Spectral Robustness for Correlation Clustering Reconstruction in Semi-Adversarial Models

Flavio Chierichetti
Sapienza University
Rome, Italy
flavio@di.uniroma1.it

Alessandro Panconesi
Sapienza University
Rome, Italy
ale@di.uniroma1.it

Giuseppe Re
Sapienza University
Rome, Italy
re@di.uniroma1.it

Luca Trevisan
Bocconi University
Milan, Italy
l.trevisan@unibocconi.it

Abstract

Correlation Clustering is an important clustering problem with many applications. We study the reconstruction version of this problem, in which one seeks to reconstruct a latent clustering that has been corrupted by random noise and adversarial modifications. Concerning the latter, there is a standard "post-adversarial" model in the literature, in which adversarial modifications come after the noise. Here, we introduce and analyse a "pre-adversarial" model, in which adversarial modifications come before the noise. Given an input coming from such a semi-adversarial generative model, the goal is to approximately reconstruct with high probability the latent clustering. We focus on the case where the hidden clusters have nearly equal size and show the following. In the pre-adversarial setting, spectral algorithms are optimal, in the sense that they reconstruct all the way to the information-theoretic threshold beyond which no reconstruction is possible. This is in contrast to the post-adversarial setting, in which their ability to restore the hidden clusters stops before the threshold, but the gap is optimally filled by SDP-based algorithms. These results highlight a heretofore unknown robustness of spectral algorithms, showing them less brittle than previously thought.

1 INTRODUCTION

The rigorous analysis of combinatorial algorithms is most often carried out as a *worst-case* analysis over all possible inputs. In some cases, worst-case analysis makes excessively pessimistic predictions of a given algorithm's running time or memory use, compared to its performance on typical data. In order to achieve more predictive rigorous analyses of algorithms, there has been interest in developing data models that go "beyond worst-case analysis," combining adversarial choices and random choices. A notable example is the framework of *smoothed analysis*, introduced by Spielman and Teng (2004), where a worst-case instance is perturbed by random noise. In a complementary way, several semi-random generative models have been studied in which a random instance is perturbed by an adversary. The monograph by Roughgarden (2020) surveys this active research program.

In unsupervised machine learning, the goal is to discover structure in data that is presented in an unstructured way. Since unsupervised machine learning is postulated on the existence of a "ground truth" or "latent structure" that we want to discover, the rigorous analysis of an unsupervised learning algorithm must be carried out according to generative models that produce both a data set and a ground truth about the data set so that one can analyze whether the algorithm is able to discover the latter from the former.

Previous work on the rigorous analysis of unsupervised machine learning algorithms has typically been done according to a fixed, purely probabilistic, generative model. This type of analysis can make excessively optimistic predictions about the performance of a given algorithm, particularly if the algorithm is "overfit" to a particular generative model. In order to study the robustness of algorithms to data coming from sources whose distribution does not perfectly fit a simple probabilistic generative model, there has been interest in going "beyond average-case" in the analysis of unsupervised learning algorithms, introducing semi-adversarial models that combine probabilistic generation and adversarial choices. For example, in *robust statistics*, one is interested in inferring the parame-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

ters of a distribution given a mix of samples from the distribution and adversarially selected outliers. Several semi-adversarial network generation models have been studied for community detection and clustering problems (see Sec. 2 for a review). Such models, as our models, give a way to understand whether the average-case analysis of the performance of an algorithm is robust to deviations of the input distribution from the assumed probabilistic generative model. The adversary in a semi-random model is not meant to model an actual natural process of data creation, but to encompass all possible bounded variations from an underlying probabilistic generative model.

There are similarities between some of the semi-adversarial models developed for various computational problems and the semi-adversarial models for unsupervised machine learning, but it is important to remark on the different uses of such models in the two settings. A computational problem usually has well-defined solutions, and the goal of analysis in semi-random models is to understand whether a polynomial-time algorithm can find an exact or an approximate solution; in an unsupervised machine learning task, one wants to find a latent structure defined in the generative model, and it is possible for the latent structure to be information-theoretically impossible to find if the noise and/or the adversarial model are too strong. When we talk about *approximation* in the context of solving optimization problems in semi-random models, we refer to how close is the *cost* of the solution found by the algorithm to the cost of an optimal solution; in an unsupervised machine learning task, the study of *approximation* refers to how close is the solution itself found by the algorithm to the ground truth.

1.1 Our Setting

We are interested in studying the *correlation clustering* problem in a semi-adversarial generative model, as an unsupervised machine learning problem. This is an important and well-studied problem about the analysis of dense Boolean matrices. In the correlation clustering problem, we have n data items, which we identify with the integers $\{1, \ldots, n\}$, and an unknown partition $C_1, \ldots, C_k$ of the items into clusters. We are given a symmetric $n \times n$ matrix $\widehat{M}$, where $\widehat{M}_{i,j} \in \{-1, +1\}$ represents a belief about items i and j being in the same cluster ($\widehat{M}_{i,j} = +1$ represents a belief that they are in the same cluster and $\widehat{M}_{i,j} = -1$ the opposite). The goal is to reconstruct the partition from $\widehat{M}$. A standard probabilistic generative model for correlation clustering is to start from a random equipartition $C_1, \ldots, C_k$ of $\{1, \ldots, n\}$, consider the "zero-error" matrix $M \in \{+1, -1\}^{n \times n}$ such that that $M_{i,j} = 1$ if and only if i and j belong to the same cluster, and obtain a matrix $\widehat{M}$ by applying random noise to M. A simple noise model is obtained by setting $\widehat{M}_{i,j} = M_{i,j}$ with probability $1/2 + \epsilon$, independently for every unordered pair $\{i, j\}$, and equal to $-M_{i,j}$ with probability $1/2 - \epsilon$, for a noise parameter $\epsilon > 0$. For constant k, this model exhibits a phase transition at $\epsilon \approx 1/\sqrt{n}$. When $\epsilon = o(1/\sqrt{n})$ it is impossible to reconstruct the partition, even in an approximate way, and when $\epsilon = \omega(1/\sqrt{n})$ it is possible to reconstruct the partition with at most $o(n)$ items being misclassified. In the regime in which reconstruction is possible, we are interested in introducing adversarial modifications in addition to random noise.

A possible semi-adversarial model, which was already studied by Makarychev et al. (2016), is to allow an adversary to modify a bounded number of entries of the matrix sampled from the probabilistic model. We refer to such a model as *post-adversarial* because the adversary operates after random choices have been made. For the correlation clustering problem, moreover, it is also natural to consider a semi-adversarial model that we call *pre-adversarial*, in which the adversary is allowed to modify the zero-error matrix in a bounded number of entries, and then random noise is applied to the matrix after these adversarial modifications. This model is somewhat in the spirit of smoothed analysis, in which noise is applied after an adversarial choice[1], but has never been studied before. In the regime $\epsilon > \omega(1/\sqrt{n})$, it is easy to see that a pre-adversary with a budget of modifying $\Omega(n^2)$ entries or a post-adversary with a budget of modifying $\Omega(\epsilon n^2)$ entries are able to force any algorithm to misclassify $\Omega(n)$ data points. Our goal is to understand whether spectral approaches can reconstruct the partition when the adversary has smaller budgets.

1.2 Our Contribution

All our results are asymptotic in n and assume a constant k number of clusters. They are also summarized in Table 1.

Optimal Pre-Adversarial Robustness of Spectral Algorithms. We show that a spectral algorithm can handle any pre-adversary making $o(n^2)$ changes, in the noise regime $\epsilon > \omega(1/\sqrt{n})$, giving polynomial-time reconstruction of the clustering with $o(n)$ misclassified items, with high probability.

Sub-Optimal and Yet Non-Trivial Post-Adversarial Robustness of Spectral Algorithms. In the post-adversarial setting, in the

[1] A notable difference is that our adversary has a limit to how many entries of the matrix M it can modify, while in smoothed analysis the first step is to select a completely adversarial instance.

Table 1: Reconstruction achieved by our Spectral Algorithm in the two semi-adversarial settings, compared to the Information-Theoretic Bounds.

Setting	**Inf.-Theor. Bounds**	**Spectral Algorithm**
Pre-Adversary	$B = o(n^2)$, $\epsilon = \omega(1/\sqrt{n})$	$B = o(n^2)$, $\epsilon = \omega(1/\sqrt{n})$
Post-Adversary	$B = o(\epsilon n^2)$, $\epsilon = \omega(1/\sqrt{n})$	$B = o(\epsilon^2 n^2)$, $\epsilon = \omega(1/\sqrt{n})$

noise regime $\epsilon > \omega(1/\sqrt{n})$, an analogous spectral algorithm can handle adversaries making $o(\epsilon^2 n^2)$ changes, delivering, as before, with high probability a reconstructed clustering with $o(n)$ misclassified items. This analysis is nearly tight, in that we can devise post-adversarial strategies with a budget of $O(\epsilon^2 n^2)$ changes which, for a wide range of values for ϵ, cause the spectral algorithm to misclassify $\Omega(n)$ items.

Optimal Post-Adversarial Robustness of SDP. Makarychev et al. (2016) already formulated an algorithm based on semidefinite programming (SDP) and showed that, in the post-adversarial setting, in the noise regime $\epsilon > \omega(1/\sqrt{n})$, the algorithm reconstructs in polynomial-time the correct clustering up to $o(n)$ misclassifications, with high probability, for all post-adversaries that have a budget of $o(\epsilon n^2)$ changes, matching an information-theoretic lower bound. We also provide an SDP-based algorithm with the same theoretical guarantees, but it is significantly different from the one in Makarychev et al. (2016).

In previous analyses of semi-adversarial settings, spectral algorithms usually performed poorly in the presence of adversaries (with some exceptions (Steinhardt, 2017)), so it is interesting that our pre-adversarial model provides an adversarial setting in which a spectral algorithm performs well all the way to information-theoretic limits. This is perhaps our main conceptual contribution.

We comment on an additional piece of intuition that comes out of our work. From previous work on semi-adversarial models, there is well-established evidence that spectral algorithms perform poorly on matrices that are very sparse, for example on adjacency matrices or Laplacian matrices of sparse random graphs modified by an adversary. The reason is that it is possible to change a small number of entries of a sparse matrix and create spurious large eigenvalues with localized eigenvectors, and doing so is a good adversarial strategy to make a spectral algorithm fail. In correlation clustering, the given matrix is dense, and so our analysis in the pre-adversarial setting can be seen as providing complementary intuition that spectral algorithms can be robust on dense random matrices. But where is the difference coming from between the optimal behaviour in the pre-adversarial setting and the sub-optimal behavior in the post-adversarial setting? We can think of the application of noise as the following process: each entry is left unchanged with probability 2ϵ, and it is replaced with a fresh random bit with probability $1 - 2\epsilon$. According to the above point of view, after the application of noise, there is only a sparse subset of ϵn^2 entries that give information about the clustering, while all the other entries give no information. So we can see that the pre-adversary operates on a dense matrix of entries that give information about the clustering, while the post-adversary operates, effectively, on a sparser one, explaining the sub-optimal robustness of spectral methods, and the existence of adversarial strategies to create localized eigenvectors with large eigenvalues.

1.3 Roadmap

In Section 2 we discuss relevant related work. In Section 3 we define the problem precisely and introduce the notation that we use. In Section 4, we present our spectral algorithm with its theoretical guarantees. We proceed by introducing our SDP-based algorithm in Section 5. Section 6 presents our lower bounds for reconstruction. Finally, in Section 7, we conclude by discussing the limitations of spectral approaches.

2 Related Work

Semi-Adversarial Models. Semi-Random (or, Semi-Adversarial) models have been the object of intense study in the recent past – see Roughgarden (2020) for a comprehensive introduction to the topic. The original motivation to go beyond the worst-case analysis of algorithms was to come up with fast algorithms that, with high probability over the random choice of the input, returned an (approximately) optimal solution, to avoid dealing with input substructures that make the problem hard but that might not be often found in practice. In fully random models, however, an algorithm is only required to solve instances coming from a given distribution. As a result, many optimal solutions to fully-random models overfit to the random model and are unlikely to behave well with real-world instances.

Researchers then introduced several *semi-random* models, *Smoothed Analysis* (Spielman and Teng, 2004) being perhaps the most famous exemplar. Several problems have been studied in this setting, e.g., planted clique (Feige and Krauthgamer, 2000) (whose optimal algorithm so far is based on a spectral algorithm), various of its generalizations, e.g., the Stochastic Block Model (Makarychev et al., 2016; Moitra et al., 2016), as well as Densest Subgraph (Bhaskara et al., 2010), Multi-Object Matching (Shi et al., 2020), and Correlation Clustering (Mathieu and Schudy, 2010). We will later say more on some of the above works, focusing on those most relevant to our work.

From a technical standpoint, the algorithmic strategies required for these semi-random models deviate significantly from the ones successfully applied to the fully-random and the smoothed analysis settings. In particular, purely spectral approaches (with no regularization) work in these settings but often fail when the adversary enters the picture (Makarychev et al., 2016; Moitra et al., 2016). In this paper, we observe similar behaviors of spectral and SDP-based methods. But, as pointed out, we also observe a certain unexpected robustness of the former which is only partial in the post-adversarial setting, but optimal in the pre-adversarial one. Possibly, the foremost difference between our work and most of the previous semi-random ones lies in its algorithmic goal: here, we are not trying to optimize an objective function over a semi-random instance – we are, rather, trying to reconstruct the unknown parameters (the unknown base clustering) of the semi-random model, given one (adversarially perturbed) sample from it. Our specific goal significantly changes the techniques employed and the overall algorithmic approach. In particular, in the context of rigorous machine learning, it has often been observed that the max-likelihood problem, when not enough samples are available, ends up with optimal solutions that are far from the unknown model parameters (see, e.g., Rubinstein and Vardi (2017)). That is, optimizing the max-likelihood objective does not, in general, return the hidden parameters of the model.

Correlation Clustering. Correlation clustering is a basic primitive in the machine learner's toolkit with applications ranging in several domains, including NLP (Van Gael and Zhu, 2007), social network analysis (Chen et al., 2012), and clustering aggregation (Gionis et al., 2007). Correlation Clustering was introduced by Bansal et al. (2004), who also presented several problems and approximation algorithms for it. Interestingly, they also considered the purely-random "seed reconstruction" version of the Correlation Clustering problem.

Clustering Reconstruction. The fully-random model closest to ours is the Stochastic Block Model. Given a *seed* partition of the nodes of a graph into clusters, the Stochastic Block Model samples a random graph as follows: a biased coin is flipped independently for each pair of nodes, using a different bias depending on whether the two nodes are in the same cluster or in different clusters. Any pair of nodes from the same cluster have a probability p of being connected by an edge; while any pair of nodes from different clusters have a probability $q < p$ of being joined by an edge[2]. The problem of reconstructing the seed partition starting from such a random graph has been studied extensively, especially in the bounded degree setting, and several spectral algorithms, as well as algorithms based on semidefinite programming and Grothendieck's inequality (Guédon and Vershynin, 2016), have been proposed for solving it.

Several *semi-adversarial* variants of the Stochastic Block Models have been studied by the community. Building on the work by Feige and Kilian (2001), Makarychev et al. (2016) (and, independently, Moitra et al. (2016)) gave algorithms to reconstruct the seed partition starting from a graph obtained by *monotone* modifications (plus a limited amount of adversarial ones) of a sample from the SBM. More precisely, in their semi-adversarial model, Nature first samples an SBM graph; then, an adversary can monotonically strengthen the random signal. Finally, the adversary has a limited budget for modifying the remaining edges. This setting includes also our *post-adversarial* setting. The algorithms to reconstruct the seed partition, similarly to those of Guédon and Vershynin (2016), are based on SDPs.

Mathieu and Schudy (2010) studied a different version of the semi-adversarial correlation clustering reconstruction problem: in their model, as in ours, one begins with a partition $C_1^*, \dots, C_k^*$ of the n nodes into clusters. Then, each pair of nodes gets *corrupted* i.i.d. with probability p: the adversary can then change the ± 1 label of each corrupted pair however it likes. They show that, if each original cluster has size $\Omega(\sqrt{n})$, and if $p \leq 1/3$, then an SDP based algorithm (together with a weighted version of the randomized rounding procedure of Ailon et al. (2008)) reconstructs the hidden clusters. An important difference between the error model of Mathieu and Schudy (2010) and ours is that the constraints that they put on their adversary are such that exact reconstruction is possible, while in both our pre-adversarial and post-adversarial settings our adversary is able to erase all information about a subset of vertices, and hence exact reconstruction is information-theoretically impossible. In their work, they also consider this semi-adversarial noise model

[2] One could see the correlation clustering distribution obtained by applying random noise to the zero-error matrix as an instance of the stochastic block model in which $q = 1/2 - \epsilon$ and $p = 1/2 + \epsilon$, and we interpret the presence/absence of an edge as a $+1/-1$. The stochastic block model, however, is typically analyzed in settings in which p and q are of the order of $1/n$ or $\log n/n$, leading to very sparse graphs.

from the point of view of approximation algorithms for the correlation clustering objective function.

Spectral Algorithms. Spectral algorithms have been extensively used for cluster reconstruction (Ng et al., 2001; Balakrishnan et al., 2011). Here, we restrict our discussion to works that apply spectral algorithms applied to (semi-)random models. In particular, Boppana (1987) introduced the spectral method for the fully-random graph bisection problem; McSherry (2001) and Coja-Oghlan (2006) improved the method to work for more general partitions, and to work with tighter gaps between the intra-cluster, and extra-cluster, probabilities.

