GRAPH CLUSTERING AND THE NUCLEAR WASSERSTEIN METRIC.

DANIEL DE ROUX AND MAURICIO VELASCO

ABSTRACT. We study the problem of learning the cluster structure of a random graph $\mathcal{G}$ from an independent sample. We propose a Wasserstein robust formulation of this optimization problem and prove that it can be reformulated as a tractable convex optimization problem. We give theoretical exact recovery guarantees for this problem when the Wasserstein metric is induced by the nuclear norm and $\mathcal{G}$ is distributed according to the stochastic block model. Finally we present our Julia implementation of the proposed algorithm and show its numerical performance on synthetic data.

1. INTRODUCTION

Let $\mathcal{G}$ be a random graph with n vertices. By a deterministic summary of $\mathcal{G}$ we mean a (deterministic) graph H^* which, on average, differs from $\mathcal{G}$ by as few edges as possible. In this article we study the problem of finding deterministic summaries *from an independent sample* of $\mathcal{G}$ of size N. More precisely we address the following problem:

Problem 1.1. *Given adjacency matrices $B_1, \dots, B_N$ of an independent sample of $\mathcal{G}$ find a symmetric matrix A^* in* $\arg\min_A \mathbb{E}_{B\sim\mathcal{G}}[\|A - B\|_1]$.

Special cases of this problem arise in cluster detection and in data summarization, both heavily studied in the literature (see Section 1.1 for details).

A possible approach to problem 1.1 is to use the samples to construct the empirical measure $\hat{\mu} := \sum_{i=1}^{N} \frac{1}{N}\delta_{B_i}$ as an approximation of the distribution of $\mathcal{G}$ and to find a minimizer A of the resulting empirical risk

$$\mathbb{E}_{B\sim\mu}[\|A - B\|_1] = \frac{1}{N}\sum_{i=1}^{N} \|A - B_i\|_1.$$

This approach is consistent and will lead to an optimal solution as the sample size $N \to \infty$. However, when the sample size N is not sufficiently large (typicaly one has few samples of very large graphs) for $\hat{\mu}$ to be a good approximation for the distribution of $\mathcal{G}$ this approach leads to overfitting. To mitigate this problem we propose a robust version of Problem 1.1. In the robust version one aims to minimize the worst-case risk when the distribution of B is allowed to vary in a ball $\mathcal{N}_\delta(\hat{\mu})$ of radius $\delta > 0$ centered at the empirical measure $\hat{\mu}$ in a suitable metric, leading to

2000 *Mathematics Subject Classification.* Primary 15A29 Secondary 15B52,52A22.

Key words and phrases. Compressed sensing, truncated moment problems, Kostlan-Shub-Smale polynomials.

Problem 1.2. *Given adjacency matrices $B_1, \ldots, B_N$ of an independent sample of $\mathcal{G}$ find a symmetric matrix $\overline{A}$ which minimizes the robust worst-case risk*

$$R_\delta(A) := \left(\sup_{\nu \in \mathcal{N}_\delta(\hat{\mu})} \mathbb{E}_{B \sim \mathcal{G}}[\|A - B\|_1] \right).$$

The robust worst-case risk is obviously dependent on the chosen metric among probability distributions. In this article we will use the Wasserstein metric $W_{\|\bullet\|}$ induced by a norm $\|\bullet\|$ on the space K of symmetric matrices with entries in $[0, 1]$. More precisely, if ν_1, ν_2 are probability measures on K and $\Pi(\nu_1, \nu_2)$ is the set of random variables (Z_1, Z_2) taking values in $K \times K$ with $Z_i \sim \nu_i$ then the Wasserstein distance between ν_1 and ν_2 is given by

$$W_{\|\bullet\|}(\nu_1, \nu_2) := \inf_{(Z_1, Z_2) \in \Pi(\nu_1, \nu_2)} \mathbb{E}[\|Z_1 - Z_2\|].$$

The seminal work of Esfahani and Kuhn [23] shows that robust formulations defined using the Wasserstein metric often lead to tractable convex optimization problems. Our first result is that this is also true for Problem 1.2 when the Wasserstein metric is induced by any semidefinitely representable norm.

Theorem 1.3. *Let K be the set of $n \times n$ symmetric matrices with entries in $[0, 1]$, let $\|\bullet\|$ be a norm on K and let $\delta > 0$. Problem 1.2 is equivalent to:*

$$\min_{(A, s_1, \ldots, s_N, \lambda) \in \mathcal{T}} \left(\lambda\delta + \frac{1}{N} \sum_{i=1}^{N} s_i \right)$$

where $\mathcal{T}$ is the set of $(A, \vec{s}, \lambda) \in K \times \mathbb{R}^n \times \mathbb{R}$ satisfying the inequalities $\lambda \geq 0$ and $\eta_{E,i}(\lambda) \leq s_i$ as $i = 1, \ldots, N$ and E ranges over the set of all $\{-1, 1\}$-symmetric matrices, where

$$\eta_{E,i}(\lambda) := \sup_{Y \in K} \left(\langle E, A - Y \rangle - \lambda \|B_i - Y\| \right)$$

The reformulation in Theorem 1.3 transforms the problem into a finite-dimensional convex optimization problem which unfortunately contains an exponential number of constraints. To address this problem we introduce:

(1) A more tractable simplification which agrees with the original problem whenever the optimal $\overline{A}$ occurs at a matrix with entries in $\{0, 1\}$.
(2) A relaxation of (1) which takes the form of a regularized empirical risk minimization problem.

leading to practical algorithms for graph summarization. More specifically, if $\mathcal{S}^{n \times n}$ denotes the space of real symmetric $n \times n$ matrices we prove the following

Theorem 1.4. *If A is a symmetric $\{0, 1\}$-matrix and $\delta > 0$ then:*

(1) The following equality holds:

$$R_\delta(A) = \min_{\mathcal{H}} \left(\lambda\delta + \frac{1}{N} \sum_{i=1}^{N} s_i + \frac{1}{N} \sum_{i=1}^{N} \|A - B_i\|_1 \right)$$

where $\mathcal{H}$ *is the set of* $(\lambda, s_1, \ldots, s_N, \Lambda, W) \in \mathbb{R} \times \mathbb{R}^N \times \mathcal{S}^{n\times n} \times \mathcal{S}^{n\times n}$ *which satisfy the inequalities:*

$$\begin{aligned}
&\|W\|_* \leq \lambda \\
&\Lambda \geq 0 \\
&2A - 11^t - W + \Lambda \geq 0 \\
&\|\Lambda\|_1 \leq s_i - \langle 2A - 11^t - W, B_i\rangle \text{ for } i = 1, \ldots, N
\end{aligned}$$

(2) The following inequality holds:

$$R_\delta(A) \leq \frac{1}{N}\sum_{i=1}^{N} \|A - B_i\|_1 + \delta\|2A - 11^t\|_*.$$

Moreover, if $\|\cdot\|$ *is a semidefinitely representable function then both* (1) *and the minimization of the right hand side of* (2) *for* $A \in K$ *are semidefinite programming problems.*

Solving the optimization problems in Theorem 1.4 leads to a new algorithm for estimating deterministic graph summaries which we call *Wasserstein robust graph summarization*. We carried out extensive numerical experiments applying this algorithm to a variety of graphs G distributed according to the stochastic block model and observed that the regularized problem with the Wasserstein metric *induced by the spectral norm* was able to recover the correct cluster structure using only very few samples outperforming all others. This leads us to propose the *Wasserstein robust nuclear norm summarization problem*, given by

$$\min_{A\in K}\left(\frac{1}{N}\sum_{i=1}^{N} \|A - B_i\|_1 + \delta\|2A - 11^t\|_*\right) \tag{1}$$

where the nuclear norm $\|B\|_*$ of a matrix B is the dual of the spectral norm and equals the sum of its singular values.

A central result of this article is an exact recovery guarantee for this algorithm. More precisely, we show that exact recovery occurs for suitable $\delta > 0$ with overwhelming probability on samples distributed according to the stochastic block model, explaining the good practical performance of Wasserstein nuclear norm summarization in cluster detection. In order to describe our exactness guarantee we need to establish notation for the parameters of the stochastic block model. Recall that a random graph $\mathcal{G}$ has distribution given by the stochastic block model in n vertices if there is a partition of $[n]$ into disjoint subsets $C_1, \ldots, C_k$, real numbers $0 \leq q < \frac{1}{2} < p_1, \ldots, p_k \leq 1$ and edges are added independently with probability p_{ij} of joining vertices i, j given by

$$p_{ij} := \begin{cases} p_t, \text{ if } \{i,j\} \subseteq C_t \\ q, \text{ else.} \end{cases}$$

Such a random graph $\mathcal{G}$ has a unique deterministic summary A^* obtained by putting edges only between vertices belonging to the same cluster.

