde{\nu}_n\gamma_n sn}$. We conclude the proof of Theorem 4 by recalling that $\nu_n\rho_n n \geq \log(n)$ and $\tilde{\nu}_n\gamma_n n \geq \log(n)$.

A.7 Proof of Corollary 1

Lemma 3 allows us to choose λ_1 by bounding the noise terms $\left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{op}$ with high probability. For the choice $\lambda_1 = 84\sqrt{\nu_n \rho_n n}$, we find that

$$\Xi \quad = \quad \left(1 + \frac{\sqrt{\nu_n \rho_n^2 n}}{84\sqrt{\nu_n \rho_n n}}\right) \leq 2.$$

Combining Lemma 3 with Theorem 4, we find that there exists an absolute constant $C > 0$ such that with probability at least $1 - 7e^{-\nu_n \rho_n n} - 3e^{-\tilde{\nu}_n \gamma_n n s}$,

$$\begin{aligned}\left\|\Delta \boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 &\leq C\left(\frac{\nu_n \rho_n k n}{\mu_n} + n\rho_n^2\left(\frac{\nu_n k}{\mu_n} + \nu_n s\right) + \tilde{\nu}_n \rho_n \gamma_n s n\right) \\ &\leq C\left(\frac{\nu_n \rho_n k n}{\mu_n} + \rho_n(\nu_n \rho_n \vee \tilde{\nu}_n \gamma_n) s n\right).\end{aligned}$$

A.8 Proof of auxiliary Lemmas

A.8.1 Proof of Lemma 1

Recall that by definition of $\widehat{\boldsymbol{S}}$,

$$\widehat{\boldsymbol{S}} \in \underset{\boldsymbol{S} \in \mathbb{R}_+^{n \times n}}{\arg\min} \left\{ \frac{1}{2} \left\|\boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{L}} - \boldsymbol{S} - \boldsymbol{S}^\top\right)\right\|_F^2 + \lambda_2 \|\boldsymbol{S}\|_{2,1} \right\} \tag{51}$$

Now, any subgradient of the objective function (51) at $\widehat{\boldsymbol{S}}$ is of the form

$$\nabla_{\boldsymbol{S}} \mathcal{F}\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{L}}\right) = 2\boldsymbol{\Omega} \odot \left(-\boldsymbol{A} + \widehat{\boldsymbol{L}} + \widehat{\boldsymbol{S}} + \widehat{\boldsymbol{S}}^\top\right) + \lambda_2 \boldsymbol{W}$$

where $\boldsymbol{W}$ is a subgradient of the $\|\cdot\|_{2,1}$-norm at $\widehat{\boldsymbol{S}}$. The matrix $\boldsymbol{W}$ obeys the following constraints :

- for any $j \in [n]$ such that the column $\widehat{\boldsymbol{S}}_{\cdot,j}$ is null, $\|\boldsymbol{W}_{\cdot,j}\|_2 \leq 1$;
- for any $j \in [n]$ such that $\widehat{\boldsymbol{S}}_{\cdot,j} \neq \boldsymbol{0}$, $\|\boldsymbol{W}_{\cdot,j}\|_2 = \frac{\widehat{\boldsymbol{S}}_{\cdot,j}}{\|\widehat{\boldsymbol{S}}_{\cdot,j}\|_2}$.

The Karush-Kuhn-Tucker conditions (see, e.g., [9], Section 5.5.3) imply that there exists $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{W} \in \partial \|\cdot\|_{2,1}$ such that

$$2\boldsymbol{\Omega} \odot \left(-\boldsymbol{A} + \widehat{\boldsymbol{L}} + \widehat{\boldsymbol{S}} + \widehat{\boldsymbol{S}}^\top\right) + \lambda_2 \boldsymbol{W} - \boldsymbol{H} = \boldsymbol{0} \tag{52}$$

$$\boldsymbol{H}_{ij} \geq 0 \text{ for any } (i,j) \in [n] \times [n] \tag{53}$$

$$\boldsymbol{H} \odot \widehat{\boldsymbol{S}} = \boldsymbol{0} \tag{54}$$

First, we prove the implication $\widehat{\boldsymbol{S}}_{\cdot,j} = \boldsymbol{0} \Rightarrow \left\|\boldsymbol{\Omega} \odot \left(\boldsymbol{A}_{j,\cdot} - \widehat{\boldsymbol{L}}_{j,\cdot} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)_+\right\|_2 \leq \frac{\lambda_2}{2}$. To do so, assume that j is such that $\widehat{\boldsymbol{S}}_{\cdot,j} = \boldsymbol{0}$. Then, equation (52) implies that

$$\lambda_2 \boldsymbol{W}_{\cdot,j} = 2\boldsymbol{\Omega} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right) + \boldsymbol{H}_{\cdot,j}.$$

Recall that $\|\boldsymbol{W}_{\cdot,j}\|_2 \leq 1$, and thus

$$\frac{2}{\lambda_2}\left\|\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right) + \frac{1}{2}\boldsymbol{H}_{\cdot,j}\right\|_2 \leq 1.$$

Moreover, by (53), $\boldsymbol{H}_{ij} \geq 0$. Therefore,

$$\begin{aligned}\frac{2}{\lambda_2}\left\|\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)_+\right\|_2 &\leq \frac{2}{\lambda_2}\left\|\left(\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right) + \frac{1}{2}\boldsymbol{H}_{\cdot,j}\right)_+\right\|_2 \\ &\leq \frac{2}{\lambda_2}\left\|\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right) + \frac{1}{2}\boldsymbol{H}_{\cdot,j}\right\|_2 \leq 1.\end{aligned}$$

This concludes the proof of the first implication.

To prove the other implication, assume that j is such that $\widehat{\boldsymbol{S}}_{\cdot,j} \neq \boldsymbol{0}$. Then $\boldsymbol{W}_{\cdot,j} = \frac{\widehat{\boldsymbol{S}}_{\cdot,j}}{\|\widehat{\boldsymbol{S}}_{\cdot,j}\|_2}$, and equation (52) becomes

$$\left(2 + \frac{\lambda_2}{\left\|\widehat{\boldsymbol{S}}_{\cdot,j}\right\|_2}\right)\widehat{\boldsymbol{S}}_{\cdot,j} = 2\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right) + \boldsymbol{H}_{\cdot,j} + 2\left(1 - \boldsymbol{\Omega}_{\cdot,j}\right) \odot \widehat{\boldsymbol{S}}_{\cdot,j}.$$

First, assume that for some $i \in [n]$, $\boldsymbol{H}_{ij} \neq 0$. Then, equation (54) implies that $\widehat{\boldsymbol{S}}_{ij} = 0$, and so

$$\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{L}}_{ij} - \widehat{\boldsymbol{S}}_{ji}\right) = -\boldsymbol{H}_{ij}/2 < 0.$$

On the other hand, assume that for $i \in [n]$, $\boldsymbol{H}_{ij} = 0$. Then, $\widehat{\boldsymbol{S}}_{ij} \geq 0$ implies that

$$\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{L}}_{ij} - \widehat{\boldsymbol{S}}_{ji}\right) + \left(1 - \boldsymbol{\Omega}_{ij}\right)\widehat{\boldsymbol{S}}_{ij} \geq 0$$

which implies that $\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{L}}_{ij} - \widehat{\boldsymbol{S}}_{ji}\right) \geq 0$. This shows that for $j \in [n]$ such that $\widehat{\boldsymbol{S}}_{\cdot,j} \neq \boldsymbol{0}$,

$$\left(2 + \frac{\lambda_2}{\left\|\widehat{\boldsymbol{S}}_{\cdot,j}\right\|_2}\right)\widehat{\boldsymbol{S}}_{\cdot,j} = 2\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)_+ + 2\left(1 - \boldsymbol{\Omega}_{\cdot,j}\right) \odot \widehat{\boldsymbol{S}}_{\cdot,j}. \tag{55}$$

Now, for all i such that $\boldsymbol{\Omega}_{ij} = 0$, equation (55) becomes $\left(2 + \frac{\lambda_2}{\|\widehat{\boldsymbol{S}}_{\cdot,j}\|_2}\right)\widehat{\boldsymbol{S}}_{ij} = 2\widehat{\boldsymbol{S}}_{ij}$, and thus $\widehat{\boldsymbol{S}}_{ij} = 0$. This remarks, combined with equation (55), implies that

$$\left(2 + \frac{\lambda_2}{\left\|\widehat{\boldsymbol{S}}_{\cdot,j}\right\|_2}\right)\widehat{\boldsymbol{S}}_{\cdot,j} = 2\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)_+.$$

This implies in particular that

$$2\left\|\left(\boldsymbol{\Omega}_{\cdot,j} \odot \left(\boldsymbol{A}_{\cdot,j} - \widehat{\boldsymbol{L}}_{\cdot,j} - \widehat{\boldsymbol{S}}_{j,\cdot}\right)\right)_+\right\|_2 = 2\left\|\widehat{\boldsymbol{S}}_{\cdot,j}\right\|_2 + \lambda_2 > \lambda_2.$$

This concludes the proof of Lemma 1.