Since spectral algorithms for clustering are often very efficient in terms of running time, especially when compared to more complex methods like the one based on semidefinite programming (Olsson et al., 2007), there has been interest in studying the robustness of spectral algorithms in several random and semi-adversarial settings (Ling, 2020; Stephan and Massoulié, 2019; Peche and Perchet, 2020; Abbe et al., 2020).

3 PRELIMINARIES

We study Correlation Clustering Reconstruction, defined as follows. We are given a complete graph of n points $\{1, \ldots, n\} =: [n]$ divided into k clusters, each of size n/k. For now, we assume n/k to be an integer. However, we will show that all our results still apply when the communities have size $n/k + o(n)$. If i, j belong to the same cluster the edge ij is labeled $+1$, otherwise the label is -1. In matrix notation, we are given a matrix M such that $M_{i,j} := +1$ if i, j are in the same cluster and -1 otherwise. The matrix M, which we call the *zero-error* matrix, is modified by random noise and adversarially, according to the following two processes. Let $0 \leq \epsilon \leq 1/2$ and $0 \leq B \leq n^2$. The quantity B is an integer and referred to as the *budget* of the adversary. We will assume that the adversarial changes are symmetric and the resulting matrix diagonalizable, since it would suffice $\leq B$ extra changes to achieve symmetry, and we are only interested in the asymptotic value of B.

Pre-Adversary. M is modified as follows. First, an adversary swaps the labels of B entries of M. The resulting matrix M' is then modified by random noise: every entry of M' is swapped with probability $1/2 - \epsilon$. The resulting matrix is denoted as M''.

Post-Adversary. Here, the process is inverted: first, we inject random noise and then let the adversary operate. First, every element of M is swapped with probability $1/2 - \epsilon$. The resulting matrix is M' (same notation, but the context will disambiguate). Second, an adversary swaps the sign of B elements in the matrix, giving rise to a matrix M''.

In both models, the Correlation Clustering Reconstruction problem is:

Given M'', reconstruct M as accurately as possible in polynomial-time.

This reconstruction goal is different from the usual optimization point of view. It is, however, of fundamental concern from the machine learning perspective. Notice that the post-adversarial setting is equivalent to the *model with outliers* from Makarychev et al. (2016). Observe that in the presence of such adversarial modifications it does not make sense to ask for Maximum Likelihood Estimation recovery of the latent clusters. Note also that asking for a high probability of perfect reconstruction is futile, for the adversary can swap the clusters of two nodes with only $B = 2n$ changes. Therefore, we focus on approximate reconstruction. Our goal is to find polynomial-time algorithms such that, with probability $1 - o(1)$, they correctly classify $n - o(n)$ vertices under the pre- and post-adversary. More precisely, let $\mathcal{P}^*$ be the $k-$partition of $[n]$ corresponding to the ground-truth clustering. We are required to output a partition $\mathcal{P}$ of the $[n]$ vertices into k non-empty sets to maximize the number of correctly classified vertices, which is defined as

$$\max_{\psi:\mathcal{P}^* \to \mathcal{P} \text{ bijective}} \sum_{S \in \mathcal{P}^*} |S \cap \psi(S)|.$$

Alternatively, we define the number of *misclassified vertices* by $\mathcal{P}$ as n minus the number of correctly classified vertices. Our goal is to correctly classify $n - o(n)$ vertices with high probability $1 - o(1)$.

3.1 The Technical Toolkit

We now describe our main technical toolkit, consisting of definitions and known facts about matrix norms, eigenvalues and eigenvectors, and concentration inequalities. The reader familiar with such background can safely skip directly to the next section.

We define $\boldsymbol{f}_i$ as the characteristic vector of the i^{th} cluster for a given zero-error matrix M: there are 1's in the positions corresponding to the elements of the ith cluster, and 0 everywhere else. We also define $\mathbf{1}$ as the vector having all coordinates equal to 1. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, the *euclidean* norms is defined as $\|\boldsymbol{x}\| := \sqrt{\sum_{i=1}^n \boldsymbol{x}_i^2}$. We also define the scalar product between two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x} \cdot \boldsymbol{y} := \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{y}_i$. Given a square matrix $M \in \mathbb{R}^{n,n}$, the *Frobenius* norm is defined as $\|M\|_F^2 := \sum_{i,j=1}^n M_{i,j}^2$. The *spectral*, or *Operator*, norm is defined as $\|M\|_{op} := \max_{\|\boldsymbol{x}\|=1} \|M\boldsymbol{x}\| = \max_{\|\boldsymbol{x}\|=\|\boldsymbol{y}\|=1} |\boldsymbol{x}^T A \boldsymbol{y}|$.

Our analyses study how eigenvectors are affected by perturbations of the matrix. The following result is eminently useful in this regard.

Theorem 3.1 (Davis-Kahan-Wedin)**.** *Let M, N be symmetric matrices in $\mathbb{R}^{n,n}$ such that M has eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ with corresponding orthogonal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, while N has eigenvalues $\lambda'_1 \geq \ldots \geq \lambda'_n$ with corresponding orthogonal eigenvectors $\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_n$. Let $k \leq n$, and let $V_M \in \mathbb{R}^{n,k}$ having $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ as columns, $V_N \in \mathbb{R}^{n,k}$ having $\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_k$ as columns. Also, suppose $\delta_k := \lambda_k - \lambda_{k+1} > 0$. Then, $\|V_M V_M^T - V_N V_N^T\|_F \leq \frac{2\sqrt{k}\cdot\|N-M\|_{op}}{\delta_k}$.*

Let us now recall some concentration inequalities.

Theorem 3.2 (Chernoff–Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be a sequence of scalar random variables with $X_i \in [a_i, b_i]\ \forall\ i \in [n]$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $\lambda > 0$, $\Pr\left(|\overline{X} - \mathbb{E}[\overline{X}]| \geq \lambda\right) \leq 2 \cdot \exp\left(-\frac{2\lambda^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$.*

Theorem 3.3 (Azuma's Inequality)**.** *Let $X_1, \ldots, X_n$ be a sequence of scalar random variables with $|X_i| \leq c_i > 0$ almost surely. Assume also that we have the martingale difference property $\mathbb{E}[X_i | X_1, \ldots, X_{i-1}] = 0$ almost surely for all $1 \leq i \leq n$. Let $S_n = \sum_{i=1}^n X_i$ and $\gamma := \sqrt{\sum_{i=1}^n c_i^2}$. Then, for any $\lambda > 0$, S_n obeys the large deviation inequality $\Pr(|S_n| \geq \lambda) \leq 2 \cdot \exp(-2\lambda^2/\gamma^2)$.*

We also make use of the well-known Power-Method to compute eigenvectors and eigenvalues of symmetric matrices (see, e.g., Golub and Van Loan (1996)). For the sake of the exposition, we ignore the small errors of these solutions since they are easily absorbed by other types of errors we deal with in our derivations.

4 THE SPECTRAL ALGORITHM

Here we present a spectral algorithm, which will be used both in the pre-adversarial and the post-adversarial setting. We also explain our proof strategy and state the theoretical guarantees, with full proofs deferred to the Appendix.

4.1 Proof Strategy

Technically, the known average-case analyses of spectral algorithms for clustering problems depend on bounding the spectral norm of the difference between the empirical matrix that we are given and the average matrix, using concentration results for random matrices; we are able to extend this analysis to the semi-adversarial setting by bounding the spectral norm of the difference of the matrix before and after the intervention of the adversary, which is easy to do by using Frobenius norm as an intermediate step.

More precisely, our spectral results are based on bounding the spectral norm of the changes caused by the adversary. An adversary that makes up to B changes to a matrix that has ± 1 entries can make changes whose spectral norm is at most $O(\sqrt{B})$. If such changes are made by a pre-adversary, the spectral norm of the changes after the application of the random noise is $O(\epsilon\sqrt{B})$, and the spectral algorithm works well provided that this is much smaller than ϵn, which is true if $B = o(n^2)$. If the changes are made by a post-adversary, then we need $O(\sqrt{B})$ to be much smaller than ϵn, and so we need the condition $B = o(\epsilon^2 n^2)$. By Theorem 3.1, such a bound on the operator norm of the difference between the zero-error matrix M and our input matrix M'' reflects upon our ability to approximately recover the main eigenvectors of M, and so the whole clustering.

4.2 Properties Of The Zero-Error Matrix

The zero-error matrix M has rank k for $k > 2$, and rank 1 for $k = 2$. We now describe its spectrum. First, 0 is an eigenvalue for M whose eigenspace, for $k > 2$, has dimension $n - k$ by the Dimension Theorem for vector spaces (it is described by a homogeneous equation whose associated matrix, M, has rank k), and has dimension $n - 1$ for $k = 2$. Second, $2 \cdot n/k$ is also an eigenvalue, and its eigenspace has dimension $k - 1$. A basis for the eigenspace of $2 \cdot n/k$ is $\{\boldsymbol{f}_i - \boldsymbol{f}_{i+1},\ i \in [k-1]\}$. If $k > 2$, we also have another eigenvalue, which is negative: $(2/k - 1)n$, whose eigenspace has dimension 1 and is spanned by the eigenvector $\mathbf{1}$ with all identical coordinates. There are no more eigenvalues, since the vector space $\mathbb{R}^n$ is the direct sum of these eigenspaces.

We show that, to reconstruct the latent clustering, it suffices to recover a good approximation of an orthogonal basis of eigenvectors for the largest eigenvalue of M. This is what we do in our spectral algorithm.

4.3 Our Algorithm

In Algorithm Spectral, we first obtain the eigenvectors of the $k - 1$ leading eigenvalues of the perturbed matrix M''. We do not know the number of clusters k, but we pick k as the smallest integer such that the $(k - 1)-$th largest eigenvalue is larger than 2 times the $k-$th largest eigenvalue in absolute value. After that, we use the eigenvalues to retrieve the cluster each element belongs to in procedure Get-Clusters: for each $i \in [n]$, this procedure computes a tentative cluster $\mathcal{S}_i$ for it, which contains all the indices whose corresponding elements in every generated eigenvector are close to the i^{th} element. We show that most of these tentative clusters are correct, meaning that they approximately reconstruct the cluster the element belongs to. Therefore, we can sample $k-1$ distinct approximate clusters and build the k^{th} cluster with the remaining elements. With high probability, this procedure returns k almost correct clusters.

Algorithm 1 SPECTRAL(M'', n, t). Input: perturbed matrix M'', input size n, separating threshold t.

```
$\mathcal{S} \leftarrow \emptyset$
$\mathcal{U} \leftarrow [n]$
Let $\{\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_{k-1}, \boldsymbol{v}''_k\}$ be the $k$ eigenvectors of the largest eigenvalues $\lambda''_1 \geq \ldots \geq \lambda''_k$ of $M$ obtained through POWER-METHOD, where $k$ is the smallest integer such that $\frac{|\lambda''_{k-1}|}{|\lambda''_k|} > 2$.
$\{\mathcal{S}_1, ..., \mathcal{S}_n\} \leftarrow$ GET-CLUSTERS($\{\boldsymbol{v}''_1, ..., \boldsymbol{v}''_{k-1}\}, t$)
for $\ell = 1, \ldots, k-1$ do
    repeat
        Sample $i \in \mathcal{U}$ U.A.R.
    until $|\mathcal{S}_i \cap \mathcal{U}| \geq \frac{n}{2k}$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{S}_i \cap \mathcal{U}\}$
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}_i$
$\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{U}\}$
return $\mathcal{S}$
```

Algorithm 2 GET-CLUSTERS($\{\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_{k-1}\}, t$). Input: orthogonal unitary vectors $\{\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_{k-1}\}$ and separating threshold $t > 0$.

```
for $i = 1, \ldots, n$ do
    for $j = 1, \ldots, n$ do
        $\mathcal{S}_{ij} \leftarrow \emptyset$
        for $h = 1, \ldots, k-1$ do
            if $|(\boldsymbol{v}''_h)_i - (\boldsymbol{v}''_h)_j| > t$ then
                $\mathcal{S}_{ij} \leftarrow \mathcal{S}_{ij} \cup \{h\}$
    $\mathcal{S}_i \leftarrow \{j \in [n] : \mathcal{S}_{i,j} = \emptyset\}$
return $\{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$
```

4.4 Algorithm Spectral Can Cope Optimally With The Pre-adversary

Let

$$\epsilon = \omega(n^{-1/2}) \quad \text{and} \quad B = o(n^2). \tag{1}$$

Recall that in this regime first the pre-adversary swaps B entries of the original matrix M, giving rise to a matrix M'. Then, M' is modified by random noise by swapping every entry with probability $1/2 - \epsilon$. The resulting matrix is denoted as M''. As we show in Section 6, if $\epsilon = o(n^{-1/2})$ or $B = \Omega(n^2)$ information-theoretic lower bounds kick in and no reconstruction is possible.

We are able to show the following, which implies the optimality of the SPECTRAL Algorithm in the whole feasible noise regime.

Theorem 4.1. *With high probability* $1 - o(1)$*, Algorithm* SPECTRAL(M'', n, t) *with* $t = \frac{1}{2\sqrt{2n}}$ *outputs* k *clusters with* $o(n)$ *misclassified vertices, where the number of misclassified vertices is at most* $\frac{2048k^5}{\epsilon^2} + \frac{128k^5B}{n} = o(n)$*. Moreover, with high probability* $1 - o(1)$*, the running time is* $\tilde{O}(n^2)$*.*

The full proof is in the Appendix.

4.5 Sub-Optimal Robustness Of Algorithm Spectral With The Post-adversary

Let

$$\epsilon = \omega(n^{-1/2}) \quad \text{and} \quad B = o(\epsilon^2 n^2). \tag{2}$$

Similar to the pre-adversarial setting, we can show sub-optimal yet non-trivial robustness of Algorithm SPECTRAL for these parameters.

Theorem 4.2. *With high probability* $1 - o(1)$*, Algorithm* SPECTRAL(M'', n, t) *with* $t = \frac{1}{2\sqrt{2n}}$ *outputs* k *clusters with* $o(n)$ *misclassified vertices, where the number of misclassified vertices is at most* $\frac{2048k^5}{\epsilon^2} + \frac{32k^5B}{\epsilon^2 n} = o(n)$*. Moreover, with high probability* $1 - o(1)$*, the running time is* $\tilde{O}(n^2)$*.*

The proof is in the Appendix. As we show in Section 7, as far as B is concerned, this analysis is nearly tight asymptotically for a wide range of values of ϵ. Indeed, when $\epsilon = \Omega((\log(n)/n)^{1/3})$, by modifying $B = \Theta(\epsilon^2 n^2)$ entries, the post-adversary can induce with high-probability a principal eigenvector that gives no information on the latent clustering. This makes spectral algorithms like Algorithm SPECTRAL fail and creates a gap between the information-theoretic threshold, which is $B = o(\epsilon n^2)$ (see Section 6), and the reach of spectral methods. We believe that the same techniques can be used to prove that the same impossibility result holds for any $\epsilon = \omega(n^{-1/2})$, but we have not been able to prove it in this work.

4.6 Going Beyond Equinumerous Clusters

We also show that our spectral algorithm and its theoretical guarantees still hold when all the communities have size $n/k + o(n)$. In this setting, the clustering differs from an equinumerous clustering only by $o(n)$ elements. Therefore, the zero-error matrix M has only $o(n^2)$ different elements from a zero-error matrix associated to an equinumerous clustering, and can be seen as a derivation of this last matrix under the action of a pre-adversary changing $o(n^2)$ entries. We have already shown that our spectral algorithm can handle such a pre-adversary effectively, so it can deal with nearly equinumerous clusters. More details can be found in the Appendix.

5 OPTIMAL ROBUSTNESS WITH SDP

Consider the post-adversarial setting with parameters

$$\epsilon = \omega(n^{-1/2}) \quad \text{and} \quad B = o(\epsilon n^2). \tag{3}$$

Makarychev et al. (2016) formulated a polynomial-time algorithm based on semidefinite programming (SDP) in their model with outliers, which is equivalent to our post-adversary, and showed that, with high probability, the algorithm reconstructs the correct clustering up to $o(n)$ misclassified vertices, matching the information-theoretic lower bound. We also provide an SDP-based algorithm with the same guarantees, but it is significantly different from the one by Makarychev et al. (2016).

5.1 Overview Of Our Algorithm

We present a novel optimal algorithm for correlation clustering reconstruction in the post-adversarial setting. Differently from what has been done for the spectral algorithm, here we assume to have access to the number of clusters k. The assumption of knowing the number of clusters in advance, despite being an additional assumption with respect to our spectral algorithm, is also present in related works to ours, as in the context of Angular Synchronization and Group Synchronization (Bandeira et al., 2017; Shi et al., 2020).

At first, our algorithm transforms the input matrix M'' into a new matrix Q that is close to a positive semidefinite matrix. Moreover, by using Q, we can move the vector $\mathbf{1}$ out of the equation because it is no longer a leading eigenvector.

Our algorithm, named Recursive-Clust, iteratively solves SDPs, with Q or one of its submatrices as the coefficient matrix, to extract a good eigenvector from the solution matrix, then uses the eigenvector to partition the set of vertices in two, and is applied recursively on each of the two subsets. The eigenvector is picked by randomly sampling an eigenvector from an orthogonal basis of eigenvectors for the solution matrix of the SDP, where each eigenvector is picked with probability proportional to its eigenvalue.