Theorem 1.5. *Suppose $B_1, \dots, B_N$ are independent and are distributed according to $\mathcal{G}$. If $\alpha = \min(|p_t - \frac{1}{2}|, |q - \frac{1}{2}|)$ and δ^* is the maximum of $a(\delta) := \frac{\delta\left(\alpha - \frac{\delta}{n}\right)^2}{\left(1+\frac{2\delta}{n}\right)}$ in $[0, \alpha n]$ then the probability that A^* is not the unique minimizer of (1) is bounded above by*

$$\exp\left(-\frac{2N(n-1)^3(\alpha n - \delta^*)^2}{n}\right) + e^{-N(\delta^* a(\delta^*))} \prod_{i \neq j} \left(1 + (e^{-4a}\widetilde{p_{ij}} + \widetilde{q_{ij}})^{\frac{N}{2}}\right).$$

which decreases exponentially with the sample size N.

The key point of the proof of Theorem 1.5, discussed at length in Section 3, is that the subdifferential of the regularization term at A^* is sufficiently rich so as to contain enough transportation matrices. This gives a geometric explanation for why the nuclear norm is a good regularizer for graph summarization problems.

Finally we focus on the practical performance of the proposed algorithm. Solving the optimization problems appearing in Theorem 1.4 typically require solving large semidefinite programs which are beyond the capacity of standard off-the-shelf software even for relatively small graphs (of say 40 vertices with $N = 4$). One possible reason is that off-the-shelf solvers often use interior point methods, which are highly accurate but often do not scale well. A better alternative, especially well suited for Wasserstein robust nuclear norm summarization is to use first order numerical optimization methods such as the alternating direction method of multipliers (ADMM). In Section 5 we adapt the ADMM algorithm to our regularized problem and present our open source Julia implementation. This implementation can run the Wasserstein nuclear norm summarization algorithm in graphs of up to 10000 vertices and $N \leq 10$ on a common laptop.

1.1. **Relation to previous work.** Understanding structural properties of graphs is a problem of high interest since graphs appear naturally in many areas. The problem of finding communities in a graph is one of the most studied tasks in machine learning, with applications to Biology [7, 14, 29], social data understanding [16, 22, 25] and other general machine learning tasks such as natural language processing [15, 27]. One of the reasons of the importance of this problem is that it allows researchers to organize datasets into sets of similar observations, an then apply learning algorithms on each of the sets. Most theoretical results in the community detection problem occur in the context of generative random graph models. Of these the most studied is the stochastic block model [1] which is particularly important since in it communities are defined a priori and thus there is a formally specified underlying true cluster structure with which the output of algorithms can be compared.

There are three main approaches for the problem of community detection on a (single) graph:

(1) Spectral clustering algorithms, widely used in the detection of sparse communities [5, 13, 20, 21].
(2) SDP approaches based on the seminal work of Goemans and Williamson [18], [2, 19, 24].

(3) Decompositions of the adjacency matrix of the graph into a matrix of low rank and a sparse (noise) matrix, using a convex relaxation based on the nuclear norm [8, 10, 9]. These methods have led to several convex optimization algorithms to find communities in graphs [4, 28, 12, 11, 26, 3].

The results in this article differ from previous work in two main accounts:

(1) Our methods apply to the problem of learning from *several* samples while, to the best of our knowledge, the above algorithms must be applied apply to a single graph. Practicioners resort to several "graph combination" methods to transform datasets consisting of several samples into one. Our methods do not require this additional pre-processing step and are able to use the additional information contained in several samples to obtain more accurate recovery (see Section 5)
(2) Formulating learning problems as Wasserstein robust optimization problems leads to regularization algorithms in a principled way. Even in the single-sample case this leads to improvements in performance when compared for instance with the sparse+low-rank approach (see Section 5).

1.2. **Acknowledgments.** We wish to thank Fabrice Gamboa, Mauricio Junca and Thierry Klein for useful conversations during the completion of this project. D. De Roux was partially supported by a grant from Facultad de Ciencias, Universidad de los Andes and by CAOBA. M. Velasco was partially supported by research funds from Universidad de los Andes.

2. Robust learning of deterministic summaries

Let $[n] := \{1, 2, \dots, n\}$. By a graph G on $[n]$ we mean a finite loopless undirected graph with vertex set $[n]$. The adjacency matrix of such a graph is a symmetric $n \times n$ matrix with entries in $\{0, 1\}$ and ones in positions (i, j) whenever vertices i and j are adjacent. Let $\mathcal{A}_n$ be the set of adjacency matrices of graphs on $[n]$ (i.e. $\{0, 1\}$-matrices of size $n \times n$ which are symmetric and have zeroes in the diagonal). Throughout the article we will use graphs and their adjacency matrices interchangeably. By a random graph $\mathcal{G}$ on $[n]$ we mean a random variable B taking values in $\mathcal{A}_n$. For a matrix B we let $\|B\|_p$ to be the p-th root of the sum of the p-th powers of its entries. We denote by $\langle A, B\rangle := \mathrm{Tr}(AB)$ the Frobenius inner product on symmetric matrices.

Definition 2.1. *A deterministic summary of a random graph $\mathcal{G}$ on $[n]$ is a graph A^* with $A^* \in \arg\min_{A\in\mathcal{A}_n} \mathbb{E}[\|A - \mathcal{G}\|_1]$.*

The deterministic summary A^* is a graph which, on average, differs from $\mathcal{G}$ by the smallest possible number of edges. If the distribution of $\mathcal{G}$ is known it is easy to find a deterministic summary (which is often unique). More precisely,

Lemma 2.2. *If $\mathbb{P}\{(i, j) \in \mathcal{G}\} \neq \frac{1}{2}$ for all $i, j \in [n]$ then the unique deterministic summary of $\mathcal{G}$ is given by A^* with*

$$A^*_{ij} = \begin{cases} 1, \text{ if } \mathbb{P}\{(i, j) \in \mathcal{G}\} > \mathbb{P}\{(i, j) \notin \mathcal{G}\} \\ 0, \text{ if } \mathbb{P}\{(i, j) \in \mathcal{G}\} < \mathbb{P}\{(i, j) \notin \mathcal{G}\} \end{cases}$$

Proof. If $A \in \mathcal{A}_n$ then the term coming from entry (i,j) in $\mathbb{E}[\|A - \mathcal{G}\|_1]$ is given by $|A_{ij} - 1|p + |A_{ij}|q$ where p (resp q) is the probability that (i,j) is (resp. is not) an edge of $\mathcal{G}$. This quantity is greater than $\min(p,q)$ and equality is achieved, for all i,j when $A = A^*$. □

Motivated by the previous Lemma we define a cluster structure on a random graph.

Definition 2.3. *A random graph $\mathcal{G}$ has a cluster structure or a community structure if it has a unique deterministic summary A^* and the corresponding graph is a disjoint union of cliques. We call these cliques the clusters of $\mathcal{G}$.*

The main problem that we address in this article is that of *learning* deterministic summaries (and in particular the problem of learning cluster structures on random graphs who have them). By this we mean that our only knowledge about the distribution of the random graph $\mathcal{G}$ is encoded in an independent sample $B_1, \ldots, B_N$ of adjacency matrices with the same distribution as $\mathcal{G}$, leading to Problem 1.1 in the Introduction.

Given the sample, define the empirical measure $\hat{\mu} := \frac{1}{N}\sum_{j=1}^{N} \delta_{B_j}$ as a sum of Dirac delta measures at the sample points. As the number of sample points increases the measure $\hat{\mu}$ converges to the distribution of $\mathcal{G}$ [17] and it is therefore reasonable to try to minimize the objective function in Problem 1.1 with respect to the measure $\hat{\mu}$ instead of $\mathcal{G}$, that is by finding a minimizer of the empirical risk

$$\overline{A} \in \arg\min_{A \in K} \mathbb{E}_{Z \sim \hat{\mu}}[\|A - Z\|]$$

Arguing as in Lemma 2.2 it is immediate that a (generally unique) minimizer $\overline{A}$ is given by counting edge frequencies, that is

$$\overline{A}_{ij} := \begin{cases} 1, \text{ if } |\{t : B_{ij}^{(t)} = 1\}| > |\{t : B_{ij}^{(t)} = 0\}| \\ 0, \text{ if } |\{t : B_{ij}^{(t)} = 1\}| < |\{t : B_{ij}^{(t)} = 0\}| \end{cases}$$

The empirical risk minimization approach has lots of advantages, it is easy to implement, scales very well and is guaranteed to be consistent (in the sense that $\overline{A}_{ij} \to \overline{A}$ as the number of samples $N \to \infty$). However it also suffers from some potential drawbacks:

(1) If the sample size N is small then the empirical measure $\hat{\mu}$ could be very far from the distribution of $\mathcal{G}$.
(2) The estimation of $\overline{A}$ is done independently edge by edge and in particular it does not use any global information, for instance the existence of a cluster structure as part of the estimation process.