A.8.2 Proof of Lemma 2

Note that for any partition of the nodes into inliers $\mathcal{I}$ and outliers $\mathcal{O}$, the solution $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ to equation (3) such that $\mathcal{O}$ is the support of the columns of $\boldsymbol{S}^*$ is unique up to diagonal terms (if it exists). Indeed, we then have $\boldsymbol{L}^* = \mathbb{E}[\boldsymbol{A}]_{|\mathcal{I}\times\mathcal{I}}$ and $\boldsymbol{S}^* = \mathbb{E}[\boldsymbol{A}]_{|\mathcal{I}\times\mathcal{O}} + 1/2\mathbb{E}[\boldsymbol{A}]_{|\mathcal{O}\times\mathcal{O}}$. Thus, it is enough to prove that the partition into inliers and outliers is unique to prove Lemma 2.

We prove Lemma 2 by contradiction. Let us assume that there exists two different sets $\mathcal{O}$ and $\tilde{\mathcal{O}}$ such that there exists two solutions $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ and $(\tilde{\boldsymbol{L}}, \tilde{\boldsymbol{S}})$ to equation (3), where $\mathcal{O}$ is the support of the columns of $\boldsymbol{S}^*$, and $\tilde{\mathcal{O}}$ that of $\tilde{\boldsymbol{S}}$, and such that $\nu_n n \geq (\max_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} \boldsymbol{\Pi}_{ij}) \vee (\max_{i\in\tilde{\mathcal{I}}} \sum_{j\in\tilde{\mathcal{I}}} \boldsymbol{\Pi}_{ij})$ and $\tilde{\nu_n} s \geq$

$(\max_{i\in\mathcal{I}} \sum_{j\in\mathcal{O}} \mathbf{\Pi}_{ij}) \vee (\max_{i\in\tilde{\mathcal{I}}} \sum_{j\in\tilde{\mathcal{O}}} \mathbf{\Pi}_{ij})$. Here, we have defined $\mathcal{O} = \{j : \boldsymbol{S}^*_{\cdot j} \neq \mathbf{0}\}$, $\tilde{\mathcal{O}} = \{j : \tilde{\boldsymbol{S}}_{\cdot j} \neq \mathbf{0}\}$, $\mathcal{I} = \{j : \boldsymbol{L}^*_{\cdot j} \neq \mathbf{0}\}$, and $\tilde{\mathcal{I}} = \{j : \tilde{\boldsymbol{L}}_{\cdot j} \neq \mathbf{0}\}$. Note that (3) implies that $|\mathcal{O}| = |\tilde{\mathcal{O}}|$, and thus there exists $j \in \mathcal{O} \cap \tilde{\mathcal{I}}$.

We obtain a contradiction by proving that the expected observed degree of j is too large for j to be an inlier. By definition of $(\boldsymbol{L}^*, \boldsymbol{S}^*)$, one has $\boldsymbol{S}^* = \mathbb{E}[\boldsymbol{A}]_{|\mathcal{I}\times\mathcal{O}} + 1/2\mathbb{E}[\boldsymbol{A}]_{|\mathcal{O}\times\mathcal{O}}$. Since $j \in \mathcal{O}$, this yields $\sum_{i\in\mathcal{I}} \mathbf{\Pi}_{ij} \boldsymbol{S}^*_{ij} = \sum_{i\in\mathcal{I}} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij}$. Under assumption 4, we find that $\sum_{i\in\mathcal{I}} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij} \geq C\rho_n\nu_n n$, where $C = 8 \times 19$. In particular, this implies that $\sum_{i\in[n]} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij} \geq 152\rho_n\nu_n n$.

Now, since $j \in \tilde{\mathcal{I}}$, for all $i \in \tilde{\mathcal{I}}$, we have $\mathbb{E}[\boldsymbol{A}]_{ij} \leq \rho_n$ and $\sum_{i\in\tilde{\mathcal{I}}} \mathbf{\Pi}_{ij} \leq \nu_n n$. Thus, $\sum_{i\in\tilde{\mathcal{I}}} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij} \leq \rho_n\nu_n n$. Similarly, $\sum_{i\in\tilde{\mathcal{O}}} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij} \leq \gamma_n\tilde{\nu}_n s$. This implies that $\sum_{i\in[n]} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij} \leq \rho_n\nu_n n + \gamma_n\tilde{\nu}_n s$. Using assumption 3, we find that $\sum_{i\in[n]} \mathbf{\Pi}_{ij} \mathbb{E}[\boldsymbol{A}]_{ij} \leq 2\rho_n\nu_n n$, and obtain a contradiction.

A.8.3 Proof of Lemma 3

Note that $\mathbf{\Omega} \odot \mathbf{\Sigma}_{|I}$ is a symmetric random matrix with independent centered entries. Moreover, for $(i,j) \in \mathcal{I}\times\mathcal{I}$, $(\mathbf{\Omega} \odot \mathbf{\Sigma})_{ij} = b_{ij}\xi_{ij}$, where we define $b_{ij} \triangleq \mathbf{\Pi}_{ij}\boldsymbol{L}^*_{ij}\left(1 - \boldsymbol{L}^*_{ij}\right)$ and $\xi_{ij} = \frac{\mathbf{\Omega}_{ij}\mathbf{\Sigma}_{ij}}{b_{ij}}$. With these notations, we see that $\max_{ij} \mathbb{E}\left[(\xi_{ij}b_{ij})^{2\alpha}\right]^{\frac{1}{2\alpha}} \leq 1$ and that $\max_i \sqrt{\sum_j b_{ij}^2} \leq \nu_n\rho_n n$. Applying Proposition 2 for $t = \sqrt{2\nu_n\rho_n n}$ and $\alpha = 3$, we find that

$$\mathbb{P}\left(\left\|(\mathbf{\Omega} \odot \mathbf{\Sigma})_{|I}\right\|_{op} \geq \sqrt{2}e^{\frac{2}{3}}\left(2\sqrt{\nu_n\rho_n n} + 42\sqrt{\log(n)}\right) + \sqrt{2\nu_n\rho_n n}\right) \leq e^{-\nu_n\rho_n n}.$$

We conclude the proof of Lemma 3 by recalling that $\log(n) \leq \nu_n\rho_n n$.