We show that, with high probability, this procedure approximately retrieves a basis of the eigenspace of the largest eigenvalue of the zero-error matrix M, allowing to reconstruct the ground-truth clustering with $o(n)$ misclassified vertices. As for the running time, this is dominated by the time needed to solve all the semidefinite programs. Since there are at most $k = O(1)$ SDPs to solve, by using known algorithms to this end (Jiang et al., 2020), we finally get the following result.

Theorem 5.1. *With high probability $1 - o(1)$, our SDP-based Algorithm outputs k clusters with $o(n)$ misclassified vertices. Moreover, with high probability $1 - o(1)$, the running time is $\tilde{O}(n^6)$.*

5.2 Going Beyond Equinumerous Clusters

We also show that our SDP-based algorithm and its theoretical guarantees still hold when all the communities have size $n/k + o(n)$. This result can be directly derived from the one for equinumerous clusters, since the errors deriving from $o(n)$ changes to an equinumerous clustering are absorbed by the post-adversarial action. More details are in the Appendix.

5.3 Spectral Algorithms Versus SDP

It is well-known (Olsson et al., 2007) that SDP approaches have a high computational cost which scales very poorly with size. By and large, this makes them interesting only at a theoretical level (for now at least). In contrast, spectral algorithms are much more efficient, scalable, and extensively used in practice.

In this work, we have also quantified the discrepancy in running time between our spectral algorithm ($\tilde{O}(n^2)$) and our SDP-based algorithm ($\tilde{O}(n^6)$). Nothing changes if we consider the optimal SDP-based algorithm from Makarychev et al. (2016). This gives further evidence about why understanding the strengths and limitations of spectral approaches in a rigorous way is so important, even when semidefinite programming allows to gain more robustness.

6 INFORMATION-THEORETIC LOWER BOUNDS

This section is dedicated to the analysis of the information-theoretic lower bounds for our semi-adversarial settings. We remark that we are not interested in recovering the exact thresholds of efficient solvability of the problems, but only in the asymptotic ones.

In the special case $B = 0$, the pre-adversarial model and the post-adversarial model are the same, and they are equal to the well-known random model (see Abbe (2017) for a comprehensive survey). By the results of Mossel et al. (2014) approximate reconstruction is *not* solvable information-theoretically for $k = 2$ clusters if $\epsilon \leq \frac{1}{\sqrt{2n}}$. If $k \geq 2$ is a constant, Banks et al. (2016) showed that approximate reconstruction is not possible in an information-theoretic setting if $\epsilon \leq \frac{c}{\sqrt{n}}$, where c is a constant eventually depending on k. Therefore, we cannot hope to solve our problem if $\epsilon = o(n^{-1/2})$, or if $\epsilon \leq \frac{c}{\sqrt{n}}$. Since we are not interested in recovering the exact thresholds of efficient solvability of the problems, but only asymptotic ones, we focus only on $\epsilon = \omega(n^{-1/2})$. Now, we focus on the budget parameter B in the two different semi-adversarial models.

Pre-Adversarial Model. We have seen that $\epsilon = \omega(n^{-1/2})$. We prove that if $B \geq \Omega(n^2)$, then the adversary could change the clusters of a *constant* fraction of all the nodes, therefore making it impossible to approximately reconstruct the clusters. More precisely, for each constant $\delta > 0$, we can take $B \leq \delta \cdot n^2$ and the adversary could completely randomize the matrix entries for $\delta \cdot n = \Theta(n)$ vertices, making approximate reconstruction impossible. For this reason, we can only consider $B = o(n^2)$.

Post-Adversarial Model. We have seen that $\epsilon = \omega(n^{-1/2})$. Makarychev et al. (2016) proved that, if $B \geq \Omega(\epsilon n^2)$, then the adversary could make it impossible to approximately reconstruct the clusters. It follows directly from their reconstruction lower bound (with $a = n(1/2 + \epsilon)$ and $b = n(1/2 - \epsilon)$). The intuitive explanation is that the random perturbation is equivalent to changing independently each element of the zero-error matrix into random $\in \{\pm 1\}$ with probability $1 - 2\epsilon$, or leaving as it is with probability 2ϵ. Therefore, by Lemma 3.2, this means that, with high probability, only $\Theta(\epsilon n^2)$ entries preserve the original information. Therefore, for a post-adversary, it would suffice to change those $\Theta(\epsilon n^2)$ entries of M' into random to disrupt the original information.

7 LIMITATIONS OF SPECTRAL APPROACHES

We show that while spectral methods can withstand pre-adversaries, they falter against post-adversaries. Consider the setting with 2 equinumerous clusters, and recall that M' is the resulting matrix after the random noise is injected. Let $\epsilon = \Omega(\log(n)^{1/3} n^{-1/3})$. We show that if the post-adversary can modify $B = \Theta(\epsilon^2 n^2)$ entries of M', it can create a spurious large eigenvalue whose corresponding eigenvector carries no information about the original clusters. Here is a post-adversarial strategy that achieves this.

Take a set $\mathcal{S}$ of $4\epsilon n$ vertices with $2\epsilon n$ from each cluster and consider the induced minor in M'. Change all elements in the $4\epsilon n \times 4\epsilon n$ minor to 1. By Lemma 3.2, with probability $\geq 1 - e^{\Theta(\epsilon^2 n)} = 1 - o(1)$, these are $8\epsilon^2 n^2 \leq B \leq 16\epsilon^2 n^2$ many changes. Consider now the set of columns of the elements in $\mathcal{S}$. This contains $n - 4\epsilon n$ sub-rows of elements outside $\mathcal{S}$, each with $4\epsilon n$ elements. By Lemma 3.3, the absolute value of the sum of the elements in each one of this sub-rows is $\leq 2\sqrt{\epsilon n \log(n)}$ with probability $\geq 1 - \frac{2}{n^2}$. Therefore, by the union bound, with probability $1 - \frac{2}{n} = 1 - o(1)$, for each sub-row we can change $\leq 2\sqrt{\epsilon n \log(n)}$ elements in such a way as to ensure that the sum of the elements is 0. If we do this for each sub-row, we just need $\leq 2n\sqrt{\epsilon n \log(n)}$ additional changes. These adversarial changes turn the matrix M' into the final matrix M''. After these changes, consider the vector $\boldsymbol{v}_{\mathcal{S}}$ having $\frac{1}{\sqrt{\epsilon n}}$ for indices of elements in $\mathcal{S}$ and 0 everywhere else. Because of our changes, we have that

$$M'' \boldsymbol{v}_{\mathcal{S}} = 4\epsilon n \cdot \boldsymbol{v}_{\mathcal{S}}.$$

Thus, $\boldsymbol{v}_{\mathcal{S}}$ is an eigenvector for M'' with eigenvalue $4\epsilon n$. The total number of changes, with high probability, has been

$$8\epsilon^2 n^2 \leq B \leq 16\epsilon^2 n^2 + 2\epsilon^{1/2} n^{3/2} \log^{1/2}(n) = O(\epsilon^2 n^2)$$

because, by hypothesis, $\epsilon = \Omega(\log(n)^{1/3} n^{-1/3})$. Therefore, with only $B = \Theta(\epsilon^2 n^2)$ post-adversarial changes, we can create an eigenvector $\boldsymbol{v}_{\mathcal{S}}$ with eigenvalue $\Theta(\epsilon n)$. Now, by what we prove in the Appendix, $\|M''\|_{op}= \Theta(\epsilon n)$, so our new eigenvalues is asymptotically of the same order of magnitude of the largest eigenvalue of M''. This vector could become the eigenvector of the largest eigenvalue of M'', even if it does not tell anything about the original clustering, making a simple spectral approach fail. Notice that $\epsilon^2 n^2 = o(\epsilon n^2)$, so the number of changes is within the information-theoretic feasibility range.

8 Conclusion

In this work, we have proposed a new semi-adversarial framework for the analysis of algorithms and applied it to the important case of spectral algorithms for correlation clustering. In our opinion, the main takeaway point is that the framework has allowed us to reveal a certain robustness of spectral algorithms which was unknown before. We hope that this result can spur follow-up work in the same spirit, to elucidate the robustness properties of machine learning algorithms for a variety of problems.

Acknowledgements

F. Chierichetti, A. Panconesi, and G. Re were supported by a Google Focused Research Award, by the PRIN project 2017K7XPAN, by BiCi — Bertinoro international Center for informatics, and by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University. L. Trevisan was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 834861).

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Abbe, E., Boix-Adsera, E., Ralli, P., and Sandon, C. (2020). Graph powering and spectral robustness. *SIAM Journal on Mathematics of Data Science*, 2(1):132–157.

Ailon, N., Charikar, M., and Newman, A. (2008). Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5).

Balakrishnan, S., Xu, M., Krishnamurthy, A., and Singh, A. (2011). Noise thresholds for spectral clustering. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Bandeira, A. S., Boumal, N., and Singer, A. (2017). Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 163(1-2):145–167.

Banks, J., Moore, C., Neeman, J., and Netrapalli, P. (2016). Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416. PMLR.

Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. *Machine Learning*, 56(1):89–113.

Belitskii, G. et al. (2013). *Matrix norms and their applications*, volume 36. Birkhäuser.

Bhaskara, A., Charikar, M., Chlamtac, E., Feige, U., and Vijayaraghavan, A. (2010). Detecting high log-densities: An $O(n^{1/4})$ approximation for densest k-subgraph. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 201–210, New York, NY, USA. Association for Computing Machinery.

Boppana, R. B. (1987). Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, pages 280–285.

Chen, Y., Sanghavi, S., and Xu, H. (2012). Clustering sparse graphs. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Coja-Oghlan, A. (2006). A spectral heuristic for bisecting random graphs. *Random Structures & Algorithms*, 29(3):351–398.

Feige, U. and Kilian, J. (2001). Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671.

Feige, U. and Krauthgamer, R. (2000). Finding and certifying a large hidden clique in a semirandom graph. *Random Structures & Algorithms*, 16(2):195–208.

Gionis, A., Mannila, H., and Tsaparas, P. (2007). Clustering aggregation. *ACM Trans. Knowl. Discov. Data*, 1(1):4–es.

Golub, G. H. and Van Loan, C. F. (1996). Matrix computations. johns hopkins studies in the mathematical sciences.

Guédon, O. and Vershynin, R. (2016). Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3):1025–1049.

Jiang, H., Kathuria, T., Lee, Y. T., Padmanabhan, S., and Song, Z. (2020). A faster interior point method for semidefinite programming. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 910–918. IEEE.

Krivine, J.-L. (1978). Constantes de grothendieck et fonctions de type positif sur les spheres. *Séminaire Analyse fonctionnelle (dit" Maurey-Schwartz")*, pages 1–17.

Ling, S. (2020). Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods. *arXiv preprint arXiv:2008.05341*.

Makarychev, K., Makarychev, Y., and Vijayaraghavan, A. (2016). Learning communities in the presence of errors. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1258–1291, Columbia University, New York, New York, USA. PMLR.

Mathieu, C. and Schudy, W. (2010). Correlation clustering with noisy input. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 712–728. SIAM.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE.

Moitra, A., Perry, W., and Wein, A. S. (2016). How robust are reconstruction thresholds for community detection? In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 828–841, New York, NY, USA. Association for Computing Machinery.

Mossel, E., Neeman, J., and Sly, A. (2014). Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 3(5).

Nesterov, Y. and Nemirovskii, A. (1990). Optimization over positive semidefinite matrices: Mathematical background and user's manual. *USSR Acad. Sci. Centr. Econ. & Math. Inst*, 32.

Nesterov, Y. and Nemirovsky, A. (1988). A general approach to polynomial-time algorithms design for convex programming. *Report, Central Economical and Mathematical Institute, USSR Academy of Sciences, Moscow*.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA. MIT Press.

Olsson, C., Eriksson, A. P., and Kahl, F. (2007). Solving large scale binary quadratic problems: Spectral methods vs. semidefinite programming. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Peche, S. and Perchet, V. (2020). Robustness of community detection to random geometric perturbations. *Advances in Neural Information Processing Systems*, 33.

Roughgarden, T. (2020). Introduction. In Roughgarden, T., editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 1–24. Cambridge University Press.

Rubinstein, A. and Vardi, S. (2017). Sorting from noisier samples. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, page 960–972, USA. Society for Industrial and Applied Mathematics.

Shi, Y., Li, S., and Lerman, G. (2020). Robust multi-object matching via iterative reweighting of the graph connection laplacian. *Advances in Neural Information Processing Systems*, 33.

Spielman, D. A. and Teng, S.-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463.

Steinhardt, J. (2017). Does robustness imply tractability? A lower bound for planted clique in the semirandom model. *CoRR*, abs/1704.05120.

Stephan, L. and Massoulié, L. (2019). Robustness of spectral methods for community detection. In *Conference on Learning Theory*, pages 2831–2860. PMLR.

Tao, T. (2012). *Topics in random matrix theory*, volume 132. American Mathematical Soc.

Van Gael, J. and Zhu, X. (2007). Correlation clustering for crosslingual link detection. In *IJCAI*, pages 1744–1749.

Vandenberghe, L. and Boyd, S. (1996). Semidefinite programming. *SIAM review*, 38(1):49–95.

Appendix

A EXTENDED PRELIMINARIES

Here we complement our preliminary section with the additional results needed for the remaining proofs.

Given a square matrix $M \in \mathbb{R}^{n,n}$, the ℓ_∞-to-ℓ_1 operator norm, is defined as,

$$\|M\|_{\infty\to 1}:= \max_{\boldsymbol{x},\boldsymbol{y}\in\{\pm 1\}^n} |\boldsymbol{x}^T M\boldsymbol{y}| = \max_{\|\boldsymbol{x}\|_\infty\le 1, \|\boldsymbol{y}\|_\infty \le 1} \boldsymbol{x}^T M \boldsymbol{y} = \max_{\|\boldsymbol{x}\|_\infty=1} |\boldsymbol{x}^T M\boldsymbol{x}| = \max_{\|\boldsymbol{x}\|_\infty=1} \|M\boldsymbol{x}\|_1.$$

Finally, the *SDP-norm*:

$$\|M\|_{SDP}:= \max_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n,\boldsymbol{y}_1,\dots,\boldsymbol{y}_n \\ \|\boldsymbol{x}_h\|=\|\boldsymbol{y}_k\|=1 \ \forall\ h,k\in[n]}} \sum_{i,j=1}^n M_{ij}\langle \boldsymbol{x}_i, \boldsymbol{y}_j\rangle = \max_{\substack{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n,\boldsymbol{y}_1,\dots,\boldsymbol{y}_n \\ \|\boldsymbol{x}_h\|\le 1,\|\boldsymbol{y}_k\|\le 1 \ \forall\ h,k\in[n]}} \sum_{i,j=1}^n M_{ij}\langle \boldsymbol{x}_i, \boldsymbol{y}_j\rangle.$$

Let us also recall some known relationships and inequalities about these norms. (see Belitskii et al. (2013) for the proofs)

Lemma A.1. *If M is an $n\times n$ real matrix with rank r, then $\|M\|_{op}^2 \le \|M\|_F^2 \le r\cdot \|M\|_{op}^2$*

Like the operator norm, the ℓ_∞-to-ℓ_1 norm is monotone with respect to inclusion.

Lemma A.2. *Let $A\in\mathbb{R}^{n,n}$, and let $B\subseteq A$ be a square sub-matrix of A. Then, $\|B\|_{\infty\to 1}\le \|A\|_{\infty\to 1}$.*

Lemma A.3. *Let $M\in\mathbb{R}^{n,n}$. Then, $\|M\|_{\infty\to 1}\le n\cdot\|M\|_{op}$.*

Theorem A.4 (Grothendieck's Inequality)**.** *There exists a constant $c\le 1.8$ such that, for every matrix $M\in\mathbb{R}^{n,n}$, it holds*

$$\|M\|_{\infty\to 1}\le \|M\|_{SDP}\le c\cdot\|M\|_{\infty\to 1}.$$

From Krivine (1978), $c \le {}^{\pi}/_{2\ln(1+\sqrt{2})} \simeq 1.782$. We also make use of the following known facts about eigenvalues.

Lemma A.5 (Weyl's Inequality)**.** *Let M, N be symmetric matrices in $\mathbb{R}^{n,n}$ with eigenvalues respectively $\mu_1\ge\ldots\ge\mu_n$ and $\nu_1\ge\ldots\ge\nu_n$. Let $\lambda_1\ge\ldots\ge\lambda_n$ be the eigenvalues of $M+N$. Then,*

$$\mu_k+\nu_n\le\lambda_k\le\mu_k+\nu_1 \ \forall\ 1\le k\le n.$$

Corollary A.5.1. *Let M, E be symmetric matrices in $\mathbb{R}^{n,n}$, where M has eigenvalues $\lambda_1\ge\ldots\ge\lambda_n$ and $M+E$ has eigenvalues $\lambda'_1\ge\ldots\ge\lambda'_n$. Then,*

$$|\lambda_k-\lambda'_k|\le\|E\|_{op} \ \forall\ 1\le k\le n.$$

Let us now recall some concentration inequalities.