To mitigate these problems we will use a robust formulation. To this end let $\| \bullet \|$ be a norm in the space of symmetric $n \times n$ matrices and let $W_{\|\bullet\|}$ be the Wasserstein distance induced by a given norm $\| \bullet \|$ on K. For a real number $\delta \geq 0$ let $\mathcal{N}_\delta(\hat{\mu})$ be the (closed) ball of radius δ centered at $\hat{\mu}$. We would like to solve the following

Problem. *Given adjacency matrices $B_1, \ldots, B_N$ of an independent sample of $\mathcal{G}$ find a symmetric matrix $\overline{A}$ which minimizes the robust worst-case risk*

$$R_\delta(A) := \left(\sup_{\nu \in \mathcal{N}_\delta(\hat{\mu})} \mathbb{E}_{B \sim \mathcal{G}}[\|A - B\|_1] \right).$$

Our first result reformulates the robust optimization above as a finite convex optimization problem by closely following the ideas of Esfahani and Kuhn. The proof is included for the reader's benefit.

Proof of Theorem 1.3. Let $\mathcal{M}$ be the set of probability measures in $K \times K$ whose marginal distribution in the first component is given by the empirical measure $\hat{\mu}$. Every such measure is of the form

$$\nu = \sum_{i=1}^{N} \frac{1}{N} \delta_{B_i} \otimes \mathbb{Q}_i$$

where $\mathbb{Q}_i$ is a probability measure on K. Conditioning on the value of X it is immediate that these measures act on functions on two sets of variables corresponding to each copy of K via the formula

$$\int_{K \times K} f(x,y) d\nu = \sum_{i=1}^{N} \frac{1}{N} \int_K f(B_i, y) d\mathbb{Q}_i(y).$$

The distribution of a random variable $Y \in K$ lies in the Wasserstein ball $\mathcal{N}_\delta(\hat{\mu})$ iff there exists a random vector $(X, Y) \in K \times K$ with distribution $\nu \in \mathcal{M}$ which satisfies $\int_{K \times K} \|X - Y\| d\nu \leq \delta$. This description will allow us to optimize over $\mathcal{N}_\delta(\hat{\mu})$. If $A \in K$ then the following equalities hold:

$$R_\delta(A) = \sup_{\nu \in \mathcal{M}} \left(\int_{K \times K} \|A - Y\|_1 d\nu : \int_{K \times K} \|X - Y\| d\nu \leq \delta \right)$$

which can be written as

$$\sup_{\nu \in \mathcal{M}} \inf_{\lambda \geq 0} \int_{K \times K} \|A - Y\|_1 d\nu + \lambda \left(\delta - \int_{K \times K} \|X - Y\| d\nu \right)$$

By strong duality for moment problems this quantity equals

$$\inf_{\lambda \geq 0} \sup_{\nu \in \mathcal{M}} \int_{K \times K} \|A - Y\|_1 d\nu + \lambda \left(\delta - \int_{K \times K} \|X - Y\| d\nu \right)$$

and conditioning on the value of X we compute the integrals obtaining

$$\inf_{\lambda \geq 0} \lambda \delta + \frac{1}{N} \sup_{\mathbb{Q}_1, \ldots, \mathbb{Q}_N} \left(\int_K \|A - Y\|_1 - \lambda \|B_i - Y\| d\mathbb{Q}_i(Y) \right).$$

Since the integrals are maximized when the $\mathbb{Q}_i$ are Dirac deltas at optima of their corresponding functions we conclude that this quantity equals

$$\inf_{\lambda \geq 0} \lambda \delta + \frac{1}{N} \sup_{Y \in K} \left(\|A - Y\|_1 - \lambda \|B_i - Y\| \right)$$

Next we move the supremum into the constraints obtaining

$$\inf_{(\lambda,s_1,\ldots,s_n)} \left(\lambda\delta + \frac{1}{N}\sum_{i=1}^{N} s_i : \sup_{Y\in K} \left(\|A-Y\|_1 - \lambda\|B_i - Y\| \right) \leq s_i \text{ , for } i = 1,\ldots,N \right).$$

Finally, if L denotes the set of symmetric matrices with entries in $\{-1,1\}$ then the following equalities hold for every symmetric matrix B

$$\|B\|_1 = \sup_{H:\|H\|_\infty \leq 1} \langle B, H\rangle = \max_{E\in L} \langle B, E\rangle.$$

Defining the concave function

$$\eta_{E,i}(\lambda) := \sup_{Y\in K} \left(\langle E, A-Y\rangle - \lambda\|B_i - Y\| \right)$$

we therefore conclude that

$$R_\delta(A) = \inf_{(\lambda,s_1,\ldots,s_n)} \left(\lambda\delta + \frac{1}{N}\sum_{i=1}^{N} s_i : \eta_{i,E}(\lambda) \leq s_i \text{ for all } E\in L \text{ and } i = 1,\ldots,N \right).$$

Letting A vary in K we prove the Theorem. □

Our second results gives a simplification of the previous problem which agrees with it when problem 1.2 has an optima which occurs at a matrix with entries in $\{0,1\}$. The proof is similiar to the proof of 1.3.

Proof of Theorem 1.4. If the symmetric matrix A has entries in $\{0,1\}$ and X, Y are matrices in K then the following equalities hold

$$\|A-Y\|_1 = \langle 2A - 11^t, A-Y\rangle = \|A-X\|_1 + \langle 2A-11^t, X-Y\rangle$$

Now suppose (X,Y) is a random vector taking values in $K\times K$ with distribution ν in the set $\mathcal{M}$ of the proof of Theorem 1.3. Taking expected values and suprema we see that

$$R_\delta(A) = \frac{1}{N}\sum_{i=1}^{N} \|A-B_i\|_1 + \sup_{\nu\in\mathcal{M}} \int_{K\times K} \langle 2A-11^t, X-Y\rangle d\nu \tag{2}$$

Since $\langle 2A-11^t, X-Y\rangle \leq \|2A-11^t\|_* \|X-Y\|$ by definition of dual norm and $\int_{K\times K}\|X-Y\|d\nu \leq \delta$ for all Y whose distribution is in the Wasserstein ball $\mathcal{N}_\delta(\hat{\mu})$ we conclude that

$$R_\delta(A) \leq \frac{1}{N}\sum_{i=1}^{N} \|A-B_i\|_1 + \delta\|2A-11^t\|_*$$

proving the inequality claimed in part (2) of the Theorem. Repeating the argument used in the proof of Theorem 1.3 we also conclude from equation (2) that

$$R_\delta(A) = \inf_{\lambda\geq 0} \left(\lambda\delta + \frac{1}{N}\sum_{i=1}^{N} \|A-B_i\|_1 + \frac{1}{N}\sum_{i=1}^{N} s_i \right)$$

where

$$\sup_{Y\in K} \left(\langle 2A-11^t, B_i - Y\rangle - \lambda\|B_i - Y\| \right) \leq s_i$$

Using the definition of dual norm this inequality can be rewritten as

$$\sup_{Y\in K}\inf_{\|W\|_*\leq\lambda}\langle 2A-11^t-W,B_i-Y\rangle\leq s_i$$

which by duality is equivalent to

$$\inf_{\|W\|_*\leq\lambda}\sup_{Y\in K}\langle 2A-11^t-W,B_i-Y\rangle\leq s_i$$

which in turn is equivalent to the existence of a symmetric matrix W with $\|W\|_*\leq\lambda$ such that

$$\sup_{Y\in K}\langle 2A-11^t-W,B_i-Y\rangle\leq s_i.$$

Moreover, letting $D=2A-11^t-W$ the previous inequality can be rewritten as

$$\sup_{Y\in K}\langle D,-Y\rangle\leq s_i-\langle D,B_i\rangle$$

and we will compute the supremum of the left-hand side using linear programming duality on the hypercube K. In $\mathcal{S}^{n\times n}$ the set K is defined by the restrictions $0\leq Y\leq 11^t$ and thus the Lagragian of our maximization problem is given by

$$L=\langle D,-Y\rangle+\langle\Lambda,11^t-Y\rangle+\langle\eta,Y\rangle=\langle D+\Lambda-\eta,-Y\rangle+\langle\Lambda,11^t\rangle$$

where the symmetric matrices $\Lambda,\eta\geq 0$. It follows that

$$\sup_{Y\in K}\langle -D,Y\rangle=\sup_{Y\in\mathcal{S}^{n\times n}}\inf_{\Lambda,\eta\geq 0}L$$

and exchanging infimum and supremum, we obtain

$$\sup_{Y\in K}\langle -D,Y\rangle=\inf_{\Lambda,\eta\geq 0}\sup_{Y\in K}L=\inf_{\Lambda,\eta\geq 0}\begin{cases}\langle\Lambda,11^t\rangle\text{ if }D+\Lambda+\eta=0\\ \infty\text{ otherwise.}\end{cases}$$

which is equivalent to $\inf_{\Lambda\geq 0}\|\Lambda\|_1$ subject to $\Lambda,D+\Lambda\geq 0$, proving part (1) of the Theorem. If $\|\bullet\|$ is a semidefinitely representable function then so is the dual norm and therefore the feasible set of the optimization problem is the intersection of an SDr set and a polyhedron and thus an SDr set. Since the objective function is linear it is therefore a semidefinite programming problem as claimed. □

Remark 2.4. We believe that the inequality in part (2) of the Theorem is in fact an equality under very mild assumptions. The idea that regularization algorithms naturally arise from Wasserstein robust formulations has appeared in the literature, see for instance [**?**].