A.8.4 Proof of Lemma 4

First, using the 2-smoothness of the least-squares data fitting term and the ϵ-smoothness of the ridge regularization, we obtain that:

$$\begin{aligned}\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \quad \leq \quad & \mathcal{F}(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) + \langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \mathbf{S}^{(t)} - \mathbf{S}^{(t-1)}\rangle \\ & + \frac{2+\epsilon}{2}\|\mathbf{S}^{(t)} - \mathbf{S}^{(t-1)}\|_F^2 + \lambda_2(\|\mathbf{S}^{(t)}\|_{2,1} - \|\mathbf{S}^{(t-1)}\|_{2,1}). \end{aligned} \tag{56}$$

Then, by definition of the proximal operator, we have that:

$$\begin{aligned}\mathbf{S}^{(t)} \quad & \in \arg\min \left(\eta\lambda_2\|\mathbf{S}\|_{2,1} + \tfrac{1}{2}\|\mathbf{S} - \mathbf{S}^{(t-1)} - \eta\boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)})\|_F^2\right) \\ & \in \arg\min \Big(\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \mathbf{S} - \mathbf{S}^{(t-1)}\rangle + \tfrac{1}{2\eta}\|\mathbf{S} - \mathbf{S}^{(t-1)}\|_F^2 \\ & \qquad + \lambda_2(\|\mathbf{S}\|_{2,1} - \|\mathbf{S}^{(t-1)}\|_{2,1})\Big). \end{aligned} \tag{57}$$

Combining (56), (57) and the fact that $\eta \leq 1/(2+\epsilon)$, we obtain that, for any $\mathbf{S} \in \mathbb{R}^{n\times n}$:

$$\begin{aligned}\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \quad \leq \quad & \mathcal{F}(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) + \langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \mathbf{S} - \mathbf{S}^{(t-1)}\rangle \\ & + \frac{1}{2\eta}\|\mathbf{S} - \mathbf{S}^{(t-1)}\|_F^2 + \lambda_2(\|\mathbf{S}\|_{2,1} - \|\mathbf{S}^{(t-1)}\|_{2,1}). \end{aligned}$$

In particular, for matrices of the form $b\tilde{\boldsymbol{S}}^{(t-1)} + (1-b)\boldsymbol{S}^{(t-1)}$, $b \in \mathbb{R}$, we obtain:

$$\begin{aligned}\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \quad \leq \quad & \mathcal{F}(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) + b\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \tilde{\mathbf{S}}^{(t-1)} - \mathbf{S}^{(t-1)}\rangle \\ & + \frac{b^2}{2\eta}\|\tilde{\mathbf{S}}^{(t-1)} - \mathbf{S}^{(t-1)}\|_F^2 + \lambda_2(\|b\tilde{\mathbf{S}}^{(t-1)} + (1-b)\mathbf{S}^{(t-1)}\|_{2,1} - \|\mathbf{S}^{(t-1)}\|_{2,1}), \end{aligned}$$

and, using the triangular inequality:

$$\begin{aligned}\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \quad &\leq \quad \mathcal{F}(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) + b\langle \boldsymbol{G}_S(\boldsymbol{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}), \tilde{\mathbf{S}}^{(t-1)} - \mathbf{S}^{(t-1)}\rangle \\ &\quad + \frac{b^2}{2\eta}\|\tilde{\mathbf{S}}^{(t-1)} - \mathbf{S}^{(t-1)}\|_F^2 + b\lambda_2(\|\tilde{\mathbf{S}}^{(t-1)}\|_{2,1} - \|\mathbf{S}^{(t-1)}\|_{2,1}). \end{aligned} \tag{58}$$

Finally, minimizing the right hand side of (58) with respect to b, we obtain the final result:

$$\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \mathcal{F}(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \leq \frac{-\eta g_S(\mathbf{S}^{(t-1)}, \boldsymbol{L}^{(t-1)})^2}{(2Q^{(t-1)})^2},$$

where we have used that $\|\tilde{\mathbf{S}}^{(t-1)} - \mathbf{S}^{(t-1)}\|_F^2 \leq (2Q^{(t-1)})^2$.

A.8.5 Proof of Lemma 5

We first observe, using a Taylor expansion of the quadratic term of the objective function (the least-squares data fitting term plus the ridge regularization term), and (26) that:

$$\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) = \mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \beta_t g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) + \frac{\beta_t^2(1+\epsilon)}{2}\|\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2.$$

Now, recall that

$$\beta_t = \min\left\{1, \frac{\langle \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}, \boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})\rangle + \lambda_1(R^{(t-1)} - \tilde{R}^{(t)})}{(1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2}\right\},$$

with $(\tilde{\boldsymbol{L}}^{(t)}, \tilde{R}^{(t)})$ defined in (27), and g_L in (31).

Case 1: $\langle \boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}), \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\rangle + \lambda_1(R^{(t-1)} - \tilde{R}^{(t)}) \geq (1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2$. Then, $\beta_t = 1$, and $g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \geq (1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2$. As a result, we observe:

$$\begin{aligned}\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) - \mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \quad &\leq \quad -\frac{1}{2}g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \\ &\leq \quad -\frac{1}{2}\frac{(g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}))^2}{g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})} \\ &\leq \quad -\frac{1}{2}\frac{(g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}))^2}{\bar{R}^{(t)}(\lambda_1 + 2M^{(t)})}, \end{aligned} \tag{59}$$

where, to obtain the last inequality, we have used that $M^{(t)} = \|\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})\|_F \geq \|\boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)})\|_{op}$, and the inequalities $R^{(t-1)} - \tilde{R}^{(t)} \leq \bar{R}^{(t)}$ and

$$\langle \boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}), \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t-1)}\rangle \leq 2M^{(t)}\bar{R}^{(t)}.$$

Case 2: $\langle \boldsymbol{G}_L(\boldsymbol{S}^{(t)}, \boldsymbol{L}^{(t-1)}), \boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\rangle + \lambda_1(R^{(t-1)} - \tilde{R}^{(t)}) < (1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\|_F^2$. Then, $\beta_t = g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})/((1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\|_F^2)$, and we obtain:

$$\begin{aligned}\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) - \mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \quad &\leq \quad -\frac{1}{2}\frac{(g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}))^2}{(1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\|_F^2} \\ &\leq \quad -\frac{1}{2}\frac{(g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}))^2}{(1+\epsilon)(2\bar{R}^{(t)})^2}, \end{aligned}$$

where, to obtain the last inequality, we used that $\|\boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\|_F^2 \leq \|\boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\|_*^2 \leq (2\bar{R}^{(t)})^2$.

We finally prove (32) as follows. We start by noticing that $\|\tilde{\boldsymbol{L}}^{(t-1)} - \boldsymbol{L}^{(t-1)}\|_F^2 = \beta_t^2 \|\boldsymbol{L}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2$. If $\beta_t = 1$, then by definition of β_t:

$$g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \geq (1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t-1)} - \boldsymbol{L}^{(t-1)}\|_F^2 = (1+\epsilon)\|\boldsymbol{L}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2.$$

Inequality (59) then implies that:

$$\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)}) - \mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) \leq -\frac{(1+\epsilon)}{2}\|\boldsymbol{L}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2.$$

If $\beta_t = g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)})/((1+\epsilon)\|\boldsymbol{L}^{(t-1)} - \tilde{\boldsymbol{L}}^{(t)}\|_F^2)$, then:

$$\begin{aligned}
\|\tilde{\boldsymbol{L}}^{(t-1)} - \boldsymbol{L}^{(t-1)}\|_F^2 &= \beta_t^2 \|\boldsymbol{L}^{(t)} - \boldsymbol{L}^{(t-1)}\|_F^2 = \frac{(g_L(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}))^2}{(1+\epsilon)\|\tilde{\boldsymbol{L}}^{(t-1)} - \boldsymbol{L}^{(t-1)}\|_F^2} \\
&\leq \frac{2}{1+\epsilon}\left(\mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t-1)}, R^{(t-1)}) - \mathcal{F}(\mathbf{S}^{(t)}, \boldsymbol{L}^{(t)}, R^{(t)})\right),
\end{aligned}$$

which proves the result.