Theorem A.6 (Markov's Inequality)**.** *Let X be a positive random variable with finite expectation. Then, for any $a>0$, it holds*

$$\Pr(X\ge a)\le\frac{\mathbb{E}[X]}{a}.$$

Theorem A.7. *Let M be a $n\times n$ real random matrix whose entries $\{M_{i,j}\}$ are independent, have all expected value 0 ($\mathbb{E}[M_{i,j}]=0\ \forall\ i,j$) and are uniformly bounded in magnitude by 1 ($|M_{i,j}|\le 1\ \forall\ i,j$). Then, for every $A\ge 4$,*

$$\Pr(\|M\|_{op}\ge A\sqrt{n})\le 2^{-An}.$$

A proof of this theorem can be found in Tao (2012).

Finally, let us recall some known facts about semidefinite programming (see Vandenberghe and Boyd (1996) for a comprehensive introduction) and the computation of eigenvectors. Concerning the former, Nesterov and Nemirovsky (1988); Nesterov and Nemirovskii (1990) showed that interior-point methods can efficiently solve semidefinite programs (SDP's), and several other methods have been developed. More precisely, there exists an algorithm (referred to in this paper as SDP-Solver) that computes the optimal solution up to an exponentially small error in the size of the input within polynomial time in the size of the input (Jiang et al., 2020). In what follows, for clarity of the exposition, we sometimes assume the solution of SDP's to be exact. This can be done w.l.o.g. since the exponentially small errors of the SDP solution are absorbed by other types of error we control in our derivations. Similarly, we make use of the well-known Power-Method to compute eigenvectors and eigenvalues of symmetric matrices (see, e.g., Golub and Van Loan (1996)). Again, for the sake of the exposition, we ignore the small errors of these solutions since they are easily absorbed by other types of errors we deal with in our derivations. Recall that the running time of the Power-Method is polynomial in the size of the input and the ratio between the largest eigenvalue and the spectral gap.

A.1 Properties of the Zero-Error Matrix

Here we provide more details about the claims on the zero-error matrix M. We will analyze its ranks, its spectrum and provide an orthogonal basis of eigenvectors for it.

The zero-error matrix M has rank k for $k > 2$, and rank 1 for $k = 2$. Recall the definition of $\boldsymbol{f}_i$, the characteristic vector of the i^{th} cluster: there are 1's in the positions corresponding to the elements of the cluster, and 0 everywhere else. For $k > 2$, the rows of M are spanned by the vectors $\{\boldsymbol{f}_i\}_{1\leq i\leq k}$ and are linearly independent, as it can be shown by induction using the Gaussian elimination. For $k = 2$, the rows of M are spanned by the vector $\boldsymbol{f}_1 - \boldsymbol{f}_2$. Let us now look at the spectrum of M.

First, 0 is an eigenvalue for M whose eigenspace has dimension $n-k$ for $k > 2$ and $n-1$ for $k = 2$ by the Dimension Theorem for vector spaces (it is described by a homogeneous equation whose associated matrix, M, has rank k for $k > 2$ and rank 1 for $k = 2$). Second, $2 \cdot {}^{n}\!/_{k}$ is also an eigenvalue whose eigenspace has dimension $k-1$. A basis for it is $\{\boldsymbol{f}_i - \boldsymbol{f}_{i+1},\ 1 \leq i \leq n-1\}$. If $k > 2$, we also have another eigenvalue: $-{}^{(k-2)}\!/_{k} \cdot n = ({}^{2}\!/_{k} - 1)n$, whose eigenspace has dimension 1 and is spanned by the eigenvector with all identical coordinates. There are no more eigenvectors, since the vector space $\mathbb{R}^n$ is the direct sum of these eigenspaces.

It is useful to find an orthogonal basis for the eigenspace of $2 \cdot {}^{n}\!/_{k}$. With the Grahm-Schmidt orthogonalization procedure, we can get an orthogonal basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ where:

$$\boldsymbol{v}_i := \frac{1}{\sqrt{n/k}} \left(\frac{1}{\sqrt{i^2+i}} \sum_{j=1}^{i} \boldsymbol{f}_j - \frac{i}{\sqrt{i^2+i}} \boldsymbol{f}_{i+1} \right) \quad \forall\ i \in [k-1]. \tag{4}$$

Equation 4 can be shown by induction. These vectors are mutually orthogonal, have identical coordinates for vertices in the same cluster and their coordinates sum to 0. For instance, for $n = k = 3$, we get,

$$\boldsymbol{v}_1 = \frac{1}{\sqrt{2}}(\boldsymbol{f}_1 - \boldsymbol{f}_2);\ \boldsymbol{v}_2 = \frac{1}{\sqrt{6}}(\boldsymbol{f}_1 + \boldsymbol{f}_2) - \frac{2}{\sqrt{6}}\boldsymbol{f}_3.$$

Notice that, for any k, any vector of the orthogonal basis detects at least one cluster. Moreover, any orthogonal vector in this subspace detects a bisection into disjoint clusters. We exploit this to reconstruct the original clusters iteratively.

Lemma A.8. *Let $\boldsymbol{v}_i$ be as in Equation 4, for $i \in [k-1]$. And let $\boldsymbol{x} := \sum_{i=1}^{k-1} \lambda_i \boldsymbol{v}_i$, where $\|x\| = \sum_{i=1}^{k-1} \lambda_i^2 = 1$. Then, there exists $i \neq j \in [k]$ such that, if x^i is the coordinate of $\boldsymbol{x}$ along the vertices of the i^{th} cluster, it holds*

$$|x^i - x^j| > \frac{1}{k \cdot \sqrt{n}}.$$

Proof. Let $y^i := \sqrt{\frac{n}{k}} \cdot x^i$ for each $i \in [k]$. Assume by contradiction that our statement is false, so $|y^i - y^j| \leq \frac{1}{k^{3/2}}$ for each $i \neq j \in [k]$. First, we prove by induction that this implies

$$\frac{|\lambda_h|}{\sqrt{h^2+h}} \leq \left(1 - 2^{-h}\right) \frac{1}{k^{3/2}} \quad \forall\ h \in [k-1].$$

Base Case: $h = 1$. By our assumption, we have that $|y^1 - y^2| \leq \frac{1}{k^{3/2}}$. However, $|y^1 - y^2| = 2\frac{|\lambda_1|}{\sqrt{2}}$, thus $\frac{|\lambda_1|}{\sqrt{2}} \leq \frac{1}{2k^{3/2}} = (1 - 2^{-1})\frac{1}{k^{3/2}}$.
Inductive Step: $(h-1) \rightarrow h$. By our assumption, we have that $|y^h - y^{h+1}| \leq \frac{1}{k^{3/2}}$. However, $|y^h - y^{h+1}| = |2\frac{\lambda_h}{\sqrt{h^2+h}} - \frac{\lambda_{h-1}}{\sqrt{h(h-1)}}| \geq 2\frac{|\lambda_h|}{\sqrt{h^2+h}} - \frac{|\lambda_{h-1}|}{\sqrt{h(h-1)}}$ by the triangle inequality. Moreover, by the inductive hypothesis, $\frac{|\lambda_{h-1}|}{\sqrt{h(h-1)}} \leq (1 - 2^{-(h-1)})\frac{1}{k^{3/2}}$. Thus, $\frac{|\lambda_h|}{\sqrt{h^2+h}} \leq \left(1 - 2^{-h}\right)\frac{1}{k^{3/2}}$.
As a consequence, we also get that

$$|\lambda_h| \leq \frac{\sqrt{h^2+h}}{k^{3/2}} \quad \forall\, h \in [k-1].$$

However, by hypothesis, $\sum_{h=1}^{k-1} |\lambda_h|^2 = 1$. Therefore,

$$1 = \sum_{h=1}^{k-1} |\lambda_h|^2 \leq \frac{1}{k^3}\sum_{h=1}^{k-1}(h^2+h) \leq \frac{1}{k^3}\sum_{h=1}^{k-1} k(k-1) \leq \frac{(k-1)^2 k}{k^3} < 1,$$

which is a contradiction. □

B THE ANALYSIS OF THE SPECTRAL ALGORITHM

Here we provide details and proofs of the theoretical guarantees for our Algorithm Spectral.

We are assuming that there are $k = O(1)$ equinumerous clusters, each of size n/k. However, we also show that all our results apply to the relaxed setting in which each cluster has size $n/k + o(n)$.

B.1 Algorithm Spectral can cope optimally with the pre-adversary

Let

$$\epsilon = \omega(n^{-1/2}) \quad \text{and} \quad B = o(n^2).$$

We now proceed by detailing the proof of Theorem 4.1. We begin with some useful facts about the norms of the zero-error matrix M, its adversarial modification M', and our input matrix M'', perturbed with random noise.

Lemma B.1. $\|M' - M\|_{op} \leq 2\sqrt{B} = o(n)$.

Proof. Define $E := M' - M$. By definition, E has B non-zero entries, each of which is either -2 or 2. Therefore, $\|E\|_F = \sqrt{4B} = 2\sqrt{B}$. By Lemma A.1, this implies that $\|E\|_{op} \leq 2\sqrt{B}$. □

Lemma B.2. $\Pr(\|M'' - \mathbb{E}[M'']\|_{op} \geq 16\sqrt{n}) \leq 2^{-4n}$.

Proof. First, notice that $\mathbb{E}[M''] = \left(\frac{1}{2} + \epsilon\right)M' + \left(\frac{1}{2} - \epsilon\right)(-M') = 2\epsilon \cdot M'$, so

$$M''_{i,j} - \mathbb{E}[M''_{i,j}] := \begin{cases} (1-2\epsilon)M'_{i,j} & \text{w. pr. } \frac{1}{2} + \epsilon; \\ -(1+2\epsilon)M'_{i,j} & \text{w. pr. } \frac{1}{2} - \epsilon. \end{cases}$$

Thus $\frac{1}{1+2\epsilon}(M'' - \mathbb{E}[M''])$ has all the elements bounded by 1 in absolute value. Moreover, we can write it as the sum of its upper triangular part, name it N, and its lower triangular part, N^T: $\frac{1}{1+2\epsilon}(M'' - \mathbb{E}[M'']) = N + N^T$. The matrix N satisfies the hypothesis of Theorem A.7, so $\Pr(\|N\|_{op} \geq 4 \cdot \sqrt{n}) \leq 2^{-4n}$. By the triangle inequality $\Pr(\|N + N^T\|_{op} \geq 8 \cdot \sqrt{n}) \leq \Pr(\|N\|_{op} + \|N^T\|_{op} \geq 8 \cdot \sqrt{n})$. However, $\|N\|_{op} = \|N^T\|_{op}$, so $\Pr(\|N + N^T\|_{op} \geq 8 \cdot \sqrt{n}) = \Pr(\|N\|_{op} \geq 4 \cdot \sqrt{n}) \leq 2^{-4n}$. Finally, $\Pr(\|M'' - \mathbb{E}[M'']\|_{op} \geq 16\sqrt{n}) = \Pr((1+2\epsilon)\|N + N^T\|_{op} \geq 16\sqrt{n}) \leq \Pr(\|N + N^T\|_{op} \geq 8 \cdot \sqrt{n}) \leq 2^{-4n}$ because $\epsilon \leq 1/2$. □

Lemma B.3. *With probability at least* $1 - 2^{-4n} = 1 - o(1)$, $\|M'' - 2\epsilon \cdot M\|_{op} \leq 16\sqrt{n} + 4\epsilon \cdot \sqrt{B} = o(\epsilon n)$.

Proof. By the triangle inequality, $\|M'' - 2\epsilon \cdot M\|_{op} \leq \|M'' - \mathbb{E}[M'']\|_{op} + \|\mathbb{E}[M''] - 2\epsilon \cdot M\|_{op}$. First, by Lemma B.2, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds $\|M'' - \mathbb{E}[M'']\|_{op} \leq 16\sqrt{n}$. Second, we have that $\mathbb{E}[M''] = 2\epsilon \cdot M'$, so $\|\mathbb{E}[M''] - 2\epsilon \cdot M\|_{op} = 2\epsilon \cdot \|M' - M\|_{op}$. However, by Lemma B.1, it holds $\|M' - M\|_{op} \leq 2\sqrt{B}$. By putting everything together, we finally get that $\|M'' - 2\epsilon \cdot M\|_{op} \leq 16\sqrt{n} + 4\epsilon \cdot \sqrt{B}$. Finally, we notice that $16\sqrt{n} + 4\epsilon \cdot \sqrt{B} = o(\epsilon n)$ by Equation 1. □

By Lemma B.3 and by Corollary A.5.1, it follows that the n eigenvalues of M'', in decreasing order, are

$$\frac{4}{k} \cdot \epsilon n + o(\epsilon n), \ldots, \frac{4}{k} \cdot \epsilon n + o(\epsilon n), o(\epsilon n), \ldots, o(\epsilon n), 2 \cdot \left(\frac{2}{k} - 1\right) \cdot \epsilon n,$$

where ${}^4/_k \cdot \epsilon n + o(\epsilon n)$ is repeated $k-1$ times, $o(\epsilon n)$ is repeated $n-k$ times ($n-1$ for $k=2$), and $2 \cdot ({}^2/_k - 1) \cdot \epsilon n$ is repeated only once (notice that this is equal to 0 for $k=2$). Moreover, $\frac{|\lambda''_{i-1}|}{|\lambda''_i|} = 1 + o(1)$ for every $i \in [k-1]$, and $\frac{|\lambda''_{k-1}|}{|\lambda''_k|} = \omega(1)$, so k is exactly the smallest positive integer for which the condition of line 3 of Algorithm 1 holds. This shows a one-to-one correspondence between the $k-1$ largest eigenvalues of the zero-error matrix M and the $k-1$ largest eigenvalues of the input matrix M''. As for the respective eigenvectors, we can use the following results.

Lemma B.4. *Let $\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_{k-1}$ be unitary eigenvectors of the largest $k-1$ eigenvalues of M'', and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ be an orthogonal basis of the largest eigenvalue of M. Let $V \in \mathbb{R}^{n,k-1}$ with $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ as columns, and $V'' \in \mathbb{R}^{n,k-1}$ with $\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_{k-1}$ as columns. Then, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds*

$$\|VV^T - V''(V'')^T\|_F \leq \frac{8k\sqrt{k}}{\epsilon\sqrt{n}} + \frac{2k\sqrt{kB}}{n} = o(1).$$

Proof. By what previously observed, $2\epsilon \cdot M$ is diagonalizable with eigenvalues ${}^4/_k \cdot \epsilon n$, which has multiplicity $k-1$, 0, which has multiplicity $n-k$ ($n-1$ for $k=2$), and, for $k>2$, $2({}^2/_k - 1) \cdot \epsilon n$ too, which has multiplicity 1. Thus, by Theorem 3.1, for any orthogonal basis of eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ of the largest eigenvalue of M, it holds

$$\|VV^T - V''(V'')^T\|_F \leq \frac{2\sqrt{k} \cdot \|M'' - 2\epsilon \cdot M\|_{op}}{\frac{4}{k} \cdot \epsilon n}.$$

Now, by Theorem B.3, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds $\|M'' - 2\epsilon \cdot M\|_{op} \leq 16\sqrt{n} + 4\epsilon \cdot \sqrt{B} = o(\epsilon n)$, so

$$\|VV^T - V''(V'')^T\|_F \leq \frac{2\sqrt{k} \cdot (16\sqrt{n} + 4\epsilon \cdot \sqrt{B})}{\frac{4}{k} \cdot \epsilon n} = \frac{8k\sqrt{k}}{\epsilon\sqrt{n}} + \frac{2k\sqrt{kB}}{n} = o(1).$$

□

As before, we can use these results to show that the eigenspace of the obtained eigenvectors of M'' is "close" to the one of the leading eigenvectors of M.

Lemma B.5. *Let $\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_{k-1}$ be the unitary eigenvectors of the largest $k-1$ eigenvalues of M'', as returned in line 3 of Algorithm 1, and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ be an orthogonal basis of the largest eigenvalue of M. Then, with high probability $\geq 1 - 2^{-4n}$, for each $\boldsymbol{v}_h, h \in [k-1]$, it holds*

$$\sum_{\ell=1}^{k-1} \langle \boldsymbol{v}_h, \boldsymbol{v}''_\ell \rangle^2 \geq 1 - \frac{64k^3}{\epsilon^2 n} - \frac{4k^3 B}{n^2} = 1 - o(1).$$

Analogously, for each $\boldsymbol{v}''_m, m \in [k-1]$, it holds

$$\sum_{h=1}^{k-1} \langle \boldsymbol{v}_h, \boldsymbol{v}''_m \rangle^2 \geq 1 - \frac{64k^3}{\epsilon^2 n} - \frac{4k^3 B}{n^2} = 1 - o(1).$$

Proof. With high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, by Lemma B.4, it holds

$$\|VV^T - V''(V'')^T\|_F \leq \frac{8k\sqrt{k}}{\epsilon\sqrt{n}} + \frac{2k\sqrt{kB}}{n}.$$

Now, we can notice that

$$\|VV^T - V''(V'')^T\|_F^2 = \sum_{i,j=1}^{n}\left(\sum_{h=1}^{k-1}((\boldsymbol{v}_h)_i(\boldsymbol{v}_h)_j - (\boldsymbol{v}''_h)_i(\boldsymbol{v}''_h)_j)\right)^2$$
$$= 2\left(k-1-\sum_{h=1}^{k-1}\sum_{\ell=1}^{k-1}\langle\boldsymbol{v}_h,\boldsymbol{v}''_\ell\rangle^2\right).$$

Now fix a generic $h \in [k-1]$. For each $h' \in [k-1]\setminus\{h\}$, it holds $\sum_{\ell=1}^{k-1}\langle\boldsymbol{v}_{h'},\boldsymbol{v}''_\ell\rangle^2 \le \|\boldsymbol{v}_{h'}\|^2 = 1$, because it is the sum of the projections of orthogonal vectors onto $\boldsymbol{v}_{h'}$. Therefore, $\|VV^T - V''(V'')^T\|_F^2 \ge 2\left(1-\sum_{\ell=1}^{k-1}\langle\boldsymbol{v}_h,\boldsymbol{v}''_\ell\rangle^2\right)$. Moreover, we have shown that with high probability

$$\|VV^T - V''(V'')^T\|_F^2 \le \left(\frac{8k\sqrt{k}}{\epsilon\sqrt{n}} + \frac{2k\sqrt{kB}}{n}\right)^2 \le \frac{128k^3}{\epsilon^2 n} + \frac{8k^3 B}{n^2} = o(1)$$

by using $(a+b)^2 \le 2(a^2+b^2)\ \forall\, a,b \in \mathbb{R}$. Therefore, we have that

$$\sum_{\ell=1}^{k-1}\langle\boldsymbol{v}_h,\boldsymbol{v}''_\ell\rangle^2 \ge 1 - \frac{64k^3}{\epsilon^2 n} - \frac{4k^3 B}{n^2} = 1 - o(1).$$

Since everything is symmetric, the symmetric version of this inequality follows analogously. □

Thanks to this result, we can analyze our spectral approach. By recovering the eigenvectors of the $k-1$ largest eigenvalues of M'', we get a very good approximation of a basis of the eigenspace of the leading eigenvector of M, which can be used to set the clusters apart. We are ready to prove the main result, Theorem 4.1.