Motivated by the inequality in part (2) of Theorem 1.4 we define the following semidefinite-programming algorithm for graph summarization given an independent sample $B_1,\ldots,B_N$ of $\mathcal{G}$,

Definition 2.5. *Given $\delta>0$, the Wasserstein robust nuclear norm graph summarization (WRNS) algorithm consists of finding a matrix $\overline{A}\in\arg\min_{A\in K}\Delta(A)$ where*

$$\Delta(A):=\delta\|2A-11^t\|_*+\frac{1}{N}\sum_{i=1}^{N}\|A-B_i\|_1$$

and $\|B\|_*$ denotes the nuclear norm of a matrix B defined as the sum of its singular values (since B is symmetric this coincides with the sum of the absolute values of the eigenvalues of B).

3. Exact community detection in the stochastic block model and the Wasserstein nuclear norm.

In this Section we will prove that there is a choice of $\delta = \delta(N)$ such that the WRNS algorithm recovers the exact cluster structure of a random graph $\mathcal{G}$ generated with the stochastic block model with very high probability.

More precisely suppose $C_1, \ldots, C_l$ are a partition of the set $[n]$ and the real numbers p_i, q satisfy the inequalities $0 \leq q < \frac{1}{2} < p_i \leq 1$ for $i = 1, \ldots, l$. We say that a random graph $\mathcal{G}$ with vertex set $[n]$ is generated by the stochastic block model determined by the clusters $C_1, \ldots, C_l$ and the numbers $p_i, \bar{p}$ if the edges of $\mathcal{G}$ are independent random variables and an edge joins vertices i, j with probability p_t if $\{i, j\} \subseteq C_t$ for some cluster C_t and with probability q if $\{i, j\}$ is not contained in any C_t. By Lemma 2.2 the graph $\mathcal{G}$ has a unique cluster (or community) structure A^* with entries in $\{0, 1\}$ given by $A^*_{ij} = 1$ iff $\{i, j\} \subseteq C_t$ for some cluster t. Given a random sample $B_1, \ldots, B_N$, a (deterministic) graph summarization algorithm produces a symmetric matrix $\overline{A}(B_1, \ldots, B_N)$ which is a random variable due to the randomness of the sample. We would like to understand how often does the algorithm recover the exact cluster structure, equivalently we would like to understand the probability that $\overline{A} = A^*$.

Our analysis of the WRNS algorithm will consist of two steps. In Section 3.1 we will find geometric conditions on the subdifferential of $\Delta(A)$ at A^* which guarantee that A^* is the unique solution of the optimization problem. These conditions provide a geometric explanation of the usefulness of the regularizer. These geometric conditions will depend on the sample and in Section **??** we will show that the probability that these conditions fail to hold decreases exponentially on the sample size N when δ is appropriately chosen.

3.1. Geometry of subdifferentials and sufficient conditions for exactness.

Let h be a real valued convex function on a vector space V with an inner product. The subdifferential of h at a point $a \in \mathrm{dom}(h)$ is the closed convex set given by

$$\partial h(a) := \{g \in V : \forall z \in \mathrm{dom}(h)\, (h(z) \geq h(a) + \langle g, z - a\rangle)\}\,.$$

Subdifferentials are a very useful tool for proving that a point x^* is a minimizer of h because this occurs if and only if $\vec{0} \in \partial h(x^*)$.

For an integer N let $B_1, \ldots, B_N$ be an independent sample of N graphs with the same distribution as $\mathcal{G}$. Recall that

$$\Delta(A) = \delta \|2A - 11^t\|_* + \frac{1}{N}\sum_{k=1}^{N} \|A - B_k\|_1$$

And define functions f, g:

$$f := \frac{1}{N}\sum_{k=1}^{N} \|A - B_k\|_1, \quad g := \delta\|2A - 11^t\|_*$$

For vertices $i, j \in [n]$ let $n_1(ij)$ (resp. $n_0(ij)$) be the random variable which counts the number of times that a given pair is (resp is not) an edge of some B_j, $j = 1, \ldots N$ and let Γ be the symmetric matrix with zero diagonal and off-diagonal entries given

by $\Gamma_{ij} = \frac{n_0(ij)-n_1(ij)}{N}$. The following Lemma summarizes basic subdifferentiability properties of our function $\Delta(A)$ at A^*.

Lemma 3.1. *The following statements hold:*

(1) The subdifferential of f at A^* is the set of symmetric matrices C satisfying the inequalities

$$\Gamma_{ij} = \frac{n_0(ij) - n_1(ij)}{N} \leq C_{ij} \leq 1, \text{ if } \{i,j\} \subseteq C_t \text{ for some } t,$$

$$-1 \leq C_{ij} \leq \Gamma_{ij} = \frac{n_0(ij) - n_1(ij)}{N} \text{ if } \{i,j\} \text{ does not belong to any cluster and}$$

$$-N \leq C_{ii} \leq N \text{ for all } i.$$

(2) The subdifferential of g at A^* is the set of symmetric matrices of the form $2\delta C$ where C has spectral norm $\|C\| \leq 1$ and satisfies $\langle C, 2A - 11^t\rangle = n$.

Proof. (1) Since the subdifferential is additive it suffices to understand the subdifferential of the absolute value. If $i,j \in C_t$ then $A^*_{ij} = 1$ and the entry ij of the subdifferential of the sum at A^* is $[-1,1]$ for each B_i containing the edge and it is 1 for each B_i for which (ij) is not an edge. ♣♣♣ Daniel: [vale la pena mencionar el caso i=j o es obvio la propiedad en la diagonal?]♣♣♣ Mauricio: [Vale la pena porfa escribirla] If i,j is not contained in any cluster then $A^*_{ij} = 0$ and the entry ij of the subdifferential of the sum at A^* is -1 for each B_i which contains the edge ij and $[-1,1]$ for each B_i which does not, proving the claim. (2) It is easy to prove that the subdifferential of any norm $\|\bullet\|$ at a point X is given by those C for which the dual norm $\|C\|_* \leq 1$ and $\langle C, X\rangle = \|X\|$. Claim (2) follows because $\|2A - 11^t\| = Tr(2A - 11^t) = n$ where the first equality holds since $2A - 11^t$ is positive semidefinite.♣♣♣ Mauricio: [Esto es obvio con solo dos clusters (la matriz es uu^t donde u es el vector con 1's en un cluster y -1's en el complemento pero hay que demostrarlo para tres o mas)]. □

By Lemma 3.1 a symmetric matrix C lies in the subdifferential if and only if it satisfies the inequalities

$$\Gamma_{ij} \leq C_{ij} \leq 1, \text{ if } \{i,j\} \subseteq C_t \text{ for some } t,$$

$$-1 \leq C_{ij} \leq \Gamma_{ij} \text{ if } \{i,j\} \text{ does not belong to any cluster and}$$

$$-N \leq C_{ii} \leq N \text{ for all } i.$$

To simplify these inequalities we let $I \subseteq [n] \times [n]$ be the set of pairs (i,j) with $i \neq j$ which simultaneously belong to some cluster C_t and the set O to be the set of pairs $(i,j)$with $i \neq j$ for which $\{i,j\}$ is not contained in any cluster. Define a linear operator $\widetilde{\bullet} : \mathcal{S}^{n\times n} \to \mathcal{S}^{n\times n}$ on symmetric matrices by the formula:

$$\widetilde{A} = \begin{cases} A_{ij} \text{ if } i = j \text{ or } ij \in I \text{ and} \\ -A_{ij} \text{ if } ij \in O. \end{cases}$$

In this language C_{ij} belongs to the subdifferential of f if and only if $\widetilde{\Gamma}_{ij} \leq \widetilde{C}_{ij}$ for $i \neq j$.

Our next Lemma gives sufficient conditions for the true cluster structure A^* to be a minimizer of the proposed optimization problem. In order to describe it we introduce the following notation.