A.8.6 Proof of Lemma 7

To prove equation (42) in Lemma 7, recall that for $j \in \mathcal{I}$, $\sum_{i\in\mathcal{I}} \mathbb{E}\left[\boldsymbol{\Omega}_{ij}\boldsymbol{\Sigma}_{ij}^2\right] \leq n\nu_n\rho_n$, that $\sum_{i\in\mathcal{I}} \mathbb{V}ar\left[\boldsymbol{\Omega}_{ij}\boldsymbol{\Sigma}_{ij}^2\right] \leq n\nu_n\rho_n$, and that $\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma} \odot \boldsymbol{\Sigma}\|_\infty \leq 1$. Applying Bernstein's inequality (21), we obtain that for any $j \in \mathcal{I}$ and $t > 0$,

$$\mathbb{P}\left(\sum_{i\in\mathcal{I}} \boldsymbol{\Omega}_{ij}\boldsymbol{\Sigma}_{ij}^2 \geq \nu_n\rho_n n + \sqrt{2t\nu_n\rho_n n} + \frac{3}{2}t\right) \leq 2e^{-t}$$

Choosing $t = 2\nu_n\rho_n n$, we find that

$$\begin{aligned}
\mathbb{P}\left(\max_{j\in\mathcal{I}} \sqrt{\sum_{i\in\mathcal{I}} \boldsymbol{\Omega}_{ij}\boldsymbol{\Sigma}_{ij}^2} \geq \sqrt{6\nu_n\rho_n n}\right) &\leq 2ne^{-2\nu_n\rho_n n} \\
\mathbb{P}\left(\left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{2,\infty} \geq \sqrt{6\nu_n\rho_n n}\right) &\leq 2e^{-\nu_n\rho_n n}
\end{aligned}$$

where we have used the union bound and $\nu_n\gamma_n n \geq \log(n)$. This proves equation (42) in Lemma 3.

To prove equation (43) in Lemma 7, note that $\left\|\boldsymbol{\Omega}_{|I}\right\|_{2,\infty} \leq \left\|\boldsymbol{\Pi}_{|I} - \boldsymbol{\Omega}_{|I}\right\|_{2,\infty} + \left\|\boldsymbol{\Pi}_{|I}\right\|_{2,\infty}$ and $\left\|\boldsymbol{\Pi}_{|I}\right\|_{2,\infty} \leq \sqrt{\nu_n n}$. Moreover, for $j \in \mathcal{I}$, $\sum_{i\in\mathcal{I}} \mathbb{E}\left[\left(\boldsymbol{\Pi}_{ij} - \boldsymbol{\Omega}_{ij}\right)^2\right] \leq \nu_n n$, $\sum_{i\in\mathcal{I}} \mathbb{V}ar\left[\left(\boldsymbol{\Pi} - \boldsymbol{\Omega}_{ij}\right)^2\right] \leq \nu_n n$, and $\left\|\boldsymbol{\Pi}_{|I} - \boldsymbol{\Omega}_{|I}\right\|_\infty \leq 1$. We apply Bernstein's inequality and find that for any $j \in \mathcal{I}$ and $t > 0$,

$$\mathbb{P}\left(\sum_{i\in\mathcal{I}} \left(\boldsymbol{\Pi}_{ij} - \boldsymbol{\Omega}_{ij}\right)^2 \geq \nu_n n + \sqrt{2t\nu_n n} + \frac{3}{2}t\right) \leq 2e^{-t}$$

Choosing $t = 2\nu_n n$ and using an union bound, we find that

$$\begin{aligned}
\mathbb{P}\left(\sup_{j\in\mathcal{I}} \sqrt{\sum_{i\in\mathcal{I}} \left(\boldsymbol{\Pi}_{ij} - \boldsymbol{\Omega}_{ij}\right)^2} \geq \sqrt{6\nu_n n}\right) &\leq 2ne^{-2n\nu_n} \\
\mathbb{P}\left(\left\|\boldsymbol{\Pi}_{|I} - \boldsymbol{\Omega}_{|I}\right\|_{2,\infty} \geq \sqrt{6\nu_n n}\right) &\leq 2e^{-\nu_n n}
\end{aligned}$$

where we have used that $\nu_n n \geq \log(n)$. This proves equation (43).

To prove equation (44), recall that for $(i,j) \in \mathcal{O} \times \mathcal{I}$, $\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} \sim$Bernoulli$(\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij})$, and that $\left\|\boldsymbol{\Pi} \odot \boldsymbol{S}^\top\right\|_\infty \leq \nu_n\gamma_n$. Then, applying Bernstein's inequality (21), we find that for any $j \in \mathcal{I}$ and any $t > 0$,

$$\mathbb{P}\left(\sum_{i\in\mathcal{O}}\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} \geq s\nu_n\gamma_n + \sqrt{2ts\nu_n\gamma_n} + \frac{3t}{2}\right) \leq 2e^{-t}.$$

Choosing $t = 2\nu_n\rho_n n$, we find that

$$\mathbb{P}\left(\sum_{i\in\mathcal{O}}\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} \geq s\nu_n\gamma_n + 2\sqrt{\gamma_n\rho_n n s}\nu_n + 3\nu_n\rho_n n\right) \leq 2e^{-t}.$$

Under Assumption 3, this implies

$$\mathbb{P}\left(\sum_{i\in\mathcal{O}}\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} \geq 6\nu_n\rho_n n\right) \leq 2e^{-2\nu_n\rho_n n}.$$

Using the union bound, and the bound $\nu_n\rho_n n \geq \log(n)$, we conclude that

$$\mathbb{P}\left(\max_{j\in\mathcal{I}}\sqrt{\sum_{i\in\mathcal{O}}\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij}} \geq \sqrt{6\nu_n\rho_n nn}\right) \leq 2ne^{-2\nu_n\rho_n n} \leq 2e^{-\nu_n\rho_n n}.$$

This concludes the proof of Lemma 7.

A.8.7 Proof of Lemma 8

Recall that for $j \in \mathcal{O}$, $\left\{\left(\left(\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} - \rho_n\right)_+\right)^2\right\}_{i\in\mathcal{I}}$ are independent random variables. Moreover, easy calculations yields that $\mathbb{E}\left[\left(\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \rho_n\right)_+\right)^2\right] = \boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2$, and that $\mathbb{V}ar\left[\left(\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \rho_n\right)_+\right)^2\right] \leq \boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2$. Applying Bernstein's inequality (21), we see that for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i\in\mathcal{I}}\mathbb{E}\left[\left(\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \rho_n\right)_+\right)^2\right] - \sum_{i\in\mathcal{I}}\left(\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \rho_n\right)_+\right)^2\right| \geq \sqrt{2t\sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2} + \frac{3t}{2}\right) \leq 2e^{-t}.$$

Choosing $t = \frac{1}{80}\sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2$, we find that

$$\mathbb{P}\left(\sum_{i\in\mathcal{I}}\left(\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \rho_n\right)\right)_+^2 \leq \sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2 - \frac{1}{2}\sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2\right) \leq 2e^{-\frac{1}{80}\sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2}.$$

When $\min_{j\in\mathcal{O}}\sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}(1-\rho_n)^2 \geq \nu_n\rho_n n$ and $\rho_n \leq \frac{1}{2}$, this implies that

$$\mathbb{P}\left(\min_{j\in\mathcal{O}}\sqrt{\sum_{i\in\mathcal{I}}\left(\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \rho_n\right)_+\right)^2} \leq \frac{1}{4}\min_{j\in\mathcal{O}}\sqrt{\sum_{i\in\mathcal{I}}\boldsymbol{\Pi}_{ij}\boldsymbol{S}^*_{ij}}\right) \leq 2se^{-\frac{\nu_n\rho_n n}{80}}.$$