Proof of Theorem 4.1. We show that, with high probability $1-o(1)$, the Algorithm 1 is well-defined, so it always succeeds in finding $i \in \mathcal{U}$ satisfying the condition of line 8, and that our solution consists in k clusters and is an approximate reconstruction of the original clusters.

First, for each eigenvector $\boldsymbol{v}''_\ell, \ell \in [k-1]$, we can define $\tilde{\boldsymbol{v}}_\ell := \sum_{h=1}^{k-1}\langle\boldsymbol{v}_h,\boldsymbol{v}''_\ell\rangle\boldsymbol{v}_h$, which is the projection of $\boldsymbol{v}''_\ell$ onto the eigenspace spanned by $\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{k-1}\}$. Now, for each cluster C, we can define $\lambda_{\ell,C}$ as the coordinate of the vertices belonging to cluster C in vector $\tilde{\boldsymbol{v}}_\ell$, which is well-defined by what said about the spectrum of the input matrix. We can also define $\mathcal{S}_{bad}^{\ell,C} := \{i \in C : |(\boldsymbol{v}''_\ell)_i - \lambda_{\ell,C}| > \frac{1}{4\sqrt{2n}}\}$, which is the set of indices of C which have been moved far from λ_C in $\boldsymbol{v}''_\ell$. By Lemma B.5, with high probability $\ge 1-2^{-4n}$ it holds $\|\boldsymbol{v}''_\ell - \tilde{\boldsymbol{v}}_\ell\|^2 \le \frac{64k^3}{\epsilon^2 n} + \frac{4k^3B}{n^2} = o(1)$, thus

$$\frac{64k^3}{\epsilon^2 n} + \frac{4k^3 B}{n^2} \ge \|\boldsymbol{v}''_\ell - \tilde{\boldsymbol{v}}_\ell\|^2 \ge \sum_{i\in\mathcal{S}_{bad}^{\ell,C}}|(\boldsymbol{v}''_\ell)_i - \lambda_{\ell,C}|^2 > \frac{|\mathcal{S}_{bad}^{\ell,C}|}{32n},$$

which implies that $|\mathcal{S}_{bad}^{\ell,C}| \le \frac{2048k^3}{\epsilon^2} + \frac{128k^3B}{n} = o(n)$. Therefore, for each cluster, all but $o(n)$ vertices have coordinates close to $\lambda_{\ell,C}$ in $\boldsymbol{v}''_\ell$.

Second, we show that, for each pair of different clusters C_1, C_2, there exists $\boldsymbol{v}''_\ell, \ell \in [k-1]$, such that $|\lambda_{\ell,C_1} - \lambda_{\ell,C_2}| > \frac{1}{\sqrt{2n}}$. Consider the orthogonal basis of the eigenspace of the main eigenvalue of M, as defined at the beginning of this section. By its definition, for each $\boldsymbol{v}''_\ell, \ell \in [k-1]$, it holds $|\lambda_{\ell,C_1} - \lambda_{\ell,C_2}| = |\langle\boldsymbol{v}_1,\boldsymbol{v}''_\ell\rangle|\cdot\sqrt{\frac{2k}{n}}$, because the coordinates of clusters C_1, C_2 only differ in vector $\boldsymbol{v}_1$, and by an amount of $\sqrt{\frac{2k}{n}}$ (they are $+\sqrt{\frac{k}{2n}}$ and $-\sqrt{\frac{k}{2n}}$). By Lemma B.5, it holds $\sum_{\ell=1}^{k-1}\langle\boldsymbol{v}_1,\boldsymbol{v}''_\ell\rangle^2 \ge 1 - \frac{64k^3}{\epsilon^2 n} - \frac{4k^3B}{n^2} = 1 - o(1)$, so there exists $\boldsymbol{v}''_\ell, \ell \in [k-1]$, such that $\langle\boldsymbol{v}_1,\boldsymbol{v}''_\ell\rangle^2 > \frac{1}{k}$, implying that $|\lambda_{\ell,C_1} - \lambda_{\ell,C_2}| = |\langle\boldsymbol{v}_1,\boldsymbol{v}''_\ell\rangle|\cdot\sqrt{\frac{2k}{n}} > \sqrt{\frac{2}{n}} > \frac{1}{\sqrt{2n}}$.

Third, consider $\mathcal{S}_{bad} := \bigcup_{\ell\in[k-1],C}\mathcal{S}_{bad}^{\ell,C}$. By the union bound and by what just proven,

$$|\mathcal{S}_{bad}| \le k^2\cdot\left(\frac{2048k^3}{\epsilon^2} + \frac{128k^3B}{n}\right) = \frac{2048k^5}{\epsilon^2} + \frac{128k^5B}{n} = o(n).$$

Moreover, for each $i \notin \mathcal{S}_{bad}$, we have that:

- if $j \notin \mathcal{S}_{bad}$ belongs to the same cluster C_1 of i, then for each $\boldsymbol{v}''_\ell, \ell \in [k-1]$ it holds $|(\boldsymbol{v}''_\ell)_i - (\boldsymbol{v}''_\ell)_j| \leq |(\boldsymbol{v}''_\ell)_i - \lambda_{\ell,C_1}| + |\lambda_{\ell,C_1} - (\boldsymbol{v}''_\ell)_j| \leq \frac{1}{2\sqrt{2n}}$ by the triangle inequality;
- if $j \notin \mathcal{S}_{bad}$ belongs to a different cluster C_2 from i, then there exists $\boldsymbol{v}''_\ell, \ell \in [k-1]$, such that $|\lambda_{\ell,C_1} - \lambda_{\ell,C_2}| > \frac{1}{\sqrt{2n}}$, implying that $|(\boldsymbol{v}''_\ell)_i - (\boldsymbol{v}''_\ell)_j| \geq |\lambda_{\ell,C_1} - \lambda_{\ell,C_2}| - |\lambda_{\ell,C_1} - (\boldsymbol{v}''_\ell)_i| - |(\boldsymbol{v}''_\ell)_j - \lambda_{\ell,C_2}| > \frac{1}{2\sqrt{2n}}$ by the triangle inequality.

We have shown that $\frac{1}{2\sqrt{2n}}$ is an appropriate distance threshold to separate elements in different clusters that do not belong to $\mathcal{S}_{bad}$.

Finally, by summing up, since $|\mathcal{S}_{bad}| = o(n)$, at each time step with high probability $\geq 1 - o(1)$ we select $i \notin \mathcal{S}_{bad}$ from line 7. In this case, $\mathcal{S}_i$ has $\frac{n}{k} + o(n)$ elements, which are the elements in its cluster plus/minus eventual elements of $\mathcal{S}_{bad}$. The elements of $\mathcal{S}_{bad}$ could be wrongly added to $\mathcal{S}_i$, or wrongly removed and associated to a different set of those. Since $k = O(1)$, with high probability $\geq 1 - o(1)$ this happens for k straight times. Under all these assumptions, only the elements in $\mathcal{S}_{bad}$ can be classified incorrectly, but they are at most $\frac{2048k^5}{\epsilon^2} + \frac{128k^5 B}{n} = o(n)$ by Eq. 1.

Polynomial Running Time. If we neglect the cost of the Power-Method and of procedure Get-Clusters, Algorithm Spectral has a linear cost in n. Get-Clusters has cost $O(n^2)$ because $k = O(1)$. Finally, as shown in Golub and Van Loan (1996), the running time of the Power-Method on the matrix $M'' \in \mathbb{R}^{n,n}$ with k largest eigenvalues $\lambda''_1 \geq \ldots \geq \lambda''_k$, is $O\left(kn^2 \cdot \frac{\lambda''_1}{\lambda''_{k-1} - \lambda''_k} \cdot \log(1/\gamma)\right)$, where γ is the ℓ_2 error between the reconstructed eigenvectors and the original ones. By what we have shown on the spectrum of M'', it holds $\frac{\lambda''_1}{\lambda''_{k-1} - \lambda''_k} = 1 + o(1)$. Moreover, any error $\gamma = 1/\text{poly}(n)$ gives a good enough result to our purpose. Thus, the total running time is $O(n^2 \log(n)) = \tilde{O}(n^2)$. □

B.2 Sub-Optimal Robustness of Algorithm Spectral with the post-adversary

Let

$$\epsilon = \omega(n^{-1/2}) \quad \text{and} \quad B = o(\epsilon^2 n^2).$$

We now proceed by detailing the proof of Theorem 4.2. As before, we begin with some useful facts about the norms of the zero-error matrix M, its random perturbation M', and our input matrix M'', modified by the adversary. The proofs are only sketched since they are identical to the ones for the pre-adversarial setting.

Lemma B.6. $\Pr(\|M' - \mathbb{E}[M']\|_{op} \geq 16\sqrt{n}) \leq 2^{-4n}$.

Proof. $\frac{1}{2}(M' - \mathbb{E}[M'])$ has all the elements bounded by 1 in absolute value. Moreover, we can write it as the sum of its upper triangular part, exactly as in Lemma B.2. Analogously, we get $\Pr(\|M' - \mathbb{E}[M']\|_{op} \geq 16\sqrt{n}) \leq 2^{-4n}$. □

Lemma B.7. $\|M'' - M'\|_{op} \leq 2\sqrt{B}$.

Proof. Define $E := M'' - M'$. By definition, E has B non-zero entries, each of which has absolute value ≤ 2. Therefore, $\|E\|_F \leq \sqrt{4B} = 2\sqrt{B}$. By Lemma A.1, this implies that $\|E\|_{op} \leq 2\sqrt{B}$. □

Now, we can bound the norm of the difference between the zero-error matrix and the input matrix M''.

Theorem B.8. *With high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds*

$$\|M'' - 2\epsilon \cdot M\|_{op} \leq 16\sqrt{n} + 2\sqrt{B} = o(\epsilon n).$$

Proof. First, by Lemma B.6, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds $\|M' - \mathbb{E}[M']\|_{op} \leq 16\sqrt{n}$. Second, by Lemma B.7 it holds $\|M'' - M'\|_{op} \leq 2\sqrt{B}$. However,

$$\mathbb{E}[M'] = 2\epsilon \cdot M.$$

Thus, by the triangle inequality, $\|M'' - 2\epsilon \cdot M\|_{op} \leq \|M'' - M'\|_{op} + \|M' - \mathbb{E}[M']\|_{op}$. By putting everything together, we get that $\|M'' - 2\epsilon \cdot M\|_{op} \leq 16\sqrt{n} + 2\sqrt{B}$. Finally, this is $o(\epsilon n)$ by Eq. 2. □

We can now proceed by showing that the eigenspace of the $k-1$ leading eigenvalues of M'' is very close to the one of M.

Lemma B.9. *Let $\boldsymbol{v}''_1, \dots, \boldsymbol{v}''_{k-1}$ be unitary eigenvectors of the largest $k-1$ eigenvalues of M'', and let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}$ be an orthogonal basis of the largest eigenvalue of M. Let $V \in \mathbb{R}^{n,k-1}$ with $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}$ as columns, and $V'' \in \mathbb{R}^{n,k-1}$ with $\boldsymbol{v}''_1, \dots, \boldsymbol{v}''_{k-1}$ as columns. Then, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds*

$$\|VV^T - V''(V'')^T\|_F \leq \frac{8k\sqrt{k}}{\epsilon\sqrt{n}} + \frac{k\sqrt{kB}}{\epsilon n} = o(1).$$

Proof. Exactly as in Lemma B.4, by Theorem 3.1, for any orthogonal basis of eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}$ of the largest eigenvalue of M, it holds

$$\|VV^T - V''(V'')^T\|_F \leq \frac{2\sqrt{k} \cdot \|M'' - 2\epsilon \cdot M\|_{op}}{\frac{4}{k} \cdot \epsilon n}.$$

Now, by Theorem B.8, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds $\|M'' - 2\epsilon \cdot M\|_{op} \leq 16\sqrt{n} + 2\sqrt{B} = o(\epsilon n)$, so

$$\|VV^T - V''(V'')^T\|_F \leq \frac{2\sqrt{k} \cdot (16\sqrt{n} + 2\sqrt{B})}{\frac{4}{k} \cdot \epsilon n} = \frac{8k\sqrt{k}}{\epsilon\sqrt{n}} + \frac{k\sqrt{kB}}{\epsilon n} = o(1).$$

□

As before, we can use these results to show that the eigenspace of the obtained eigenvectors of M'' is "close" to the one of the leading eigenvectors of M.

Lemma B.10. *Let $\boldsymbol{v}''_1, \dots, \boldsymbol{v}''_{k-1}$ be the unitary eigenvectors of the largest $k-1$ eigenvalues of M'', as returned in line 3 of Algorithm* Spectral*, and let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}$ be an orthogonal basis of the largest eigenvalue of M. Then, with high probability $\geq 1 - 2^{-4n}$, for each $\boldsymbol{v}_h, h \in [k-1]$, it holds*

$$\sum_{\ell=1}^{k-1} \langle \boldsymbol{v}_h, \boldsymbol{v}''_\ell \rangle^2 \geq 1 - \frac{64k^3}{\epsilon^2 n} - \frac{k^3 B}{\epsilon^2 n^2} = 1 - o(1).$$

Analogously, for each $\boldsymbol{v}''_m, m \in [k-1]$, it holds

$$\sum_{h=1}^{k-1} \langle \boldsymbol{v}_h, \boldsymbol{v}''_m \rangle^2 \geq 1 - \frac{64k^3}{\epsilon^2 n} - \frac{k^3 B}{\epsilon^2 n^2} = 1 - o(1).$$

Proof. The proof is identical to the one of Lemma B.5, and relies on Lemma B.9. □

We have all the necessary bounds to prove Theorem 4.2.

Proof of Theorem 4.2. The proof is analogous to the one of Theorem 4.1.

We get that $|\mathcal{S}_{bad}|$, defined exactly in the same way, is bounded by

$$|\mathcal{S}_{bad}| \leq \frac{2048k^5}{\epsilon^2} + \frac{32k^5 B}{\epsilon^2 n} = o(n).$$

We derive that $\frac{1}{2\sqrt{2n}}$ is an appropriate distance threshold to separate elements in different clusters that do not belong to $\mathcal{S}_{bad}$.

Since $|\mathcal{S}_{bad}| = o(n)$, at each time step with high probability $\geq 1 - o(1)$ we select $i \notin \mathcal{S}_{bad}$ from line 7. In this case, $\mathcal{S}_i$ has $\frac{n}{k} + o(n)$ elements, which are the elements in its cluster plus/minus eventual elements of $\mathcal{S}_{bad}$. The elements of $\mathcal{S}_{bad}$ could be wrongly added to $\mathcal{S}_i$, or wrongly removed and associated to a different set of those. Since $k = O(1)$, with high probability $\geq 1 - o(1)$ this happens for k straight times. Under all these assumptions, only the elements in $\mathcal{S}_{bad}$ can be classified incorrectly, but they are at most $\frac{2048k^5}{\epsilon^2} + \frac{32k^5 B}{\epsilon^2 n} = o(n)$ by Eq. 2.

Polynomial Running Time. Exactly as in the pre-adversarial setting, the total running time is still $O(n^2 \log(n)) = \tilde{O}(n^2)$ with high probability. □

B.3 The Spectral Algorithm beyond Equinumerous Clusters

We show that our spectral algorithm and its theoretical guarantees still hold when all the communities have size ${}^{n}/_{k} + o(n)$.

Recall that the parameters of the pre-adversarial setting satisfy $B = o(n^2), \epsilon = \omega({}^{1}/_{\sqrt{n}})$ (Eq. 1). Given a ground-truth $k-$clustering with all the clusters having size ${}^{n}/_{k} + o(n)$, we can move $o(n)$ points to a different cluster for each of the k clusters, and obtain an equinumerous $k-$clustering. This is equivalent to changing $B' = o(n^2)$ entries in the zero-error matrix M associated to the ground-truth clustering, to obtain a new matrix $\widehat{M}$ representing a close equinumerous clustering[3]. It now suffices to reconstruct the clustering associated to $\widehat{M}$ with $o(n)$ misclassified vertices, since at most other $o(n)$ errors are made when considering $\widehat{M}$ instead of M.