Definition 3.2. *Let δ be a positive real number. For a symmetric matrix Γ define the quantities*

$$b(\Gamma,\delta) := \sum_{i\neq j} \max\left(\widetilde{\Gamma_{ij}} + \frac{2\delta}{n}, 0\right) \text{ and } a(\Gamma,\delta) := \sum_{i\neq j} \max\left(-\widetilde{\Gamma_{ij}} - \frac{2\delta}{n}, 0\right)$$

The quantity $b(\Gamma,\delta)$ (resp. $a(\Gamma,\delta)$) measures the total amount by which the matrix $\widetilde{-\frac{2\delta}{n}11^t}$ fails (resp. succeeds) to be in the subdifferential of Lemma 3.1 in the sense that it sums over all ij the amount by which the inequalities $\widetilde{\Gamma_{ij}} \leq \frac{2\delta}{n}11^t$ fail (resp. succeed). The key point of the following Theorem is that if the inequality fails by less than it succeeds then the subdifferential of the spectral norm is sufficiently rich so as to allow us to redistribute these quantities. In this sense the following Theorem explains the success of the spectral norm in cluster recovery algorithms.

Recall that

$$\Delta(A) = \delta\|2A - 11^t\|_* + \frac{1}{N}\sum_{k=1}^{N}\|A - B_k\|_1$$

And define functions f, g:

$$f := \frac{1}{N}\sum_{k=1}^{N}\|A - B_k\|_1, \quad g := \delta\|2A - 11^t\|_*$$

In this vocabulary, we have that

$$\partial(\Delta) = \partial(f)(A^*) + \partial(g)(A^*) \tag{3}$$

Theorem 3.3. *[Alternative proof of theorem 6.2] Assume there are only two clusters. Let $\delta > 0$ with $\left(\frac{\delta}{n} + b(\Gamma,\delta)\right) < \frac{N}{2}$. If $b(\Gamma,\delta) < \min(\delta, a(\Gamma,\delta))$ then A^* is a minimizer of the optimization problem $\min_A \Delta(A)$.*

Proof. We will show that there exists a matrix C such that $-C \in \partial(g)(A^*)$ for which $\widetilde{\Gamma}_{ij} \leq \widetilde{C_{ij}}$ for $i \neq j$. This implies that $C_{ij} \in \partial(f)(A^*)$ and therefore $0 = C - C$ belongs to the subdifferential $\partial(\Delta)(A^*)$ and thus A^* is a minimizer of $\Delta(A)$.

Recall that $-C \in \partial(g)(A^*)$ if and only if

$$\langle H^*, C\rangle = -2\delta n \text{ and } \|C\| \leq 2\delta$$

where $H^* = 2A - 11^t$ and $\|\bullet\|$ is the spectral norm.

Notice that both these conditions are satisfied by setting $C^0 = -\frac{2\delta}{n}H^*$ as

$$\frac{-2\delta}{n}\langle H^*, H^*\rangle = \frac{-2\delta}{n}n^2 = -2\delta n \text{ and } \|C^0\| = \frac{2\delta}{n}\|H^*\| = 2\delta.$$

However this choice of C^0 will not, in general, satisfy the inequalities $\widetilde{\Gamma}_{ij} \leq \widetilde{-\frac{2\delta}{n}H^*}_{ij}$ for $i \neq j$. Therefore, we will correct our candidate matrix C^0 so that it satisfies these inequalities and still belongs to $\partial(g)(A^*)$.

To do this, we will construct a matrix K and add it to C^0. Crucially, K will satisfy that $KH^* = H^*K = 0$ so that we can control the spectral norm of $C^0 + K$.

Let $i < j$ and define the symmetric matrix e^{ij} as follows:

$$e^{ij} = \begin{cases} 1 \text{ if } (i,j) \in I. \\ -1 \text{ if } (i,j) \in O. \\ -1 \textit{ in the entry } ii \text{ and in } jj. \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

Observe that for any (i,j) the matrix e^{ij} satisfies the following properties:

(1) $H^*e^{ij} = 0 = e^{ij}H^*$.
(2) The inequality $\| - e^{ij} + e^{st}\| \leq 2$ holds for all ij and st. This is immediate noting that the spectral norm is bounded by the product of the induced $1-norm$ and the induced $\infty-$norm. (The inequality is strict only if $|\{i,j\} \cap \{s,t\}| \geq 1$ and in this case it can take values of $\sqrt{3}$ and 0).

The idea is to use that $b(\Gamma,\delta) < a(\Gamma,\delta)$ so there exists a way to redistribute the quantity $b(\Gamma,\delta)$ by subtracting it from the ij for which $\frac{2\delta}{n} < \widetilde{\Gamma}_{ij}$ and adding it into those st for which $\widetilde{\Gamma}_{st} \leq \frac{2\delta}{n}$.

Let U be the set of entries $\{i,j\}$ where $\widetilde{\Gamma}_{ij} + \frac{2\delta}{n} \leq 0$ and V be the set of entries $\{i,j\}$ where $-(\widetilde{\Gamma}_{ij} + \frac{2\delta}{n}) > 0$. Observe that

$$b(\Gamma,\delta) = \sum_{ij \in U} (\widetilde{\Gamma}_{ij} + \frac{2\delta}{n}) \text{ and } a(\Gamma,\delta) = \sum_{ij \in V} -(\widetilde{\Gamma}_{ij} + \frac{2\delta}{n}).$$

By hypothesis $b(\Gamma,\delta) < a(\Gamma,\delta)$. Let $l_1,..,l_k$ be an enumeration of V were k is it's cardinality. For each entry $ij \in U$, there exists nonnegative coefficients $\gamma^{ij}_{l_1},..,\gamma^{ij}_{l_k}$ such that:

$$\widetilde{\Gamma}_{ij} + \frac{2\delta}{n} - \gamma^{ij}_{l_1} - ... - \gamma^{ij}_{l_k} < 0.$$

and that such that for each γ_{l_p} with $p \in \{1,..,k\}$

$$-(\widetilde{\Gamma}_{l_p} + \frac{2\delta}{n}) - \sum_{ij \in U} \gamma^{ij}_{l_p} > 0. \tag{5}$$

For $ij \in U$ define the matrix

$$W^{ij} := \gamma^{ij}_{l_1}(\widetilde{e^{ij} - e^{l_1}}) + ... + \gamma^{ij}_{l_k}(\widetilde{e^{ij} - e^{l_k}})$$

Given that $ij \in U$, $l_1,..,l_k \in V$ and the sets U and V are disjoint, the support of e^{ij} is disjoint from the support of any of the matrices $e^{l_1},..,e^{l_k}$ (except probably at the diagonal). In particular, if $ij \in I$ the entry ij of the matrix W^{ij}, namely W^{ij}_{ij} is equal to:

$$\gamma^{ij}_{l_1}(1-0) + ... + \gamma^{ij}_{l_k}(1-0) = \gamma^{ij}_{l_1} + ... + \gamma^{ij}_{l_k}.$$

and if $ij \in O$,

$$\gamma^{ij}_{l_1}(-1-0) + ... + \gamma^{ij}_{l_k}(-1-0) = -\gamma^{ij}_{l_1} - ... - \gamma^{ij}_{l_k}.$$

Define the matrix C^1 as:

$$C^1 := C^0 + \sum_{ij \in U} W^{ij}$$

We will now verify that $C^1 \in \partial(f)(A^*)$. For $ij \in V$, we have by definition of V and by equation 5 that $\widetilde{\Gamma}_{ij} < \widetilde{C}^1_{ij}$.

Let $ij \in I \cap U$. We have that $\Gamma_{ij} + \frac{2\delta}{n} > 0$. Now, the entry of C^1 in ij is given by:

$$C^1_{ij} = -\frac{2\delta}{n} H_{ij} + W^{ij}_{ij} = -\frac{2\delta}{n} + \gamma^{ij}_{l_1} + ... + \gamma^{ij}_{l_k}$$

By the construction of the γ's, we have that

$$\Gamma_{ij} + \frac{2\delta}{n} - \gamma^{ij}_{l_1} - ... - \gamma^{ij}_{l_k} < 0$$

it follows that

$$\Gamma_{ij} - C^1_{ij} < 0$$

so that

$$C^1_{ij} > \Gamma_{ij}.$$

For $ij \in O \cap U$, we have that $\frac{2\delta}{n} - \Gamma_{ij} > 0$. The entry of C^1 in ij is given by:

$$C^1_{ij} = -\frac{2\delta}{n} H_{ij} + W^{ij}_{ij} = \frac{2\delta}{n} - \gamma^{ij}_{l_1} - ... - \gamma^{ij}_{l_k}$$

Since

$$\frac{2\delta}{n} - \Gamma_{ij} - \gamma^{ij}_{l_1} - ... - \gamma^{ij}_{l_k} < 0$$

it follows that

$$-C_{ij} > -\Gamma_{ij}.$$

and that $\widetilde{\Gamma}_{ij} < \widetilde{C}^1_{ij}$ for $i \neq j$.