A.8.8 Proof of Lemma 9

Let $\partial\left\|\cdot\right\|_*$ and $\partial\left\|\cdot\right\|_{2,1}$ denote respectively the sub-differentials of $\left\|\cdot\right\|_*$ and $\left\|\cdot\right\|_{2,1}$ norms. Recall that $\left(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{L}}\right)$ minimizes $\mathcal{F}$. The standard optimality condition over a convex set states that for any admissible matrix $(\boldsymbol{S}, \boldsymbol{L})$, there exists $\widehat{\boldsymbol{V}} \in \partial\left\|\widehat{\boldsymbol{S}}\right\|_{2,1}$ and $\widehat{\boldsymbol{W}} \in \partial\left\|\widehat{\boldsymbol{L}}\right\|_*$ such that

$$
\begin{aligned}
-\left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top - \widehat{\boldsymbol{L}}\right) \middle| \boldsymbol{S} - \widehat{\boldsymbol{S}} + \boldsymbol{S}^\top - \widehat{\boldsymbol{S}}^\top + \boldsymbol{L} - \widehat{\boldsymbol{L}} \right\rangle & \\
+\lambda_1 \left\langle \widehat{\boldsymbol{W}} \middle| \boldsymbol{L} - \widehat{\boldsymbol{L}} \right\rangle + \lambda_2 \left\langle \widehat{\boldsymbol{V}} \middle| \boldsymbol{S} - \widehat{\boldsymbol{S}} \right\rangle & \geq 0
\end{aligned} \tag{60}
$$

Applying equation (60) for the admissible matrices $\left(\widehat{\boldsymbol{S}}, \boldsymbol{L}^*\right)$, we find that there exists $\widehat{\boldsymbol{W}} \in \partial \left\|\widehat{\boldsymbol{L}}\right\|_*$ such that

$$
-\left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top - \widehat{\boldsymbol{L}}\right) \middle| \Delta \boldsymbol{L} \right\rangle + \lambda_1 \left\langle \widehat{\boldsymbol{W}} \middle| \Delta \boldsymbol{L} \right\rangle \geq 0. \tag{61}
$$

Recall that $\boldsymbol{\Sigma}_{|I} \triangleq \boldsymbol{A}_{|I} + \operatorname{diag}(\boldsymbol{L}^*) - \boldsymbol{L}^*$, that $\Delta \boldsymbol{L} \triangleq \boldsymbol{L}^* - \widehat{\boldsymbol{L}}$, and that $\boldsymbol{\Omega} \odot \operatorname{diag}(\boldsymbol{M}) = 0$ for any matrix $\boldsymbol{M}$. Thus, equation (61) becomes

$$
-\left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{\Sigma}_I + \Delta \boldsymbol{L} + \boldsymbol{A}_{|O} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right) \middle| \Delta \boldsymbol{L} \right\rangle + \lambda_1 \left\langle \widehat{\boldsymbol{W}} \middle| \Delta \boldsymbol{L} \right\rangle \geq 0. \tag{62}
$$

Developing equation (62), we find that

$$
\begin{aligned}
-\left\langle \boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I} \middle| \Delta \boldsymbol{L} \right\rangle - \left\langle \boldsymbol{\Omega} \odot \Delta \boldsymbol{L} \middle| \Delta \boldsymbol{L} \right\rangle - \left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle & \\
+\left\langle \boldsymbol{\Omega} \odot \left(\widehat{\boldsymbol{S}} + \widehat{\boldsymbol{S}}^\top\right)_{|I} \middle| \Delta \boldsymbol{L} \right\rangle + \lambda_1 \left\langle \widehat{\boldsymbol{W}} \middle| \Delta \boldsymbol{L} \right\rangle & \leq 0.
\end{aligned}
$$

We have proved in Theorem 3 that $\widehat{\boldsymbol{S}}_{|I} = \widehat{\boldsymbol{S}}_{|I}^\top = \boldsymbol{0}$ with probability at least $1 - 6e^{-\nu_n \rho_n n}$. Therefore, when equation (17) holds,

$$
\|\boldsymbol{\Omega} \odot \Delta \boldsymbol{L}\|_F^2 \leq \left| \left\langle \boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I} \middle| \Delta \boldsymbol{L} \right\rangle \right| + \left| \left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle \right| + \lambda_1 \left\langle \widehat{\boldsymbol{W}} \middle| \Delta \boldsymbol{L} \right\rangle.
$$

Using the duality of the $\|\cdot\|_*$-norm and the $\|\cdot\|_{op}$-norm, we find that

$$
\|\boldsymbol{\Omega} \odot \Delta \boldsymbol{L}\|_F^2 \leq \left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{op} \|\Delta \boldsymbol{L}\|_* + \left| \left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle \right| + \lambda_1 \left\langle \widehat{\boldsymbol{W}} \middle| \Delta \boldsymbol{L} \right\rangle.
$$

Next, we bound the term $\left| \left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle \right|$ using the following Lemma.

Lemma 12. *With probability at least* $1 - 2e^{-\tilde{\nu}_n \gamma_n s n}$,

$$
\left| \left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle \right| \leq 16 \tilde{\nu}_n \gamma_n \rho_n n s.
$$

Proof. See Section A.8.11. □

Finally, we bound $\left\langle \widehat{\boldsymbol{W}} \middle| \Delta \boldsymbol{L} \right\rangle$. Note that by definition of the subgradient, $\left\langle \widehat{\boldsymbol{W}} \middle| \boldsymbol{L}^* - \widehat{\boldsymbol{L}} \right\rangle \leq \|\boldsymbol{L}^*\|_* - \left\|\widehat{\boldsymbol{L}}\right\|_*$. Using the separability of the spectral norm on orthogonal subspaces and the identity $\mathcal{P}_{\boldsymbol{L}^*}(\boldsymbol{L}^*) = \boldsymbol{L}^*$, we find that

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{L}}\right\|_* &= \left\|\mathcal{P}_{\boldsymbol{L}^*}^\perp(\Delta \boldsymbol{L}) + \mathcal{P}_{\boldsymbol{L}^*}(\Delta \boldsymbol{L}) - \boldsymbol{L}^*\right\|_* \\
&= \left\|\mathcal{P}_{\boldsymbol{L}^*}^\perp(\Delta \boldsymbol{L})\right\|_* + \left\|\mathcal{P}_{\boldsymbol{L}^*}(\Delta \boldsymbol{L}) - \boldsymbol{L}^*\right\|_* \\
&\geq \left\|\mathcal{P}_{\boldsymbol{L}^*}^\perp(\Delta \boldsymbol{L})\right\|_* + \|\boldsymbol{L}^*\|_* - \left\|\mathcal{P}_{\boldsymbol{L}^*}(\Delta \boldsymbol{L})\right\|_*.
\end{aligned}
$$

Combining this result with Lemma 12, we find that with probability at least $1 - 6e^{-\nu_n \rho_n n} - 2e^{-\tilde{\nu}_n \gamma_n s n}$,

$$
\|\boldsymbol{\Omega} \odot \Delta \boldsymbol{L}\|_F^2 \leq \left\|\boldsymbol{\Omega} \odot \boldsymbol{\Sigma}_{|I}\right\|_{op} \left(\left\|\mathcal{P}_{\boldsymbol{L}^*}(\Delta \boldsymbol{L})\right\|_* + \left\|\mathcal{P}_{\boldsymbol{L}^*}^\perp(\Delta \boldsymbol{L})\right\|_*\right) + 16 \tilde{\nu}_n \gamma_n \rho_n s n + \lambda_1 \left(\left\|\mathcal{P}_{\boldsymbol{L}^*}(\Delta \boldsymbol{L})\right\|_* - \left\|\mathcal{P}_{\boldsymbol{L}^*}^\perp(\Delta \boldsymbol{L})\right\|_*\right).
$$

Recall that by definition, $\Psi \geq 16\tilde{\nu}_n\gamma_n\rho_n ns$. Thus, when $\lambda_1 \geq 3\left\|\boldsymbol{\Omega}\odot\boldsymbol{\Sigma}_{|I}\right\|_{op}$,

$$\|\boldsymbol{\Omega}\odot\Delta\boldsymbol{L}\|_F^2 \quad\leq\quad \frac{\lambda_1}{3}\left(5\left\|\mathcal{P}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_* - \left\|\mathcal{P}^{\perp}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_*\right)+\Psi.$$