We notice that the $B' = o(n^2)$ changes to entries of M are equivalent to the action of a pre-adversary with a budget of $B' = o(n^2)$ changes over the matrix $\widehat{M}$. As a consequence, the perturbation of M in the pre-adversarial setting with parameters B, ϵ following Eq. 1, is equivalent to a pre-adversarial perturbation of $\widehat{M}$ with parameters $B + B', \epsilon$. Since $B + B' = o(n^2)$, Theorem 4.1 still holds and we can reconstruct the clustering for $\widehat{M}$ with $o(n)$ misclassified vertices with high probability. Instead, a perturbation of M in the post-adversarial setting with parameters ϵ, B following Eq. 2, is equivalent to a pre-adversarial perturbation of $\widehat{M}$ with parameters $B', {}^{1}/_{2}$, followed by a post-adversarial perturbation with parameters ϵ, B. Notice that the pre-adversarial perturbation only consists of the B' adversarial changes to the zero-error matrix M. Therefore, since $B' = o(n^2)$, our spectral algorithm can handle both such semi-adversarial setting, and Theorem 4.2 still holds for $\widehat{M}$.

C THE ANALYSIS OF THE SDP-BASED ALGORITHM

Here we provide details and proofs of the theoretical guarantees for our Algorithm based on semidefinite programming, which is used to achieve optimal reconstruction in the post-adversarial setting.

We are assuming that there are $k = O(1)$ equinumerous clusters, each of size ${}^{n}/_{k}$. However, we also show that all our results apply to the relaxed setting in which each cluster has size ${}^{n}/_{k} + o(n)$.

C.1 A Positive Semidefinite Zero-Error Matrix

We have seen that the original matrix M has a negative eigenvalue $-{}^{k-2}/_{k} \cdot n$ for $k > 2$, with corresponding unitary eigenvector $\boldsymbol{z}$ whose coordinates are all equal to ${}^{1}/_{\sqrt{n}}$. This does not carry any information about the clusters. By removing it from the spectral decomposition, we can consider a different zero-error,

$$P := \frac{k}{2(k-1)} \cdot \left(M + n\left(1 - \frac{2}{k}\right) \boldsymbol{z}\boldsymbol{z}^T \right),$$

whose entries are:

$$P_{i,j} := \begin{cases} 1 & \text{if } i, j \text{ are in the same cluster;} \\ -\frac{1}{k-1} & \text{otherwise.} \end{cases}$$

This matrix has rank $k-1$ and k distinct rows, the last of which is the opposite of the sum of the previous $k-1$ ones. Its spectrum consists of:

- the positive eigenvalue ${}^{n}/_{k-1}$, whose eigenspace has dimension $k-1$, with a basis given by $\{\boldsymbol{f}_i - \boldsymbol{f}_{i+1}\}_{i \in [k-1]}$. Notice that this is also the subspace of vectors having all the same coordinates for vertices in the same cluster and having sum of coordinates equal to 0;
- 0, whose eigenspace has dimension $n - k + 1$ and is the complementary to the previous eigenspace. This subspace is described by the equation $P\boldsymbol{x} = \boldsymbol{0}$.

[3] This is not possible if ${}^{n}/_{k}$ is not an integer. However, if this is the case, we could add just other $\leq k - 1 = O(1)$ extra vertices to the matrix to make it true. This would involve just extra $\leq 2k \cdot n = \Theta(n) = o(n^2)$ changes to the zero-error matrix, so it has a neglectable effect.

An orthogonal basis for the eigenspace of ${}^{n}/_{k-1}$ has already been found in Equation 4. It is also useful to find the value of some norms for the matrix P.

Lemma C.1. *We have that* $\|P\|_F = {}^{n}/_{\sqrt{k-1}}$, $\|P\|_{op} = {}^{n}/_{k-1}$, *and* $\|P\|_{SDP} = {}^{n^2}/_{k-1}$. *Moreover, when k is even* $\|P\|_{\infty\to 1} = {}^{n^2}/_{k-1}$, *while when k is odd,* $\|P\|_{\infty\to 1} = n^2(k+1)/k^2$.

Proof. The first two equations follow by what we have just said on the spectrum of P. As for the third one, it is easy to observe that the ± 1 values in the corresponding norm should be symmetric. Let α be the number of $+1$ and $\beta = k - \alpha$ be the number of -1 in the optimal solution for the case $n = k$ (when $k < n$, it just suffices to multiply everything by $\frac{n^2}{k^2}$). We have that

$$\|P\|_{\infty\to 1} = 1 \cdot k + \frac{1}{k-1} \cdot [\alpha \cdot (\beta - \alpha + 1) + \beta \cdot (\alpha - \beta + 1)] = k + \frac{\alpha + \beta - (\alpha - \beta)^2}{k-1}.$$

Now, since $\alpha + \beta = k$, we get $k + 1 + \frac{1}{k-1} - \frac{(2\alpha - k)^2}{k-1}$, which is maximized when α is as close as possible to $k/2$, yielding different values for k even/odd, respectively $k + 1 + \frac{1}{k-1} = \frac{k^2}{k-1}$ and $k+1$. The SDP norm of P is the maximum of its Frobenius scalar product with a set of positive semidefinite matrices which contains P itself, so it is

$$\|P\|_{SDP} = P \bullet P = k + \frac{k^2 - k}{(k-1)^2} = k + 1 + \frac{1}{k-1} = \frac{k^2}{k-1}.$$

□

We finally need to assess how P changes after a random perturbation as the one described in the random model for the original matrix M. We define

$$P'_{i,j} := \begin{cases} M'_{i,j} & \text{if } M'_{i,j} > 0; \\ \frac{M'_{i,j}}{k-1} & \text{otherwise.} \end{cases}$$

Equivalently, we can write

$$P'_{i,j} = \begin{cases} P_{i,j} & \text{w. pr. } \frac{1}{2} + \epsilon; \\ -P_{i,j} + (1 - \frac{1}{k-1}) & \text{w. pr. } \frac{1}{2} - \epsilon. \end{cases}$$

We can see that this turns 1 into $-\frac{1}{k-1}$ w. pr. $\frac{1}{2} - \epsilon$ and vice versa. Observe that:

$$\mathbb{E}[P'] = \left(\frac{1}{2} - \epsilon\right)\left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1} + 2\epsilon \cdot P, \tag{5}$$

where $\mathbb{1}$ is the $n \times n$ matrix with all entries equal to 1. We can now define $Q := P' - \left(\frac{1}{2} - \epsilon\right)\left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1}$. By what we have said, this gives $\mathbb{E}[Q] = 2\epsilon \cdot P$, so Q can be used as a random perturbation of the matrix $2\epsilon \cdot P$.

C.2 A Novel Optimal Algorithm with Recursive Semidefinite Programming

In this algorithm, we use the positive semidefinite matrix P instead of M, with its random perturbation P' and its post-adversarial perturbation P'', which can be obtained from M'' by turning its negative entries to $-\frac{1}{k-1}$. We also recall that

$$\mathbb{E}[P'] = 2\epsilon \cdot P + \left(\frac{1}{2} - \epsilon\right)\left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1}.$$

We can define

$$Q := P'' - \left(\frac{1}{2} - \epsilon\right)\left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1} = \frac{k}{2(k-1)} \cdot M'' + \epsilon\left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1}.$$

By doing so, Q can be seen as a perturbation of $2\epsilon \cdot P$, which is a positive semidefinite matrix. Thus, it can be effectively used as the input matrix for an SDP that aims to reconstruct the clusters. Notice that this definition of Q is different from the one in the main paper and requires the knowledge of the parameter ϵ. For now, we use this last definition of Q and assume to have access to the parameter ϵ to avoid overcomplicating the proofs. However, at the end of this section, we will argue on how to do without this assumption.

C.2.1 General Properties

Here we prove some norm inequalities involving the matrices P, P', P'', and Q.

Lemma C.2.

$$\Pr(\|P' - \mathbb{E}[P']\|_{\infty\to 1} \geq 16n\sqrt{n}) \leq 2^{-4n}.$$

Proof. First, it holds

$$P'_{i,j} - \mathbb{E}[P'_{i,j}] := \begin{cases} (1-2\epsilon)P_{i,j} - \left(\frac{1}{2}-\epsilon\right)\left(1-\frac{1}{k-1}\right) & \text{w. pr. } \frac{1}{2}+\epsilon; \\ -(1+2\epsilon)P_{i,j} + \left(\frac{1}{2}+\epsilon\right)\left(1-\frac{1}{k-1}\right) & \text{w. pr. } \frac{1}{2}-\epsilon. \end{cases}$$

Thus $\frac{1}{1+2\epsilon}(P' - \mathbb{E}[P'])$ has all the elements bounded by 1 in absolute value. By Lemma B.2, it follows that $\Pr(\|P' - \mathbb{E}[P']\|_{op} \geq 16\sqrt{n}) \leq 2^{-4n}$. Now, by Lemma A.3, it holds $\|P' - \mathbb{E}[P']\|_{op} \geq \frac{1}{n} \cdot \|P' - \mathbb{E}[P']\|_{\infty\to 1}$, so

$$\Pr(\|P' - \mathbb{E}[P']\|_{\infty\to 1} \geq 16n\sqrt{n}) \leq \Pr(\|P' - \mathbb{E}[P']\|_{op} \geq 16\sqrt{n}) \leq 2^{-4n}.$$

□

We can also bound the norm displacement after the post-adversary intervention.

Lemma C.3.

$$\|P'' - P'\|_{\infty\to 1} \leq 2B = o(\epsilon n^2).$$

Proof. Let $P'' = P' + E$, where E, the matrix of adversarial changes, has $B = o(\epsilon n^2)$ non-zero entries, all with absolute value $1 + 1/k-1 \leq 2$. Therefore, $\|E\|_{\infty\to 1} \leq 2B$. □

Consider the auxiliary matrix Q, defined as:

$$Q = P'' - \left(\frac{1}{2} - \epsilon\right)\left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1}.$$

Lemma C.4. *With high probability* $\geq 1 - 2^{-4n} = 1 - o(1)$*, it holds*

$$\|Q - 2\epsilon \cdot P\|_{\infty\to 1} \leq 16n\sqrt{n} + 2B = o(\epsilon n^2).$$

Proof. By definition of Q and P', we get that

$$Q - 2\epsilon \cdot P = (P'' - P') + (P' - \mathbb{E}[P']).$$

By Lemma C.2, with probability $\geq 1 - 2^{-4n}$ it holds $\|P' - \mathbb{E}[P']\|_{\infty\to 1} \leq 16n\sqrt{n}$; by Lemma C.3, it holds $\|P'' - P'\|_{\infty\to 1} \leq 2B$. By putting everything together and using the triangle inequality, we finally get that with high probability ($\geq 1 - 2^{-4n}$)

$$\|Q - 2\epsilon \cdot P\|_{\infty\to 1} \leq 16n\sqrt{n} + 2B = o(\epsilon n^2).$$

□

C.2.2 A Recursive SDP-Based Approach

Consider the following SDP:

$$\begin{array}{lll} \text{maximize} & \displaystyle\sum_{j=1}^{n} Q_{ij}\langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle & \\ \text{subject to} & \|\boldsymbol{x}_i\| = 1, \boldsymbol{x}_i \in \mathbb{R}^n & i = 1, \ldots, n \\ & \|\boldsymbol{y}_i\| = 1, \boldsymbol{y}_i \in \mathbb{R}^n & i = 1, \ldots, n \end{array} \tag{6}$$

The maximum of this SDP is equal to $\|Q\|_{SDP}$. Given the optimal solution $\{\boldsymbol{x}_i^*\}_{i\in[n]}, \{\boldsymbol{y}_i^*\}_{i\in[n]}$, consider the matrix X where $X_{ij} := \langle \boldsymbol{x}_i^*, \boldsymbol{y}_j^* \rangle \ \forall\, i, j \in [n]$. Then, $\|Q\|_{SDP} = Q \bullet X$, where $\bullet$ represents the Kronecker (element-wise) product. Since Q is symmetric, by a well-known characteristic of SDPs, X is symmetric too. The following lemma holds.

Lemma C.5. *With high probability ($\geq 1 - 2^{-4n}$), it holds*

$$\|Q\|_{SDP} = Q \bullet X \geq \frac{2}{k-1} \cdot \epsilon n^2 - 29n\sqrt{n} - 4B = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2).$$

Proof. By the triangle inequality, we have that

$$\|Q\|_{SDP} \geq 2\epsilon \cdot \|P\|_{SDP} - \|Q - 2\epsilon \cdot P\|_{SDP}.$$

Now, by Lemma C.4, with high probability $\geq 1 - 2^{-4n} = 1 - o(1)$, it holds $\|Q - 2\epsilon \cdot P\|_{\infty \to 1} \leq 16n\sqrt{n} + 2B = o(\epsilon n^2)$ so, by Theorem A.4, we get that

$$\|Q - 2\epsilon \cdot P\|_{SDP} \leq 1.8 \cdot (16n\sqrt{n} + 2B) \leq 29n\sqrt{n} + 4B = o(\epsilon n^2).$$

Moreover, by Lemma C.1, we get that $\|P\|_{SDP} = \frac{n^2}{k-1}$. By substituting these above, and exploiting Eq. 3, we finally get that

$$\|Q\|_{SDP} \geq \frac{2}{k-1} \cdot \epsilon n^2 - 29n\sqrt{n} - 4B = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2).$$

□

Since X is symmetric, we can consider its spectral decomposition into orthogonal eigenvectors:

$$X = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}^i (\boldsymbol{u}^i)^T.$$

Now, consider $\boldsymbol{u}^*$ picked randomly in $\{\boldsymbol{u}^i, i \in [n]\}$, where $\boldsymbol{u}^i$ is picked with probability proportional to λ_i. We will show that, with high probability, $\boldsymbol{u}^*$ gives a separation of the vertices into two sets, each containing at least one original cluster, and putting almost always together vertices belonging to the same cluster.

Lemma C.6. *Consider $Q \in \mathbb{R}^{n,n}$ such that $\|Q - 2\epsilon \cdot P\|_{\infty \to 1} \leq f(n, B, \epsilon) = o(\epsilon n^2)$. Let X be the (symmetric positive semidefinite) solution matrix of SDP 6 w.r.t. Q, let $\{\boldsymbol{u}^i\}_{i \in [n]}$ an orthogonal basis of eigenvectors for X with eigenvalues respectively $\{\lambda_i\}_{i \in [n]}$. Pick $\boldsymbol{u}^* \in \{\boldsymbol{u}^i\}_{i \in [n]}$ randomly, where each $\boldsymbol{u}^i$ is chosen with probability $\frac{\lambda_i}{n}$. Then, with high probability $\geq 1 - 2k \cdot \sqrt{\frac{f(n,B,\epsilon)}{\epsilon n^2}} - 2^{-4n} = 1 - o(1)$, there exists $\boldsymbol{v}$ eigenvector of P with eigenvalue $\frac{n}{k-1}$ such that*

$$\|\boldsymbol{u}^* - \boldsymbol{v}\|^2 \leq 4k \cdot \sqrt{\frac{f(n, B, \epsilon)}{\epsilon n^2}} = o(1).$$

Proof. First, by definition of X it holds $\|Q\|_{SDP} = Q \bullet X$. Now, by the triangle inequality, $|\|Q\|_{SDP} - 2\epsilon \cdot \|P\|_{SDP}| \leq \|Q - 2\epsilon \cdot P\|_{SDP}$. However, by Theorem A.4, $\|Q - 2\epsilon \cdot P\|_{SDP} \leq 1.8 \cdot \|Q - 2\epsilon \cdot P\|_{\infty \to 1} \leq 1.8 \cdot f(n, B, \epsilon) = o(\epsilon n^2)$. Thus, by Lemma C.1, we get that $\|Q\|_{SDP} \geq 2\epsilon \|P\|_{SDP} - \|Q - 2\epsilon \cdot P\|_{SDP} \geq \frac{2}{k-1} \cdot \epsilon n^2 - 1.8 \cdot f(n, B, \epsilon) = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2)$. Now, recall that $\|Q - 2\epsilon \cdot P\|_{SDP} \leq 1.8 \cdot f(n, B, \epsilon)$, so $|(Q - 2\epsilon \cdot P) \bullet X| \leq \|Q - 2\epsilon \cdot P\|_{SDP} \leq 1.8 \cdot f(n, B, \epsilon) = o(\epsilon n^2)$. So, by the triangle inequality, we also get that $(2\epsilon P) \bullet X \geq Q \bullet X - |(Q - 2\epsilon \cdot P) \bullet X| \geq \|Q\|_{SDP} - \|Q - 2\epsilon \cdot P\|_{SDP}$, implying that $(2\epsilon P) \bullet X \geq \frac{2}{k-1} \cdot \epsilon n^2 - 3.6 \cdot f(n, B, \epsilon) = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2)$, i.e. that

$$P \bullet X \geq \frac{n^2}{k-1} - \frac{1.8}{\epsilon} \cdot f(n, B, \epsilon) \geq \frac{n^2}{k-1} - \frac{2}{\epsilon} \cdot f(n, B, \epsilon) = \frac{n^2}{k-1} - o(n^2). \tag{7}$$