It remains to show that the entries of the diagonal of C^1 are bounded by n. The diagonal entries $\{ss\}$ of C^1 are given by

$$-\frac{2\delta}{n} + \sum_{ij \in U} W^{ij}_{ss}.$$

Notice that by the definition of W^{ij}, each of the matrices $(\widetilde{e^{ij} - e^{l_p}})$ has, in the worst case, a -2 in the entry ss so the entry ss of C^1 is bounded below by

$$-\frac{2\delta}{n} + -2\left(\sum_{ij \in U} \sum_{l_p \in V} \gamma^{ij}_{l_p}\right)$$

the quantity in the parentheses is all the weight that we have to distribute, i.e $b(\Gamma, \delta)$. Therefore,

$$C^1 \geq -\frac{2\delta}{n} - 2(b(\Gamma, \delta)) = -2\left(\frac{\delta}{n} + b(\Gamma, \delta)\right).$$

By hypothesis, $\frac{N}{2} > \left(\frac{\delta}{n} + b(\Gamma, \delta)\right)$ so we obtain that

$$C^1_{ss} > -N.$$

It is obvious that $N > C^1_{ss}$. We conclude that $C^1 \in \partial(f)(A^*)$.

For each ij, $H^*e^{ij} = 0 = e^{ij}H^*$ so the equality $\langle H^*, C^1\rangle = \langle H^*, -\frac{2\delta}{n}H^*\rangle = -2\delta n$ holds and moreover

$$\|C^1\| = \max\left(\left\|-\frac{2\delta}{n}H^*\right\|, \left\|\sum_{ij\in U} W^{ij}\right\|\right).$$

The operator norm of the first term in the maximum equals 2δ and that of the second term is bounded by $2b(\Gamma,\delta)$ by the triangle inequality and the definition of $b(\Gamma,\delta)$. We conclude that $\|C\|$ is bounded by 2δ because $b(\Gamma,\delta) \leq \delta$. As a result $-C \in \partial(g)(A^*)$ proving the Theorem. Note that C^1 satisfies all the inequalities that define the membership to $\partial(f)(A^*)$ in 3.1 strictly, so C^1 belongs is interior point of $\partial(f)(A^*)$.

□

We will now prove the general case when there are more than two clusters. The proof will be similar the proof of the previous theorem. We will start with the candidate matrix $\frac{-2\delta}{n}H^*$ and correct it by adding transport matrices that assure that the corrected matrix belong to the subdifferential of $\Delta(A)$ at A^*. The difficulty in applying the tools of the previous theorem to solve the general case is that the matrices K that transport weight from one cluster to another do not, in general, satisfy the relation $KH^* = H^*K = 0$ when there are more than two clusters. This problem can be solved by splitting the matrix H^* into matrices than only take into account 2 clusters, and using the previous theorem. We begin recalling the following simple result:

Claim 3.4. *Let $a_i, b_i \geq 0,\ i = 1,\ldots,p$ be two non-negative, finite sequences of real numbers such that*

$$\sum_{i=1}^{p} a_i \geq \sum_{i=1}^{p} b_i.$$

Then there exist a finite sequence of reals c_i such that

- $\sum_{i=1}^{p} c_i = 0$.
- $a_i \geq b_i + c_i\ \forall i$.

Now we proceed to do the proof. For clusters $C_s \neq C_t$ define the matrix $H^{C_sC_t}$ whose entries are given by:

$$H^{C_iC_j}_{uv} = \begin{cases} 1 \text{ if } u,v \in C_s \text{ or } u,v \in C_t. \\ -1 \text{ if } u \in C_s, v \in C_t \text{ or } u \in C_s, v \in C_t. \\ 0 \text{ in any other case.} \end{cases}$$

Notice that this matrix has 4 blocs. Two with only 1 and two with only -1. Moreover, its spectral norm is equal to $|C_s + |C_t|$.

Lemma 3.5. *Assume there are l clusters. Suppose that $b(\Gamma,\delta) < min(\delta, a(\Gamma,\delta))$. Then, A^* is a minimizer of the optimization problem $\min_A \Delta(A)$.*

Proof. First of all, observe that

$$\frac{-2\delta}{n}H^* = \frac{-2\delta}{n}\frac{1}{l-1}\sum_{1\le s<t\le l} H^{C_sC_t}.$$

For each, $C_s \neq C_t$ construct a transport matrix Δ_{st} as in the previous theorem, as to assure that $\Delta_{st}H^{C_sC_t} = H^{C_sC_t}\Delta_{st} = 0$. This can be done since $H^{C_sC_t}$ takes into account only two clusters. Recall that the total amount of weight to be corrected is $b(\Gamma,\delta)$. Let $w_{s,t}$ the weight to be distributed from cluster s to cluster t. In the notation of the previous theorem, w_{st} is just the sum of the $\gamma_{l_p}^{ij}$ where $l_p \in C_s$ and $ij \in C_t$.

Let

$$C := -\frac{2\delta}{n}H^* + \sum_{i<j}\Delta_{ij} = \frac{-2\delta}{n}\frac{1}{l-1}\sum_{1\le i<j\le l} H^{C_iC_j} + \sum_{i<j}\Delta_{ij}$$

Finally, assume that for each $i<j$, $\frac{1}{l-1}||H^{C_iC_j}|| \ge ||\Delta_{ij}||$.

Then,

$$\begin{aligned}\|C\| &= \left\|\frac{-2\delta}{n}\frac{1}{l-1}\sum_{1\le i<j\le l} H^{C_iC_j} + \sum_{i<j}\Delta_{ij}\right\| \\ &\le \frac{2\delta}{n(l-1)}\sum_{1\le i<j\le l}\left\|H^{C_iC_j} + \frac{n(l-1)}{2\delta}\Delta_{ij}\right\| \\ &= \frac{2\delta}{n(l-1)}\sum_{1\le i<j\le l}\max\left(\left\|H^{C_iC_j}\right\|, \left\|\frac{n(l-1)}{2\delta}\Delta_{ij}\right\|\right)\end{aligned}$$

Now notice that

$$\begin{aligned}\delta \ge b(\Gamma,\delta) \text{ therefore } (l-1)n &\ge \frac{(l-1)n}{\delta}b(\Gamma,\delta) \\ \text{which implies that } \sum_{i<j}\left\|H^{C_iC_j}\right\| &\ge \frac{(l-1)n}{\delta}\sum_{i<j} w_{i,j} \\ &\ge \frac{(l-1)n}{2\delta}\sum_{i<j}\left\|\Delta_{i,j}\right\|.\end{aligned}$$

By the claim, we can assume without loss of generality that for each $i<j$,

$$\left\|H^{C_i,C_j}\right\| \ge \frac{(l-1)n}{\delta}\left\|\Delta_{i,j}\right\|.$$

This implies that the last sum reduces to

$$\frac{2\delta}{n(l-1)}\sum_{1\le i<j\le l}\left\|H^{C_iC_j}\right\| = \sum_{1\le i<j\le l}\frac{2\delta(|C_i|+|C_j|)}{n(l-1)} = \frac{2\delta(l-1)}{n(l-1)}\sum_{1\le i<j\le l}|C_i|+|C_j| = 2\delta.$$

And so $\|C\| \le 2\delta$.

□

♣♣♣ Daniel: [toca revisar que $< h^*, C >$ es igual a $-2\delta n$ o eso es obvio?]

Uniqueness of the minimizer. In this brief section we discuss an important corollary: A^* is the unique minimizer of the optimization problem $min_A\Delta(A)$. We begin proving a well known lemma.

Lemma 3.6. *Let f be a convex function defined over a region D. Let $\hat{x}$ be a point in it's domain such that the subdifferential of f at $\hat{x}$ is full dimensional and 0 belongs to it's interior. Then, $\hat{x}$ is the unique minimizer of f.*

Proof. Let $B_\epsilon(0)$ be a ball of radius ϵ centered in 0 and contained in $\partial(f)(\hat{x})$. Let $x \in D$ with $x \neq \hat{x}$. Let $Q \in \partial(f)(\hat{x})$. By the property of the elements of the subdifferential at a point,

$$f(x) \geq f(\hat{x}) + \langle Q, x - \hat{x} \rangle .$$

This property holds for every $Q \in \partial(f)(\hat{x})$, so taking supremum we obtain that

$$f(x) \geq \sup_{Q \in \partial(f)(\hat{x})} f(\hat{x}) + \langle Q, x - \hat{x} \rangle ..$$

Since $B_\epsilon(0) \subseteq \partial(f)(\hat{x})$ we obtain that

$$f(x) \geq f(\hat{x}) + \sup_{Q \in B_\epsilon(0)} \langle Q, x - \hat{x} \rangle = f(\hat{x}) + \epsilon \|x - \hat{x}\|.$$

As $x \neq \hat{x}$, $\epsilon\|x - \hat{x}\| > 0$. Therefore, $f(x) > f(\hat{x})$ for all $x \in D$ different of $\hat{x}$.

□

Remark 3.7. Under the conditions of theorem 3.3, The matrix C^1 we constructed satisfies the inequalities given by 3.1 strictly, so C^1 is an interior point of $\partial(f)(A^*)$. It follows that the matrix C constructed in 3.5 also satisfies these inequalities strictly and thefore it is also an interior point of $\partial(f)(A^*)$.