This proves equation (47) in Lemma 9. This result also implies that

$$\left\|\mathcal{P}^{\perp}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_* \quad\leq\quad 5\left\|\mathcal{P}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_* + \frac{3\Psi}{\lambda_1}.$$

Recall that $\boldsymbol{L}^*$ is of rank k and so $\mathcal{P}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)$ is of rank at most k. Therefore,

$$\begin{aligned}\|\Delta\boldsymbol{L}\|_* &\leq 6\left\|\mathcal{P}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_* + \frac{3\Psi}{\lambda_1} \leq 6\sqrt{k}\left\|\mathcal{P}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_F + \frac{3\Psi}{\lambda_1}\\ &\leq 6\sqrt{k}\left\|\Delta\boldsymbol{L}\right\|_F + \frac{3\Psi}{\lambda_1} \leq 6\sqrt{k}\left\|\Delta\boldsymbol{L}_{|I}\right\|_F + 6\sqrt{k(sn+s^2)}\rho_n + \frac{3\Psi}{\lambda_1}\\ &\leq 6\sqrt{k}\left\|\Delta\boldsymbol{L}_{|I}\right\|_F + 6\sqrt{3ksn}\rho_n + \frac{3\Psi}{\lambda_1}.\end{aligned}$$

where we have used that $\left\|\Delta\boldsymbol{L}_{|O}\right\|_F \leq \sqrt{|O|}\left\|\Delta\boldsymbol{L}_{|O}\right\|_\infty \leq \sqrt{s^2+2sn}\rho_n$. This completes the proof of Lemma 9.

A.8.9 Proof of Lemma 10

For ease of notations, let $\alpha = 36^2\frac{\nu_n\rho_n^2 kn}{\mu_n}$. To prove Lemma 9, we consider the following two cases.

Case 1: $\left\|\Delta\boldsymbol{L}_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} \leq \alpha$. Then the result is immediate.

Case 2: $\left\|\Delta\boldsymbol{L}_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} > \alpha$. Let $r>0$ a constant to be specified later. We consider the following sets

$$\mathcal{S}^r = \left\{\boldsymbol{M}\in\mathbb{R}^{n\times n}_{sym} : \|\boldsymbol{M}\|_\infty \leq \rho_n, \left\|\boldsymbol{M}_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} \geq \alpha, \|\boldsymbol{M}\|_* \leq \sqrt{r}\left\|\boldsymbol{M}_{|I}\right\|_F + \sqrt{3rsn}\rho_n + \frac{3\Psi}{\lambda_1}\right\}.$$

Recall that the random noise matrix $\boldsymbol{\Gamma}$ is defined as follows: for any $(i,j)\in[n]\times[n]$, $i<j$, $\boldsymbol{\Gamma}_{ij} = \boldsymbol{\Gamma}_{ji} = \boldsymbol{\Omega}_{ij}\epsilon_{ij}$ where $(\epsilon_{ij})_{1\leq i<j\leq n}$ is a Rademacher sequence. Now, we define β_r as follows :

$$\beta_r \triangleq \mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right]\left(\frac{64r\rho_n^2}{\mu_n}\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] + 15\sqrt{srn}\rho_n^2 + \frac{32\Psi\rho_n}{\lambda_1}\right).$$

Lemma 13. *With probability larger than $1-e^{-\nu_n\rho_n n}$, simultaneously for any $\boldsymbol{M}\in\mathcal{S}^r$,*

$$\frac{1}{2}\|\boldsymbol{M}\|^2_{L_2(\boldsymbol{\Pi})} \leq \left\|\boldsymbol{\Omega}\odot\boldsymbol{M}_{|I}\right\|_F^2 + \beta_r$$

Proof. See Section A.8.12. □

Recall that β was defined in equation (49), and note that $\beta = \beta_{36k}$. Then, equation (48) in Lemma 9 implies that $\Delta\boldsymbol{L}\in\mathcal{S}^{36k}$ with probability at least $1-6e^{-\nu_n\rho_n n} - 2e^{-\tilde{\nu}_n\gamma_n sn}$. Combining equation (47) in Lemma 9 and Lemma 13, we find that with probability at least $1-7e^{-n\nu_n\rho_n} - 2e^{-\tilde{\nu}_n\gamma_n sn}$,

$$\frac{1}{2}\left\|\Delta\boldsymbol{L}_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} \quad\leq\quad \frac{5\lambda_1}{3}\left\|\mathcal{P}_{\boldsymbol{L}^*}\left(\Delta\boldsymbol{L}\right)\right\|_* + \Psi + \beta.$$

The matrix $\boldsymbol{L}^*$ is of rank at most k. Therefore,

$$\begin{aligned}\left\|\Delta\boldsymbol{L}_{|I}\right\|^2_{L_2(\boldsymbol{\Pi})} &\leq \frac{10\lambda_1\sqrt{k}}{3}\left\|\Delta\boldsymbol{L}\right\|_F + 2\Psi + 2\beta \leq \frac{50\lambda_1^2 k}{9\mu_n} + \frac{\mu_n}{2}\|\Delta\boldsymbol{L}\|_F^2 + \Psi + \beta\\ &\leq \frac{50\lambda_1^2 k}{9\mu_n} + \frac{\mu_n}{2}\left\|\Delta\boldsymbol{L}_{|I}\right\|_F^2 + \frac{3}{2}\mu_n\rho_n^2 sn + \Psi + \beta\end{aligned}$$

where we have used that $\left\|\Delta \boldsymbol{L}_{|O}\right\|_F^2 \leq 3\rho_n^2 ns$. Using equation (14), we find that

$$\left\|\Delta \boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 \quad \leq \quad \frac{1}{2}\left\|\Delta \boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 + \frac{\mu_n}{2}\rho_n^2 n + \frac{3}{2}\mu_n\rho_n^2 sn + \frac{50\lambda_1^2 k}{9\mu_n} + \Psi + \beta.$$

Thus

$$\left\|\Delta \boldsymbol{L}_{|I}\right\|_{L_2(\boldsymbol{\Pi})}^2 \quad \leq \quad 8\mu_n\rho_n^2 sn + \frac{100\lambda_1^2 k}{9\mu_n} + 2\Psi + 2\beta.$$

We conclude the proof of Lemma 10 by recalling that $\mu_n \leq \nu_n$.

A.8.10 Proof of Lemma 11

To prove Lemma 11, we use Proposition 1. For $(i,j) \in I$, set $b_{ij} = \sqrt{\boldsymbol{\Pi}_{ij}}$, and $\xi_{ij} = \frac{\epsilon_{ij}\boldsymbol{\Omega}_{ij}}{b_{ij}}$, and for $i \in \mathcal{I}$ set $b_{ii} = 0$. Note that for any $(i,j) \in \mathcal{I}$, $\boldsymbol{\Gamma}_{ij} = b_{ij}\xi_{ij}$, and that $\{\xi_{ij}\}_{i\leq j}$ is a sequence of independent symmetric random variables with unit variance. Moreover, for any $(i,j) \in I$, $|b_{ij}\xi_{ij}| \leq 1$, so for any $\alpha \geq 3$, $\left(\mathbb{E}\left[\left(\xi_{ij}b_{ij}\right)^{2\alpha}\right]\right)^{\frac{1}{2\alpha}} \leq 1$. Finally, note that for any $i \in \mathcal{I}$,

$$\sqrt{\sum_{j\in\mathcal{I}} b_{ij}^2} = \sqrt{\sum_{j\in\mathcal{I}} \boldsymbol{\Pi}_{ij}} \leq \sqrt{\nu_n n}.$$

Applying Proposition 1, we find that

$$\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] \leq e^{\frac{2}{3}}\left(\sqrt{\nu_n n} + 42\sqrt{\log(n)}\right)$$

We conclude this proof by recalling that $\nu_n n \geq \log(n)$.