Now, we can use the spectral decomposition $X = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}^i (\boldsymbol{u}^i)^T$ is positive semidefinite, and $\sum_{i=1}^{n} \lambda_i = \mathrm{tr}(X) = \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 = n$. Therefore, $\{\lambda_i / n\}_i$ can be seen as a probability distribution. So it holds

$$P \bullet X = n \sum_{i=1}^{n} \frac{\lambda_i}{n} (\boldsymbol{u}^i)^T P \boldsymbol{u}^i.$$

However, $P = \frac{n}{k-1} \sum_{j=1}^{k-1} \boldsymbol{v}_j \boldsymbol{v}_j^T$ and it is positive semidefinite, so $|(\boldsymbol{u}^i)^T P \boldsymbol{u}^i| = \frac{n}{k-1} \sum_{j=1}^{k-1} \langle \boldsymbol{v}_j, \boldsymbol{u}^i \rangle^2$ for each vector $\boldsymbol{u}^i$, implying that

$$\sum_{i=1}^{n} \frac{\lambda_i}{n} \sum_{j=1}^{k-1} \langle \boldsymbol{v}_j, \boldsymbol{u}^i \rangle^2 = \frac{k-1}{n^2} \cdot P \bullet X \geq 1 - \frac{2(k-1)}{\epsilon n^2} \cdot f(n, B, \epsilon) \geq 1 - \frac{2k}{\epsilon n^2} \cdot f(n, B, \epsilon).$$

We can notice that the LHS is exactly $\mathbb{E}\left[\sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}^*\rangle^2\right]$. So, we have shown that

$$\mathbb{E}\left[\sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}^*\rangle^2\right] = \sum_{i=1}^{n}\frac{\lambda_i}{n}\sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}^i\rangle^2 \geq 1 - \frac{2k}{\epsilon n^2}\cdot f(n,B,\epsilon) = 1 - o(1). \tag{8}$$

Moreover, for each vector $\boldsymbol{u}$, the quantity $\sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}\rangle^2$ is the squared norm of its projection onto the eigenspace of the eigenvalue $\frac{n}{k-1}$ of P, so it is always a quantity in $[0,1]$. Therefore, we can define the positive random variable $\chi := 1 - \sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}^*\rangle^2$. We have that $\chi \geq 0$ and $\mathbb{E}[\chi] \leq \frac{2k}{\epsilon n^2}\cdot f(n,B,\epsilon) = o(1)$. Thus, by Theorem A.6, we have that with high probability $\geq 1 - \sqrt{\frac{2k}{\epsilon n^2}\cdot f(n,B,\epsilon)} \geq 1 - 2k\sqrt{\frac{f(n,B,\epsilon)}{\epsilon n^2}}$, it holds $\chi \leq \sqrt{\frac{2k}{\epsilon n^2}\cdot f(n,B,\epsilon)}$, implying that

$$\sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}^*\rangle^2 \geq 1 - \sqrt{\frac{2k}{\epsilon n^2}\cdot f(n,B,\epsilon)} = 1 - o(1).$$

Now, let

$$\boldsymbol{v}' := \sum_{j=1}^{k-1}\langle \boldsymbol{v}_j, \boldsymbol{u}^*\rangle \boldsymbol{v}_j;\ \ \boldsymbol{v} := \frac{\boldsymbol{v}'}{\|\boldsymbol{v}'\|}$$

be the normalized projection of $\boldsymbol{u}^*$ onto the eigenspace of the eigenvalue $\frac{n}{k-1}$ of P. It holds (using that $\sqrt{1-x} \geq 1 - x\ \forall\ x \in [0,1]$)

$$\|\boldsymbol{v} - \boldsymbol{u}^*\|^2 = \langle \boldsymbol{v} - \boldsymbol{u}^*, \boldsymbol{v} - \boldsymbol{u}^*\rangle = 2 - 2\langle \boldsymbol{v}, \boldsymbol{u}^*\rangle = 2 - \frac{2}{\|\boldsymbol{v}'\|}\langle \boldsymbol{v}', \boldsymbol{u}^*\rangle = 2 - 2\|\boldsymbol{v}'\| \leq$$

$$2 - 2\sqrt{1 - \sqrt{\frac{4k}{\epsilon n^2}\cdot f(n,B,\epsilon)}} \leq 2\sqrt{\frac{2k}{\epsilon n^2}\cdot f(n,B,\epsilon)} \leq 4k\cdot\sqrt{\frac{f(n,B,\epsilon)}{\epsilon n^2}} = o(1).$$

□

As a consequence of Lemma C.6, Lemma C.4 and Lemma A.8, we can use $\boldsymbol{u}^*$ to separate $[n]$ into two smaller sets with minimal separation of vertices in the same cluster and with at least one cluster on each side. We see how through Algorithm Recursive-Clust$([n], k, f, 1)$ where $f = f(n,B,\epsilon) = 16n\sqrt{n} + 2B = o(\epsilon n^2)$. Before that, we need a formal definition.

Definition C.1. Let $P \in \mathbb{R}^{n,n}$ matrix and let $\mathcal{S}_1, \mathcal{S}_2 \subseteq [n]$. We define $P^{\mathcal{S}_1,\mathcal{S}_2} \in \mathbb{R}^{|\mathcal{S}_1|,|\mathcal{S}_2|}$ be the sub-matrix of P restricted only to the rows in $\mathcal{S}_1$ and to the columns in $\mathcal{S}_2$.

In Algorithm 3, we recursively solve semidefinite programs.

First, we need to sample an appropriate threshold value for the eigenvector to separate the vertices (line 9), because we are getting an approximate eigenvector of the eigenspace of the leading vector of P, but we do not know exactly which approximate eigenvector we are getting. This is done by procedure Get-Threshold. The threshold could be at any point in between the maximum and the minimum value of an eigenvector of P. However, since we only get an approximate eigenvector, we need to consider more robust order statistics to establish the feasible range for thresholds.

Second, we need to fix the cardinality of the bisection (line 14) because we want the size of each partition to be an integer multiple of n/k.

Third, we need to carry the information about the number of clusters in each partition: this will be used to scale the negative elements of the input matrix Q of SDP 6 (line 5), so that this input matrix is always positive semidefinite.

We also need to carry an estimate f' of the distance in norm ℓ_∞-to-ℓ_1 between the scaled matrix Q and the scaled original matrix $2\epsilon \cdot P$, whose negative entries are scaled like the ones of Q.

We show that Algorithm 3, with high probability, always splits the solution into two "smaller" solutions that we can solve recursively, i.e. that the bisection of the input set $\mathcal{S}$ satisfies the condition of line 12. Moreover, we also show that the produced solutions only mislabel $o(n)$ vertices at each step. Before that, we prove an auxiliary lemma.

Algorithm 3 Recursive-Clust$(\mathcal{S}, k', f, \gamma)$: input $\mathcal{S}$ set of indices, k' number of clusters in $\mathcal{S}$, $f = f(n, B, \epsilon) = o(\epsilon n^2)$ such that $\|Q^{\mathcal{S},\mathcal{S}} - 2\epsilon \cdot P^{\mathcal{S},\mathcal{S}}\|_{\infty\to 1} \leq f$ (after having their negative entries multiplied by γ), γ rescaling factor for the negative entries of Q. Global variables n, k, Q, ϵ.

```
$\delta \leftarrow 4k \cdot \sqrt{\frac{f}{\epsilon n^2}}$
$n' \leftarrow |\mathcal{S}|$
**if** $n' = n/k$ **then**
    **return** $\{\mathcal{S}\}$
Let $Q_{k'}$ be the matrix obtained from $Q^{\mathcal{S},\mathcal{S}}$ by multiplying its negative coordinates by $\gamma$
Let $X$ be the solution matrix of SDP 6 for $Q_{k'}$ obtained through SDP-Solver
Let $\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n\}$ be an orthogonal basis of eigenvectors of $X$ with eigenvalues respectively $\{\lambda_i, i \in [n']\}$ obtained through Power-Method
Sample $\boldsymbol{u}^* \in \{\boldsymbol{u}^i, i \in [n']\}$ with probability distribution $\{\frac{\lambda_i}{n'}, i \in [n']\}$
$t \leftarrow$ Get-Threshold$(\boldsymbol{u}^*, n', \delta)$ is the separating threshold according to vector $\boldsymbol{u}^*$
Let $\mathcal{S}_1 := \{i \in \mathcal{S} : \boldsymbol{u}^*_i < t\}$
$k'' := \lfloor \frac{|\mathcal{S}_1|}{n/k} \rceil$ (closest integer function $\lfloor \cdot \rceil$)
**if** $k'' \in \{0, k'\}$ **then**
    **abort** (the algorithm failed)
$\mathcal{S}' \leftarrow \{i \in [n'] : \boldsymbol{u}^*_i$ is among the $k'' \cdot \frac{n}{k}$ smallest coordinates of $\boldsymbol{u}^*_i$ (ties broken arbitrarily)$\}$
$f' \leftarrow k \cdot f + 4k\delta^{1/3} \cdot \epsilon(n')^2 = o(\epsilon n^2)$
$\gamma' \leftarrow \frac{k'-1}{k''-1}$ scaling factor for $\mathcal{S}'$ because it now contains $k''$ clusters instead of $k'$
$\gamma'' \leftarrow \frac{k'-1}{k'-k''-1}$ scaling factor for $\mathcal{S} \setminus \mathcal{S}'$ because it contains the remaining $k' - k''$ clusters
$\mathcal{C}_1 \leftarrow$ Recursive-Clust$(\mathcal{S}', k'', f', \gamma')$
$\mathcal{C}_2 \leftarrow$ Recursive-Clust$(\mathcal{S} \setminus \mathcal{S}', k' - k'', f', \gamma'')$
**return** $\mathcal{C}_1 \cup \mathcal{C}_2$
```

Algorithm 4 Get-Threshold

```
**Procedure** Get-Threshold$(\boldsymbol{u}, n', \delta)$
Let $\pi$ be the ordering permutation of vector $\boldsymbol{u}$, i.e. the permutation on $[n']$ s.t. $\boldsymbol{u}_{\pi(i)} \leq \boldsymbol{u}_{\pi(j)}\ \forall\ 1 \leq i \leq j \leq n'$

$t_{\min} \leftarrow \boldsymbol{u}_{\pi(\lceil \delta^{1/3} \cdot n' \rceil)}$
$t_{\max} \leftarrow \boldsymbol{u}_{\pi(n' - \lceil \delta^{1/3} \cdot n' \rceil)}$
Pick $t \in [t_{\min}, t_{\max}]$ Uniformly At Random as the separating threshold for vector $\boldsymbol{u}$
**return** $t$
```

Lemma C.7. *Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ s.t. $\|\boldsymbol{u}-\boldsymbol{v}\|^2 \le \delta$. Suppose that the coordinates of $\boldsymbol{v}$ can be partitioned into k groups $P_1, \dots, P_k$ of $\frac{n}{k}$ elements each, such that all the coordinates in the same group P_i are equal. Then, for each group P_i, there cannot be $\ge \delta^{1/3} \cdot n$ elements $j \in P_i$ such that $|\boldsymbol{u}_j - \boldsymbol{v}_j| > \frac{\delta^{1/3}}{\sqrt{n}}$.*

Proof. Assume there is $P_i' \subseteq P_i$ such that $|P_i'| \ge \delta^{1/3} \cdot n$ and $|\boldsymbol{u}_j - \boldsymbol{v}_j| > \frac{\delta^{1/3}}{\sqrt{n}}$ $\forall\, j \in P_i'$. Then,

$$\delta \ge \|\boldsymbol{v}-\boldsymbol{u}\|^2 \ge \sum_{j \in P_i'} (\boldsymbol{u}_j - \boldsymbol{v}_j)^2 > (\delta^{1/3} \cdot n) \cdot \left(\frac{\delta^{1/3}}{\sqrt{n}}\right)^2 \ge \delta,$$

which is a contradiction. □

In other words, in $\boldsymbol{u}$, all but $\le \delta^{1/3} \cdot n$ elements of a group are within a distance $\le \frac{\delta^{1/3}}{\sqrt{n}}$ from their coordinate in $\boldsymbol{v}$. The previous lemma is used to show that almost all the coordinates of approximate eigenvectors of P are very close to the coordinates of the actual eigenvector of P. In order to get closer to the proof of the effectiveness of Algorithm 3, we state precise guarantees on what happens in the first round. To extend this to the recursive sub-problems, we need to make some adjustments to take into account the previous classification errors too.

Lemma C.8. *Consider the invocation of* Recursive-Clust$([n], k, f, 1)$ *where $f = f(n, B, \epsilon) = 16n\sqrt{n} + 2B = o(\epsilon n^2)$, and let $\delta := 4k \cdot \sqrt{\frac{f}{\epsilon n^2}} = 4k \cdot \sqrt{\frac{16}{\epsilon\sqrt{n}} + \frac{2B}{\epsilon n^2}} = o(1)$. With high probability $\ge 1 - 2^{-4n} - 6k^2 \cdot \delta^{1/3} = 1 - o(1)$, the first sampled threshold t does not satisfies the condition of line 12 in Algorithm 3, so the algorithm does not fail. Moreover:*

- *for each cluster, either $\mathcal{S}'$ or $\mathcal{S} \setminus \mathcal{S}'$ contains $\le 2\delta^{1/3} \cdot n = o(n)$ of its vertices, meaning that there are $\le 2k\delta^{1/3} \cdot n = o(n)$ misplaced vertices overall in the first recursive step;*
- *let $\mathcal{A}$ be one of the sub-sets on which the algorithm is applied recursively (the same holds for the other subset), and let $\mathcal{A}^*$ be the union of the $k_{\mathcal{A}}$ clusters having $\ge \frac{n}{k} - 2\delta^{1/3} \cdot n$ elements in $\mathcal{A}$. Let $Q_{k_{\mathcal{A}}}$ be the matrix obtained from Q by multiplying the negative entries by $\frac{k_{\mathcal{A}}-1}{k-1}$ and let $P_{k_{\mathcal{A}}}$ be the analogous matrix obtained from P. Then,*

$$\|Q_{k_{\mathcal{A}}}^{\mathcal{A}^*, \mathcal{A}^*} - 2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A}, \mathcal{A}}\|_{\infty \to 1} \le f' := k \cdot f(n, B, \epsilon) + 4k\delta^{1/3} \cdot \epsilon n^2 = o(\epsilon n^2).$$

Proof. By Lemma C.4, with high probability $\ge 1 - 2^{-4n}$ it holds $\|Q - 2\epsilon \cdot P\|_{\infty \to 1} \le f(n, B, \epsilon) = 16n\sqrt{n} + 2B = o(\epsilon n^2)$, and we consider this to be true from now on (by the union bound, the small probability of this to be false will sum up with the other encountered small probabilities). Therefore, by Lemma A.2, for each set of indices $\mathcal{S} \subseteq [n]$, it also holds $\|Q^{\mathcal{S}, \mathcal{S}} - 2\epsilon \cdot P^{\mathcal{S}, \mathcal{S}}\|_{\infty \to 1} \le 16n\sqrt{n} + 2B = o(\epsilon n^2)$. Now, by Lemma C.6, with high probability $\ge 1 - \delta/2 = 1 - o(1)$, there exists an eigenvector $\boldsymbol{v}$ of the leading eigenvalue $\frac{n}{k-1}$ of P such that $\|\boldsymbol{v} - \boldsymbol{u}^*\|^2 \le \delta = o(1)$. Now, let $v_{\max} := \max_{i \in [n]} \boldsymbol{v}_i$ and $v_{\min} := \min_{i \in [n]} \boldsymbol{v}_i$. By Lemma A.8, it holds $|v_{\max} - v_{\min}| > \frac{1}{k\sqrt{n}}$. By Lemma C.7, it follows that $|t_{\max} - v_{\max}| \le \frac{\delta^{1/3}}{\sqrt{n}}$ and $|t_{\min} - v_{\min}| \le \frac{\delta^{1/3}}{\sqrt{n}}$. As a consequence, we have that

$$\Pr\left(t \in \left[v_{\min} + \frac{\delta^{1/3}}{\sqrt{n}}, v_{\max} - \frac{\delta^{1/3}}{\sqrt{n}}\right]\right) \ge \frac{|v_{\max} - v_{\min}| - 2\frac{\delta^{1/3}}{\sqrt{n}}}{|v_{\max} - v_{\min}| + 2\frac{\delta^{1/3}}{\sqrt{n}}} \ge$$

$$1 - \frac{4\frac{\delta^{1/3}}{\sqrt{n}}}{\frac{1}{k\sqrt{n}} + 2\frac{\delta^{1/3}}{\sqrt{n}}} \ge 1 - 4k \cdot \delta^{1/3} = 1 - o(1).$$

However, if $t \in \left[v_{\min} + \frac{\delta^{1/3}}{\sqrt{n}}, v_{\max} - \frac{\delta^{1/3}}{\sqrt{n}}\right]$, by Lemma C.7, we separate almost exactly the clusters corresponding to the largest and the smallest coordinate of $\boldsymbol{v}$: $\ge \frac{n}{k} - \delta^{1/3} \cdot n$ elements of each cluster are split correctly according to the threshold, which makes the condition of line 12 not satisfied and the algorithm does not fail. Now, we need to show that the bisection misplaces $o(n)$ vertices for each cluster. First, notice that, with high probability $\ge 1 - \delta/2 = 1 - o(1)$, by Lemma C.7, for each cluster C with coordinate v_C in $\boldsymbol{v}$, all but $\delta^{1/3} \cdot n = o(n)$ elements of C have coordinates of $\boldsymbol{u}^*$ in the interval $\left[v_C - \frac{\delta^{1/3}}{\sqrt{n}}, v_C + \frac{\delta^{1/3}}{\sqrt{n}}\right]$. Therefore, by the union bound over all the clusters,

$$\Pr\left(\nexists\, C : t \in \left[v_C - \frac{\delta^{1/3}}{\sqrt{n}}, v_C + \frac{\delta^{1/3}}{\sqrt{n}}\right]\right) \geq \frac{t_{\max} - t_{\min} - 2k \cdot \frac{\delta^{1/3}}{\sqrt{n}}}{t_{\max} - t_{\min}} \geq 1 - 2k^2\delta^{1/3}.$$

Thus, with probability $\geq 1 - 2k^2\delta^{1/3} = 1 - o(1)$, we are outside each of those cluster intervals, meaning that, for each cluster, we can misplace $\leq \delta^{1/3} \cdot n$ vertices, for a total of $k\delta^{1/3} \cdot n$ total misplaced vertices according to the threshold bisection at t. Finally, the process of line 14, can bring other $k\delta^{1/3} \cdot n$ mistakes (extra $\delta^{1/3} \cdot n$ for each clusters), for a total of $2k\delta^{1/3} \cdot n = o(n)$ misplaced vertices. By the union bound, everything holds with probability $\geq 1 - 2^{-4n} - \delta/2 - 2k(k+2) \cdot \delta^{1/3} \geq 1 - 2^{-4n} - 6k^2 \cdot \delta^{1/3}$ (for sufficiently small δ).