Corollary 3.8. *Under the conditions of theorem 3.3, A^* is the unique minimizer of the optimization problem $min_A\Delta(A)$.*

Proof. By 3.7, the matrix C constructed in 3.5 is an interior point of $\partial(f)(A^*)$. Therefore, as the subdifferential of Δ at A^* is equal to the Minkowski sum of the subdifferentials of g and f at A^*,$0 = C - C$ is an interior point of $\partial(\Delta)(A^*)$. It follows by lemma 3.6 that A^* is the unique minimizer of the optimization problem $min_A\Delta(A)$.

□

Using Theorem 3.3 we now estimate the probabilities of perfect recovery of the correct cluster structure.

3.2. **A bound for recovery probabilities.** In this section we will bound the probability that the correct A^* is not an optimal solution of our proposed optimization problem. A key tool will be the following version of Hoeffding's inequality: If $X_1, \ldots, X_T$ are independent random variables with values in $[c_i, d_i]$ and $\Lambda_T := \sum_{i=1}^T X_i$ then the following inequality holds for all $t \geq 0$

$$\mathbb{P}\{\Lambda_T - \mathbb{E}[\Lambda_T] \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^T (d_i - c_i)^2}\right).$$

Theorem 3.9. *Suppose $B_1, \ldots, B_N$ are independent and have the same distribution as $\mathcal{G}$. If $\alpha = \min(|p_t - \frac{1}{2}|, |q - \frac{1}{2}|)$ and δ^* is the maximum of $a(\delta) := \frac{\delta\left(\alpha - \frac{\delta}{n}\right)^2}{\left(1+\frac{2\delta}{n}\right)}$ in $[0, \alpha n]$ then the probability that A^* is not a minimizer of (??) is bounded above by*

$$\exp\left(-\frac{2N(n-1)^3(\alpha n - \delta^*)^2}{n}\right) + e^{-N(\delta^* a(\delta^*))} \prod_{i \neq j} \left(1 + (e^{-4a} p_{ij} + q_{ij})^{\frac{N}{2}}\right).$$

Moreover this quantity decreases exponentially with the sample size N.

Proof. By Lemma 3.5 and the union bound the probability that A^* is not an optimal solution of problem (1.2) is bounded above by

$$\mathbb{P}\{b(\Gamma, \delta) \geq a(\Gamma, \delta)\} + \mathbb{P}\{b(\Gamma, \delta) \geq \delta\}. \tag{6}$$

and we will find upper bounds for the individual terms in (6). For $t = 1, \ldots, N$ and $i, j \in [n]$ define

$$Z_{ij}^{(t)} := \begin{cases} -1, \text{ if } (B_t)_{ij} = 1 \\ +1, \text{ if } (B_t)_{ij} = 0 \end{cases}$$

and note that for every $i \neq j$ the equality $\sum_{t=1}^N \frac{Z_{ij}^{(t)}}{N} = \Gamma_{ij}$ holds. As a result

$$b(\Gamma, \delta) - a(\Gamma, \delta) = \sum_{i \neq j} \left(\widetilde{\Gamma_{ij}} + \frac{2\delta}{n}\right) = \frac{2\delta n(n-1)}{n} + \sum_{t=1}^N \sum_{i \neq j} \frac{\widetilde{Z_{ij}^{(t)}}}{N}$$

and therefore if $M := \mathbb{E}\left[\sum_{t=1}^N \sum_{i \neq j} \frac{\widetilde{Z_{ij}^{(t)}}}{N}\right]$ then the number M is negative and is given by the formula

$$M = (2q-1)2|O| + \sum_{i=1}^{l} 2\binom{c_i}{2}(1 - 2p_i) \leq -2\alpha n(n-1)$$

We can therefore bound the probability in the first term with

$$\mathbb{P}\left\{\frac{2\delta n(n-1)}{n} + \sum_{t=1}^N \sum_{i \neq j} \frac{\widetilde{Z_{ij}^{(t)}}}{N} \geq 0\right\} = \mathbb{P}\left\{\sum_{t=1}^N \sum_{i \neq j} \frac{\widetilde{Z_{ij}^{(t)}}}{N} - M \geq -2\delta(n-1) - M\right\} \leq$$

$$\leq \exp\left(-2\frac{(-M - 2\delta(n-1))^2}{Nn(n-1)(\frac{2}{N})^2}\right) = \exp\left(-\frac{N}{2}\left(1 - \frac{1}{n}\right)\left(\frac{-M}{n-1} - 2\delta\right)^2\right)$$

where the inequality follows from Hoeffding's inequality applied to the $Nn(n-1)$ independent random variables $\frac{\widetilde{Z_{ij}^{(t)}}}{N}$ which have values in $\left[-\frac{1}{N}, \frac{1}{N}\right]$. The inequality applies whenever $-\frac{M}{2(n-1)} > \delta > 0$. In particular whenever $0 < \delta < \alpha n$ we have

$$\mathbb{P}\{b(\Gamma, \delta) - a(\Gamma, \delta)\} \leq \exp\left(-\frac{2N(n-1)^3(\alpha n - \delta)^2}{n}\right).$$

Bounding the second term is more involved. Recall that

$$b(\Gamma,\delta)=\sum_{i\neq j}\max\left(\widetilde{\Gamma_{ij}}+\frac{2\delta}{n},0\right).$$

Let $Y_{ij}:=\widetilde{\Gamma_{ij}}+\frac{2\delta}{n}$ and let $X_{ij}:=\max(Y_{ij},0)$. In order to prove a concentration inequality for the variables X_{ij} we begin by studying their moment generating functions $m_{X_{ij}}(t)$. Note that for every real number t the equality

$$\exp(tX_{ij})=1_{\{Y_{ij}\leq 0\}}+1_{\{Y_{ij}\geq 0\}}\exp tY_{ij}$$

holds. Now $Y_{ij}=1+\frac{2\delta}{n}-2\frac{n_{ij}}{N}$ where n_{ij} is a binomial random variable with parameters N and p_{ij} given by

$$p_{ij}:=\begin{cases}p_t, \text{ if } \{i,j\}\subseteq C_t\\ 1-q, \text{ else}\end{cases}$$

As a result taking expected values on both sides of the expression above we conclude that

$$m_{X_{ij}}(t)\leq\mathbb{P}\{Y_{ij}\leq 0\}+e^{t\left(1+\frac{2\delta}{n}\right)}\mathbb{E}\left(e^{-\frac{2t}{N}n_{ij}}1_{\{Y_{ij}\geq 0\}}\right).$$

Using the Cauchy-Schwartz inequality and the known formula for the moment generating function of a binomial random variable it follows that

$$m_{X_{ij}}(t)\leq\mathbb{P}\{Y_{ij}\leq 0\}+e^{t\left(1+\frac{2\delta}{n}\right)}\mathbb{E}\left(e^{-\frac{4t}{N}n_{ij}}\right)^{\frac{1}{2}}\mathbb{P}\{Y_{ij}\geq 0\}^{\frac{1}{2}}=$$

$$=\mathbb{P}\{Y_{ij}\leq 0\}+e^{t\left(1+\frac{2\delta}{n}\right)}\left(e^{-\frac{4t}{N}}p_{ij}+q_{ij}\right)^{\frac{N}{2}}\mathbb{P}\{Y_{ij}\geq 0\}^{\frac{1}{2}}$$

where $q_{ij}:=1-p_{ij}$. By Hoeffding's inequality on Bernoulli random variables we know that

$$\mathbb{P}\{Y_{ij}\geq 0\}\leq\exp\left(-\frac{N}{2}\left(-\frac{2\delta}{n}-(1-2p_{ij})\right)^2\right)\leq\exp\left(-2N(\alpha-\delta/n)^2\right)$$

so if $t=aN$ the inequality

$$e^{t\left(1+\frac{2\delta}{n}\right)}\exp\left(-N(\alpha-\delta/n)^2\right)\leq 1$$

holds whenever $a\leq\frac{(\alpha-\delta/n)^2}{\left(1+\frac{2\delta}{n}\right)}$ and for all such a we have

$$m_{X_{ij}}(aN)\leq\left(1+(e^{-4a}p_{ij}+q_{ij})^{\frac{N}{2}}\right)$$

We define $a(\delta):=\frac{(\alpha-\delta/n)^2}{\left(1+\frac{2\delta}{n}\right)}$ and will use it to prove a moment concentration inequality for $b(\Gamma,\delta)$ which will give us a bound on the second term in (6). For every $t>0$ we have

$$\mathbb{P}\left\{b(\Gamma,\delta)\geq\delta\right\}=\mathbb{P}\left\{\exp\left(t\sum_{i\neq j}X_{ij}\right)\geq e^{t\delta}\right\}\leq e^{-t\delta}\prod_{i\neq j}\mathbb{E}[e^{tX_{ij}}]=e^{-t\delta}m_{X_{ij}}(t)$$

Choosing $t = a(\delta)N$ and using the previous inequality we see that

$$\mathbb{P}\left\{b(\Gamma,\delta) \geq \delta\right\} \leq e^{-N\delta a(\delta)} \prod_{i\neq j} \left(1 + (e^{-4a}p_{ij} + q_{ij})^{\frac{N}{2}}\right)$$

Which decreases exponentially in N for any $0 \leq \delta \leq n\alpha$. The rate of decrease of the first term is controlled by the positive factor

$$\delta a(\delta) = \frac{\delta\left(\alpha - \frac{\delta}{n}\right)^2}{\left(1 + \frac{2\delta}{n}\right)}.$$

and we let δ^* be a maximizer of this function. □

4. An algorithm for Wasserstein nuclear norm summarization

Solving the optimization problems appearing in Theorem 1.4 require solving large semidefinite programs which are beyond the capacity of standard off-the-shelf software even for relatively small graphs (of say 70 vertices with $N = 4$). One possible reason is that off-the-shelf solvers often use interior point methods, which are highly accurate but often do not scale well. A better alternative, especially well suited for solving is to use first order numerical optimization methods such as ADMM.