A.8.11 Proof of Lemma 12

To prove Lemma 12, note that $\left\|\Delta \boldsymbol{L}\right\|_\infty \leq \rho_n$, and therefore

$$\left|\left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle\right| \quad \leq \quad 2\rho_n \sum_{(i,j)\in O} \left|\boldsymbol{\Omega}_{ij}\left(\boldsymbol{A}_{ij} - \widehat{\boldsymbol{S}}_{ij} - \widehat{\boldsymbol{S}}_{ji}\right)\right|. \tag{63}$$

Recall that $\widehat{\boldsymbol{L}}$ and $\widehat{\boldsymbol{S}}$ have non-negative entries, and that $\widehat{\boldsymbol{L}}$ and $\boldsymbol{A}$ are symmetric. Therefore, equation (55) implies that $\left\{\widehat{\boldsymbol{S}}_{ij} = 0 \text{ or } \widehat{\boldsymbol{S}}_{ji} = 0\right\} \Rightarrow \boldsymbol{A}_{ij} = 0$, and that $\widehat{\boldsymbol{S}}_{ij} + \widehat{\boldsymbol{S}}_{ji} \leq \boldsymbol{A}_{ij}$. Thus, equation (63) implies

$$\left|\left\langle \boldsymbol{\Omega} \odot \left(\boldsymbol{A} - \widehat{\boldsymbol{S}} - \widehat{\boldsymbol{S}}^\top\right)_{|O} \middle| \Delta \boldsymbol{L} \right\rangle\right| \quad \leq \quad 2\rho_n \sum_{(i,j)\in O} \boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij}. \tag{64}$$

To conclude the proof of Lemma 12, we first prove the following result:

$$\mathbb{P}\left(\sum_{(i,j)\in O} \boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} \geq 8\tilde{\nu}_n\gamma_n sn\right) \quad \leq \quad \exp(-\tilde{\nu}_n\gamma_n sn). \tag{65}$$

We use Bernstein's inequality to obtain equation (65). Note that $\{\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij}\}_{(i,j)\in O, i<j}$ is a sequence of independent Bernoulli random variables such that for any $i \in [n]$, $\sum_{j\in\mathcal{O}} \mathbb{E}\left[\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij}\right] \leq \tilde{\nu}_n\gamma_n s$, $\sum_{j\in\mathcal{O}} \mathbb{V}ar\left[\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij}\right] \leq \tilde{\nu}_n\gamma_n s$, and $\left(\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} - \mathbb{E}\left[\boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij}\right]\right) \in [-1,1]$. Hence, applying Bernstein's inequality (21), we find that for any $t > 0$,

$$\mathbb{P}\left(\sum_{(i,j)\in O} \boldsymbol{\Omega}_{ij}\boldsymbol{A}_{ij} \geq 2\tilde{\nu}_n\gamma_n sn + \sqrt{2t \times \tilde{\nu}_n\gamma_n sn} + \frac{3t}{2}\right) \quad \leq \quad 2\exp(-t).$$

Choosing $t = 2\tilde{\nu}_n\gamma_n sn$, we obtain equation (65). We conclude the proof of Lemma 12 by combining equations (64) and (65).

A.8.12 Proof of Lemma 13

To prove Lemma 13, we show that the probability of the following "bad" event is small :

$$\mathcal{B} \triangleq \{\exists \boldsymbol{M} \in \mathcal{S}^r \text{ such that } \left| \left\| \boldsymbol{\Omega} \odot \boldsymbol{M}_{|I} \right\|_F^2 - \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \right| \geq \frac{1}{2} \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 + \beta_r \}.$$

We use a standard peeling argument to control the probability of the event $\mathcal{B}$. For $T > \alpha$, define

$$\mathcal{S}(T) \triangleq \left\{ \boldsymbol{M} \in \mathcal{S}^r : \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \leq T \right\}, \; Z(T) = \sup_{\boldsymbol{M} \in \mathcal{S}(T)} \left| \left\| \boldsymbol{\Omega} \odot \boldsymbol{M}_{|I} \right\|_F^2 - \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \right|, \text{ and}$$

$$\mathcal{B}(T) \triangleq \left\{ \exists \boldsymbol{M} \in \mathcal{S}(T) : \left| \left\| \boldsymbol{\Omega} \odot \boldsymbol{M}_{|I} \right\|_F^2 - \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \right| \geq \frac{T}{4} + \beta_r \right\} = \left\{ Z(T) \geq \frac{T}{4} + \beta \right\}.$$

For $l \geq 1$, define also $\mathcal{S}_l \triangleq \left\{ \boldsymbol{M} \in \mathcal{S}^r : 2^{l-1}\alpha < \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \leq 2^l \alpha \right\} \subset \mathcal{S}\left(2^l \alpha\right)$ and

$$\begin{aligned} \mathcal{B}_l &\triangleq \left\{ \exists \boldsymbol{M} \in \mathcal{S}_l : \left| \left\| \boldsymbol{\Omega} \odot \boldsymbol{M}_{|I} \right\|_F^2 - \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \right| \geq \frac{\left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2}{2} + \beta_r \right\} \\ &\subset \left\{ \exists \boldsymbol{M} \in \mathcal{S}_l : \left| \left\| \boldsymbol{\Omega} \odot \boldsymbol{M}_{|I} \right\|_F^2 - \left\| \boldsymbol{M}_{|I} \right\|_{L_2(\boldsymbol{\Pi})}^2 \right| \geq \frac{2^{l-1}}{2} \alpha + \beta_r \right\} \subset \mathcal{B}\left(2^l \alpha\right). \end{aligned}$$

Since $\mathcal{S}^r \subset \underset{l \geq 1}{\cup} \mathcal{S}_l$, it is easy to see that $\mathcal{B} \subset \underset{l \geq 1}{\cup} \mathcal{B}_l$. To control the probability of the events $\mathcal{B}_l$, it is enough to control the probability of the events $\mathcal{B}(T)$, which is done in the following lemma.

Lemma 14. *For any $T \geq \alpha$, we have $\mathbb{P}\left(\mathcal{B}(T)\right) \leq \exp(-\frac{T}{36^2 \rho_n})$.*

Proof. See Section A.8.13. □

We apply Lemma 14 to find

$$\begin{aligned} \mathbb{P}\left(\mathcal{B}\right) &\leq \sum_{l \geq 1} \mathbb{P}\left(\mathcal{B}_l\right) \leq \sum_{l \geq 1} \exp\left(-\frac{2^l \alpha}{36^2 \rho_n} \right) \\ &\leq \sum_{l \geq 1} \exp\left(-\frac{2l\alpha}{36^2 \rho_n} \alpha \right) = \frac{\exp\left(-\frac{2\alpha}{36^2 \rho_n}\right)}{1 - \exp\left(-\frac{2\alpha}{36^2 \rho_n}\right)} = \frac{\exp\left(-2\frac{\nu_n \rho_n k n}{\mu_n}\right)}{1 - \exp\left(-2\frac{\nu_n \rho_n k n}{\mu_n}\right)} \end{aligned}$$

Note that $\frac{\nu_n \rho_n k n}{\mu_n} \geq \nu_n \rho_n n \geq \log(n) \geq 1$, and so $\mathbb{P}\left[\mathcal{B}\right] \leq \frac{1}{2} \exp\left(-2\nu_n \rho_n n\right) \leq \exp\left(-\nu_n \rho_n n\right)$. This concludes the proof of Lemma 13.