We now focus on the correctness of the estimate f' of the ℓ_∞-to-ℓ_1 norm of the generated subsets. First, we notice that $|\mathcal{A}| = |\mathcal{A}^*| = k_{\mathcal{A}} \cdot \frac{n}{k}$. Now, by what just proved, we can assume that $\mathcal{A}\Delta\mathcal{A}^* \leq 2k\delta^{1/3} \cdot n = o(n)$. By the triangle inequality

$$\|2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A},\mathcal{A}} - Q_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1} \leq \|2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A},\mathcal{A}} - 2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1} + \|2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*} - Q_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1}.$$

By Lemma C.4 we have that, under the previously mentioned events holding with high probability, $\|Q - 2\epsilon \cdot P\|_{\infty\to1} \leq 16n\sqrt{n} + 2B = o(\epsilon n^2)$. Therefore, by Lemma A.2, it follows that

$$\|2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*} - Q_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1} \leq \frac{k-1}{k_{\mathcal{A}}-1} \cdot \|2\epsilon \cdot P^{\mathcal{A}^*,\mathcal{A}^*} - Q^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1} \leq 16k \cdot n\sqrt{n} + 2k \cdot B = o(\epsilon n^2).$$

Moreover,

$$\|2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A},\mathcal{A}} - 2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1} = 2\epsilon \cdot \|P_{k_{\mathcal{A}}}^{\mathcal{A},\mathcal{A}} - P_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1}.$$

Since $\mathcal{A}\Delta\mathcal{A}^* \leq 2k\delta^{1/3} \cdot n = o(n)$ and the entries of $P_{k_{\mathcal{A}}}$ are bounded in absolute value by 1, we get that $\|P_{k_{\mathcal{A}}}^{\mathcal{A},\mathcal{A}} - P_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*}\|_{\infty\to1} \leq (2k\delta^{1/3} \cdot n) \cdot (2n) = 4k\delta^{1/3}n^2 = o(n^2)$. By putting everything together, we finally get that

$$\|Q_{k_{\mathcal{A}}}^{\mathcal{A}^*,\mathcal{A}^*} - 2\epsilon \cdot P_{k_{\mathcal{A}}}^{\mathcal{A},\mathcal{A}}\|_{\infty\to1} \leq k \cdot f + 4k\delta^{1/3}n^2 \leq 16k \cdot n\sqrt{n} + 2k \cdot B + 4k\delta^{1/3}n^2 \cdot \epsilon n^2 = o(\epsilon n^2).$$

□

We are now ready to extend the previous lemma to any recursive invocation of Recursive-Clust.

Theorem C.9. *Consider a generic invocation of* Recursive-Clust$(\mathcal{S}, k', f, \gamma)$ *originated from the first invocation of* Recursive-Clust$([n], k, 16n\sqrt{n} + 2B, 1)$*, and let* $\delta := 4k \cdot \sqrt{\frac{f}{\epsilon n^2}}$*. With high probability* $\geq 1 - o(1)$*:*

- *for each cluster* C*, either* $|C \cap \mathcal{S}| \leq o(n)$ *or* $|C \cap \mathcal{S}| \geq \frac{n}{k} - o(n)$*, and there are exactly* k' *clusters satisfying the second condition;*
- *let* $\mathcal{S}^*$ *be the union of the* k' *clusters having* $\geq \frac{n}{2k}$ *elements in* $\mathcal{S}$*, let* $Q_{k'}$ *be the matrix obtained from* Q *by multiplying the negative entries by* γ*, and let* $P_{k'}$ *be the analogous matrix obtained from* P*. Then,*

$$\|Q_{k'}^{\mathcal{S},\mathcal{S}} - 2\epsilon \cdot P_{k'}^{\mathcal{S}^*,\mathcal{S}^*}\|_{\infty\to1} \leq f = o(\epsilon n^2).$$

- *if* $k' > 1$*, the first sampled threshold* t *does not satisfies the condition of line 12 in Algorithm 3, so the algorithm does not fail.*

Proof. We show the Theorem by induction on the number of recursive calls each invocation of Recursive-Clust comes from.

Base Case. We start with the first invocation, i.e. Recursive-Clust$([n], k, 16n\sqrt{n} + 2B, 1)$. First, each one of the k clusters has $\frac{n}{k}$ elements in common with $[n]$. Second, by Lemma C.4, it holds $\|Q - 2\epsilon \cdot P\|_{\infty\to1} \leq 16n\sqrt{n} + 2B = o(\epsilon n^2)$, as desired. Finally, by Lemma C.8, with high probability $1 - o(1)$ the algorithm samples an appropriate threshold t and it does not fail. Thus, everything holds in the first invocation of Recursive-Clust.

Inductive Step: from $(\mathcal{S}, k', f, \gamma)$ to $(\mathcal{S}', k'', f', \gamma')$. This follows the exact same steps of the proof of Lemma C.8, which can also be seen as a special case, proving the inductive step from the first invocation of

Recursive-Clust to its direct calls. We quickly go through these steps. Let $n' := |\mathcal{S}|$. We start from the fact that $Q_{k'}$, thanks to the scaling by γ', becomes positive semidefinite. Moreover, let $\mathcal{S}^*$ be defined as in the statement of the lemma. By inductive hypothesis we get that, with high probability, $\|Q_{k'}^{\mathcal{S},\mathcal{S}} - 2\epsilon \cdot P_{k'}^{\mathcal{S}^*,\mathcal{S}^*}\|_{\infty\to 1}\leq f = o(\epsilon n^2)$. This is the only necessary ingredient to show that, by Lemma C.6, with high probability there exists $\boldsymbol{v}$ eigenvector of the leading eigenvalue of $P_{k'}$ such that

$$\|\boldsymbol{u}^* - \boldsymbol{v}\|^2\leq 4k' \cdot \sqrt{\frac{f}{\epsilon(n')^2}} = o(1).$$

We now notice that $P_{k'}$ is the positive semidefinite "correlation matrix" of a set of k' clusters with size $\frac{n}{k}$. Apart from a normalization factor that depends on k, k', it has the same eigenvectors and eigenvalues of the positive semidefinite "correlation matrix" P' of a set of k' clusters with size $\frac{n}{k'}$, so we can proceed as before, ignoring these $\Theta(k) = \Theta(1)$ normalization factors. From now on, we can proceed exactly as in the proof of Lemma C.8: first, we can use $\boldsymbol{u}^*$ to effectively proceed with the recursive calls to the sub-problems. Let $(\mathcal{S}', k'', f', \gamma')$ be the input of one of these sub-problems. In the exact same way of Lemma C.8, we get that $\|Q_{k''}^{\mathcal{S}',\mathcal{S}'} - 2\epsilon \cdot P_{k''}^{(\mathcal{S}')^*,(\mathcal{S}')^*}\|_{\infty\to 1}\leq f' = o(\epsilon n^2)$, where we have used a coherent notation on the sub-problem. The remaining properties follow exactly as in the proof of Lemma C.8. Notice that each recursive call comes from at most k chained invocations of Recursive-Clust, so all the estimates about the norms and the small probabilities (e.g., of failure of the algorithm) can be affected by a factor of $poly(k) = \Theta(1)$, which does not affect the asymptotic estimates. □

We can conclude that with high probability $\geq 1 - o(1)$ all the recursive calls are successful and that the total number of misplaced nodes is $o(n)$, achieving the desired result.

Theorem C.10. *With probability $1 - o(1)$, Algorithm 3 outputs k clusters and correctly classifies $n - o(n)$ vertices.*

Running Time. Let us analyse the running time of Algorithm 3. First, by Theorem C.10, there are $\leq k = O(1)$ recursive executions of procedure Recursive-Clust. Each execution of procedure Recursive-Clust takes time $\tilde{O}(n^2)$ if we neglect the time needed to solve the respective semidefinite program, and this follows analogously to the Spectral Algorithm. Solving $\mathcal{SDP}$ through SDP-Solver up to negligible error takes polynomial time. More precisely, since we are using the interior point method from Jiang et al. (2020) as SDP-Solver, it takes running time $O(n^6 \log(n))$ with $1/poly(n)$ error. The resulting running time is, therefore, dominated by the time needed to solve $O(k) = O(1)$ SDPs, which is $O(n^6 \log(n)) = \tilde{O}(n^6)$.

How to do without knowing ϵ. Here, we argue how we can do without the assumption of knowing the parameter ϵ. This parameter is only used to define Q (at the beginning of this section) and get rid of the eigenvector with all equal coordinates. However, we can also define Q in an alternative way as $\tilde{Q}$:

$$\tilde{Q} := Q - \epsilon\left(1 - \frac{1}{k-1}\right)\cdot \mathbb{1} = P'' - \frac{1}{2}\left(1 - \frac{1}{k-1}\right)\cdot \mathbb{1} = \frac{k}{2(k-1)}\cdot M''.$$

Then, we can get rid of the eigenvector $\mathbf{1}$ by adding an additional constraint to SDP 6, which is the following:

$$\sum_{i,j=1}^{n} \langle \boldsymbol{x}_i, \boldsymbol{y}_j\rangle = 0. \tag{9}$$

By doing so, we obtain a new SDP.

$$\begin{aligned}
\text{maximize} \quad & \sum_{j=1}^{n} \tilde{Q}_{ij}\langle \boldsymbol{x}_i, \boldsymbol{y}_j\rangle \\
\text{subject to} \quad & \sum_{i,j=1}^{n} \langle \boldsymbol{x}_i, \boldsymbol{y}_j\rangle = 0 \\
& \|\boldsymbol{x}_i\|= 1, \boldsymbol{x}_i \in \mathbb{R}^n \quad i = 1, \ldots, n \\
& \|\boldsymbol{y}_i\|= 1, \boldsymbol{y}_i \in \mathbb{R}^n \quad i = 1, \ldots, n
\end{aligned} \tag{10}$$

Now, consider an optimal solution matrix of SDP 10, and name it $\tilde{X}$ ($\tilde{X}_{i,j} := \langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle$). Eq. 9 is equivalent to $\tilde{X} \bullet \mathbb{1} = 0$, so it constrains $\tilde{X}$ to be orthogonal to the matrix $\mathbb{1} = \mathbf{1}\mathbf{1}^T$ or, equivalently, $\tilde{X}\mathbf{1} = 0$. We show that an optimal solution $\tilde{X}$ of SDP 10 also satisfies Lemma C.5.

Lemma C.11. *With high probability ($\geq 1 - 2^{-4n}$), it holds*

$$Q \bullet \tilde{X} \geq \frac{2}{k-1} \cdot \epsilon n^2 - 58n\sqrt{n} - 8B = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2).$$

Proof. First, since $\tilde{Q} = Q - \epsilon \left(1 - \frac{1}{k-1}\right) \cdot \mathbb{1}$ and $\tilde{X} \bullet \mathbb{1} = 0$, we have that

$$\tilde{Q} \bullet \tilde{X} = Q \bullet \tilde{X}. \tag{11}$$

Now, pick X' as an optimal solution for SDP 6. Then, $\tilde{X}' := \frac{1}{1 - \frac{1}{n}(X' \bullet \mathbb{1})} \left(X' - \frac{1}{n}(X' \bullet \mathbb{1})\mathbb{1}\right)$ is a feasible solution for SDP 10. By optimality of $\tilde{X}$, this implies that $\tilde{Q} \bullet \tilde{X}' \leq \tilde{Q} \bullet \tilde{X}$ but, from Eq. 11, we can derive the following

$$Q \bullet \tilde{X} = \tilde{Q} \bullet \tilde{X} \geq \tilde{Q} \bullet \tilde{X}' = Q \bullet \tilde{X}'. \tag{12}$$

However, by definition of $\tilde{X}'$,

$$Q \bullet \tilde{X}' = \frac{Q \bullet X'}{1 - \frac{1}{n}(X' \bullet \mathbb{1})} - \frac{(Q \bullet \mathbb{1}) \cdot \frac{1}{n}(X' \bullet \mathbb{1})}{1 - \frac{1}{n}(X' \bullet \mathbb{1})}.$$

We now need to provide a lower bound to the RHS of this last equation. Since $\|Q - 2\epsilon \cdot P\|_{SDP} \leq 29n\sqrt{n} + 4B = o(\epsilon n^2)$ (Lemma C.5), $P \bullet \mathbb{1} = 0$, and $\mathbb{1}$ is a feasible solution to SDP 6, it holds

$$Q \bullet \mathbb{1} = (Q - 2\epsilon \cdot P) \bullet \mathbb{1} \leq \|Q - 2\epsilon \cdot P\|_{SDP} \leq 29n\sqrt{n} + 4B = o(\epsilon n^2). \tag{13}$$

Moreover, by Lemma C.5 and by optimality of X', we have that

$$Q \bullet X' \geq \frac{2}{k-1} \cdot \epsilon n^2 - 29n\sqrt{n} - 4B = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2). \tag{14}$$

Finally, $\frac{1}{n}(X' \bullet \mathbb{1}) \leq o(1)$ by Lemma C.6, because almost all the eigenvectors of X', weighted by their eigenvalues, are nearly orthogonal to the vector $\mathbf{1}$. Therefore, we can say that, for sufficiently large n,

$$\frac{1}{n}(X' \bullet \mathbb{1}) \leq \frac{1}{2}. \tag{15}$$

By substituting Eq. 12, 13, 14 and 15 into Eq. 11, we finally get that

$$Q \bullet \tilde{X} \geq Q \bullet \tilde{X}' \geq Q \bullet X' - (Q \bullet \mathbb{1}) \geq \frac{2}{k-1} \cdot \epsilon n^2 - 58n\sqrt{n} - 8B = \frac{2}{k-1} \cdot \epsilon n^2 - o(\epsilon n^2).$$

□

By the previous lemma, the proofs for the SDP-based algorithm follow analogously. This shows that our approach still holds without assuming the knowledge of the parameter ϵ.

C.3 Going Beyond Equinumerous Clusters for Algorithm Recursive-Clust

We show that our SDP-based algorithm and its theoretical guarantees still hold when all the communities have size $n/k + o(n)$.

In detail, recall that the parameters of the post-adversarial setting satisfy $\epsilon = \omega(1/\sqrt{n}), B = o(\epsilon n^2)$. Given a ground-truth $k-$clustering with all the clusters having size $n/k + o(n)$, we can move $o(n)$ points to a different cluster for each of the k clusters, and obtain an equinumerous $k-$clustering. This is equivalent to changing $B' = o(n^2)$ entries in the zero-error matrix M associated to the ground-truth clustering, to obtain a new matrix $\widehat{M}$ representing a close equinumerous clustering. It now suffices to reconstruct the clustering associated to $\widehat{M}$ with $o(n)$ misclassified vertices, since at most other $o(n)$ errors are made when considering $\widehat{M}$ instead of M.

Now, recall that the random perturbation is equivalent to leaving each entry of the matrix unchanged with probability 2ϵ, and replacing it with a fresh random bit with probability $1-2\epsilon$. Hence, the random perturbation turns into random bits a $1-\Theta(\epsilon)$ fraction of the newly modified elements with high probability by Theorem 3.2. Thus, only $B' = o(\epsilon n^2)$ entries of the perturbed matrix M' follow a different distribution from the ones of the matrix $\widehat{M}'$, which is obtained from $\widehat{M}$ in the same way as M' from M. As a consequence, the perturbation of M in the post-adversarial setting with parameters ϵ, B following Eq. 3, is equivalent to a post-adversarial perturbation of $\widehat{M}$ with parameters $\epsilon, B+B'$. Since $B+B' = o(\epsilon n^2)$, Theorem C.10 still holds for $\widehat{M}$. Thus, we can reconstruct the corresponding clustering with $o(n)$ misclassified vertices with high probability.