A.8.13 Proof of Lemma 14

Recall that $Z(T) = 2 \sup_{\boldsymbol{M} \in \mathcal{S}(T)} \left| \sum_{(i,j) \in I} \boldsymbol{M}_{ij}^2 \left(\boldsymbol{\Omega}_{ij} - \boldsymbol{\Pi}_{ij} \right) \right|$, since all matrices in $\mathcal{S}$ are symmetric. In order to bound $Z(T)$, we begin by controlling the deviation of $Z(T)$ from its expectation. To do this, we apply Bousquet's Theorem 6 to the random variable $Z(T) = 2\rho_n \sup_{\boldsymbol{M} \in \mathcal{S}(T)} \left| \sum_{(i,j) \in I} f_{ij}^{\boldsymbol{M}}(\boldsymbol{\Omega}_{ij}) \right|$ where we set $f_{ij}^{\boldsymbol{M}}(\boldsymbol{\Omega}_{ij}) \triangleq \frac{(\boldsymbol{\Omega}_{ij} - \boldsymbol{\Pi}_{ij}) \boldsymbol{M}_{ij}^2}{\rho_n}$. The set of functions $\left\{ f_{ij}^{\boldsymbol{M}}, \boldsymbol{M} \in \mathcal{S}(T) \right\}$ is separable and we can apply Theorem 6 (see, e.g., [24], Section 2.1). Note that for any $(i,j) \in I$, $\mathbb{E}\left[f_{ij}^{\boldsymbol{M}}(\boldsymbol{\Omega}_{ij}) \right] = 0$, $\left| f_{ij}^{\boldsymbol{M}}(\boldsymbol{\Omega}_{ij}) \right| \leq 1$, $\mathbb{E}\left[\left(\boldsymbol{\Omega}_{ij} - \boldsymbol{\Pi}_{ij} \right)^2 \right] \leq \boldsymbol{\Pi}_{ij}$ and $\left\| \boldsymbol{M} \right\|_\infty \leq \rho_n$ so

$$v \triangleq 2 \sup_{\boldsymbol{M} \in \mathcal{S}(T)} \sum_{(i,j) \in I} \mathbb{E}\left[f_{ij}^{\boldsymbol{M}}(X_{ij})^2 \right] \leq 2 \sum_{(i,j) \in I} \boldsymbol{\Pi}_{ij} \frac{\boldsymbol{M}_{ij}^4}{\rho_n^2} \leq 2 \sup_{\boldsymbol{M} \in \mathcal{S}(T)} \sum_{(i,j) \in I} \boldsymbol{\Pi}_{ij} \boldsymbol{M}_{ij}^2 \leq T.$$

Theorem 6 implies that

$$\mathbb{P}\left(\frac{Z_T}{2\rho_n} > \frac{\mathbb{E}[Z_T]}{2\rho_n} + \frac{x}{3} + \sqrt{2x\left(\frac{2\mathbb{E}[Z_T]}{2\rho_n} + T\right)}\right) \leq \exp(-x)$$

$$\mathbb{P}\left(Z_T > \mathbb{E}[Z_T] + \frac{2\rho_n x}{3} + 2\rho_n x + 2\mathbb{E}[Z_T] + 2\rho_n\sqrt{2xT}\right) \leq \exp(-x)$$

where we used $2\sqrt{ab} \leq a + b$. Setting $x = \frac{T}{36^2\rho_n}$ and noticing that $\rho_n \leq 1$ leads to

$$\mathbb{P}\left(Z_T > 2\mathbb{E}[Z_T] + \frac{T}{8}\right) \leq \exp(-\frac{T}{36^2\rho_n}). \tag{66}$$

In a second time, in order to bound $\mathbb{E}[Z_T]$, we apply a standard symetrization argument (see, e.g., [37], Theorem 2.1). We obtain that

$$\mathbb{E}[Z(T)] \leq 4\mathbb{E}\left[\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left|\sum_{(i,j)\in I}\epsilon_{ij}\boldsymbol{M}_{ij}^2\boldsymbol{\Omega}_{ij}\right|\right] \tag{67}$$

where $(\epsilon_{ij})_{1\leq i<j\leq n}$ is a Rademacher sequence. For $i < j$, define $\phi_{ij} : x \to \frac{x^2}{2\rho_n}$. Recall that for any (i,j), $\boldsymbol{\Omega}_{ij} \in \{0,1\}$, and so $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Omega}_{ij}^2$. With these notations, equation (67) becomes

$$\mathbb{E}[Z(T)] \leq 8\rho_n\mathbb{E}\left[\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left|\sum_{i<j}\epsilon_{ij}\phi_{ij}\left(\boldsymbol{\Omega}_{ij}\boldsymbol{M}_{ij}\right)\right|\right].$$

We note that for $\boldsymbol{M} \in \mathcal{S}(T)$, $\|\boldsymbol{M}\|_\infty \leq \rho_n$. Therefore, the functions ϕ_{ij} are 1-Lipschitz functions on $[-\rho_n, \rho_n]$ vanishing at 0. We apply Talagrand's contraction principle (see, e.g., Theorem 2.2 in [37]) and find that

$$\mathbb{E}[Z(T)] \leq 16\rho_n\mathbb{E}\left[\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left|\sum_{(i,j)\in I}\epsilon_{ij}\boldsymbol{M}_{ij}\boldsymbol{\Omega}_{ij}\right|\right] = 8\rho_n\mathbb{E}\left[\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left|\left\langle\boldsymbol{M}\middle|\boldsymbol{\Gamma}_{|I}\right\rangle\right|\right]$$

where for any (i,j), $\boldsymbol{\Gamma}_{ij} = \epsilon_{ij}\boldsymbol{\Omega}_{ij}$. By the duality of the $\|\cdot\|_*$-norm and $\|\cdot\|_{op}$-norm, and by definition of $\mathcal{S}^r$, we find that

$$\begin{aligned}\mathbb{E}[Z(T)] &\leq 8\rho_n \sup_{\boldsymbol{M}\in\mathcal{S}(T)}\|\boldsymbol{M}\|_*\,\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] \\ &\leq 8\rho_n\left(\sqrt{r}\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left\|\boldsymbol{M}_{|I}\right\|_F + \sqrt{3rsn}\rho_n + \frac{3\Psi}{\lambda_1}\right)\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right].\end{aligned}$$

Using equation (14), we find that

$$\begin{aligned}\mathbb{E}[Z(T)] &\leq 8\rho_n\left(\sqrt{r}\left(\frac{1}{\sqrt{\mu_n}}\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left\|\boldsymbol{M}_{|I}\right\|_{L_2(\boldsymbol{\Pi})} + \sqrt{n}\rho_n\right) + \sqrt{3rsn}\rho_n + \frac{3\Psi}{\lambda_1}\right)\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] \\ &\leq \left(\frac{8\rho_n\sqrt{r}}{\sqrt{\mu_n}}\sup_{\boldsymbol{M}\in\mathcal{S}(T)}\left\|\boldsymbol{M}_{|I}\right\|_{L_2(\boldsymbol{\Pi})} + 8\sqrt{nr}\rho_n^2 + 8\sqrt{3srn}\rho_n^2 + \frac{32\Psi\rho_n}{\lambda_1}\right)\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right].\end{aligned}$$

Using the definition of $\mathcal{S}(T)$, we find that

$$\begin{aligned}\mathbb{E}[Z(T)] &\leq \left(\frac{8\rho_n\sqrt{rT}}{\sqrt{\mu_n}} + 8\sqrt{rn}\rho_n^2 + 8\sqrt{3srn}\rho_n^2 + \frac{32\Psi\rho_n}{\lambda_1}\right)\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] \\ &\leq \frac{T}{16} + \mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right]\left(\frac{64r\rho_n^2}{\mu_n}\mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{|I}\right\|_{op}\right] + 15\sqrt{srn}\rho_n^2 + \frac{32\Psi\rho_n}{\lambda_1}\right) \\ &= \frac{T}{16} + \beta^r.\end{aligned} \tag{68}$$

Combining equation (66) and equation (68) yields the desired result.

A.9 Proof of Lemma 6

Consider the following chain of inequality:

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \gamma_k \frac{A_k}{A_{k+1}} \geq \gamma_k,$$

since $A_{k+1} \leq A_k$. Thus, we obtain

$$\frac{1}{A_{k+1}} - \frac{1}{A_1} = \sum_{i=1}^{k} \left(\frac{1}{A_{i+1}} - \frac{1}{A_i} \right) \geq \sum_{i=1}^{k} \gamma_i,$$

which gives the result after reshuffling the terms.