In this section, we provide an algorithm specially adapted for the Wassesrtein nuclear norm summarization. It is based on a variation of the **Alternating Direction Method of Multipliers** (ADMM) called **Global Variable consensus with regularization**. Our algorithm is derived using the theory given in [6].

4.1. Global Variable consensus with regularization.

The general form of Global Variable consensus with regularization is given by

$$\begin{aligned} &\text{Minimize } \sum_{i=1}^{n} f_i(x_i) + g(z) \\ &\text{Subject to: } x_i - z = 0, \quad i = 1,..,N. \end{aligned} \tag{7}$$

To efficiently solve this optimization problem, ADMM does iterative rounds of minimization of the primal and dual variables. For certain functions, (as is the case for the l_1 norm and crucially the nuclear norm) these minimizations problems have actual analytical minimizers.

The general algorithm to solve this optimization problem is given in [6] and consists of the following steps:

(1) initialize x_i^0, y_i^0 for $i = 1,..,N$, $\rho > 0$ and z^0.
(2) Set Set $\bar{x}^k = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y}^k = \frac{1}{N}\sum_{i=1}^{N} y_i$.
(3) $x_i^{k+1} = \arg\min_{x_i}\left(f_i(x_i) + \left\langle y_i^k, x_i - z_k\right\rangle) + \frac{\rho}{2}\|x_i - z^k\|_2^2\right)$.
(4) $z^{k+1} = \arg\min_z\left(g(z) + \frac{N\rho}{2}\|z^k - \bar{x}^{k+1} - \frac{1}{\rho}\bar{y}^k\|_2^2\right)$.
(5) $y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1})$.

This algorithm converges under very general circumstances, wich can be easily checked for problem 1 : ♣♣♣ Daniel: [las condiciones son que f y g sean propias, cerradas y convexas. mas aun, el lagrangiano no aumentado L_0 debe tener un punto de silla. deberiamos incluir una demostracion de esto? mas explicitamente, la condicion es que existen (x^*, y^*, z^*) no necesariamente unicos tales que $L_0(x^*, z^*, y) \leq L_0(x^*, z^*, y^*) \leq L_0(x, z, y^*)$. donde $L_0(x, y, z) = f(x) + g(z) + \langle y, Ax - Bz - C\rangle$ en el problema de optmizacion general $minf(x) + g(z)$ sujeto a $Ax + Bz = c$.]

4.2. **Adaptation to the Wasserstein nuclear norm summarization.**

Definition 4.1. *Let f be a convex function, $\rho > 0$. The proximal operator of f at w is defined as:*

$$prox_{f,\rho}(w) = \arg\min_z f(z) + \frac{\rho}{2}\|z - w\|_2^2$$

Claim 4.2. *if $\|\cdot\|_*$ denotes the nuclear norm, and if w is a matrix with singular value decomposition USV^t then,*

$$UP_\epsilon V^t = \arg\min_x \epsilon\|x\|_* + \frac{1}{2}\|x - w\|_2^2.$$

Where

$$P_\epsilon = \begin{cases} x - \epsilon \text{ if } x > \epsilon \\ x + \epsilon \text{ if } x < -\epsilon \\ 0 \text{ in any other case.} \end{cases}$$

and P_ϵ is applied component-wise.P_ϵ is usually called the soft-thresholding operator.

Claim 4.3. *if $\|\cdot\|_1$ denotes the l_1 norm, then*

$$prox_{\|\cdot\|_1,\rho}(w) = \arg\min_x \|x\|_1 + \frac{\rho}{2}\|x - w\|_2^2 = P_{\frac{1}{\rho}}(w).$$

Where $P_{\frac{1}{\rho}}$ is applied component-wise. Moreover, if c is any $n \times n$ matrix, then

$$P_{\frac{1}{\rho}}(w) + c = \arg\min_x \|x - c\|_1 + \frac{\rho}{2}\|x - w\|_2^2.$$

Let $B_1, .., B_N$ of an $i.i.d$ sample of the random graph $\mathcal{G}$. Let n be the number of vertices of $\mathcal{G}$ and 1 denote the vector [1,..,1] of length n. To use the formulation given in 7 to solve the problem we need the following change of variables:

- $C_i = 2B_i - 11^t$.
- $Z = 2A - 11^t$.
- $f_i(x_i) = \frac{1}{2}\|x_i - C_i\|_1$.
- $g(z) = \|z\|_*$.

Our optimization problem reduces to the one given in 7 and an algorithm to solve it is given by:

(1) initialize $\rho > 0 x_i^0 = 0, y_i^0 = 0$, for $i = 1, .., N$ and $z_0 = \frac{1}{N}\sum_{i=1}^N C_i$.
(2) Set $\bar{x}^k = \frac{1}{N}\sum_{i=1}^N x_i$ and $\bar{y}^k = \frac{1}{N}\sum_{i=1}^N y_i$.
(3) $x_i^{k+1} = P_{\frac{1}{2\rho}}(z^k - \frac{1}{\rho}y_i^k - C_i) + C_i$

(4) $z^{k+1} = UP_{\frac{\lambda}{N\rho}}(S)V^t$ with $USV^t = SVD(\bar{x}^{k+1} + \frac{1}{\rho}\bar{y}^k)$.
(5) $y^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1})$.

Using this algorithm, we tested the recovery of graphs generated by the stochastic block model.The simulations are in sink with our theorical results. Results are shown in figures ... .

5. Some numerical examples using the stochastic block model.

♣♣♣ Mauricio: [Aca creo que deberia haber ejemplos de tres tipos:

(1) De grafos chiquitos, digamos 20 vertices en los que se vea como cambia el óptimo en la medida en que el parámetro δ va cambiando.
(2) En grafos medianos sus gráficas comparativas de performance para diferentes métricas con pocos samples
(3) En grafos grandes algo de cross-validation? Para ver como escoger el δ? Me imagino que si uno va viendo que tan lejos esta de ser entera la solucion entonces uno puede encontrar un rango para δ donde da resultados chéveres.

]

6. Preliminaries

6.1. **Preliminaries on graphs and norms.** By a graph G we mean a finite loopless undirected graph. We say that G is weighted if it is endowed with a function $w : E(G) \to \mathbb{R}$ which assigns to every edge a real number in $[0, 1]$. If G has n vertices then it is completely specified by its adjacency matrix $A \in \{0, 1\}^{n\times n}$ defined by $A_{ij} = 1$ if and only if vertices i, j are connected. If G is weighted then we use the term adjacency matrix of G to denote the matrix with entries $A_{i,j} = w(i, j)$.

If A is a matrix then we use $\| \bullet \|$, $\| \bullet \|_1$ to denote its operator norm and ℓ^1-norm respectively.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.
[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.
[3] N. Ailon, Y. Chen, and H. Xu. Breaking the small cluster barrier of graph clustering. In *International Conference on Machine Learning*, pages 995–1003, 2013.
[4] B. P. Ames and S. A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011.
[5] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.
[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
[7] I. Cabreros, E. Abbe, and A. Tsirigos. Detecting community structures in hi-c genomic data. In *Information Science and Systems (CISS), 2016 Annual Conference on*, pages 584–589. IEEE, 2016.

[8] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[9] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[10] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[11] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[12] Y. Chen, S. Sanghavi, and H. Xu. Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212, 2012.

[13] P. Chin, A. Rao, and V. Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423, 2015.

[14] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, et al. Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366, 2007.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[16] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[17] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

[18] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[19] O. Guédon and R. Vershynin. Community detection in sparse networks via grothendiecks inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[20] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[21] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

[22] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67. Springer, 2007.

[23] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171(1-2, Ser. A):115–166, 2018.

[24] A. Montanari and S. Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827. ACM, 2016.

[25] M. E. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[26] S. Oymak and B. Hassibi. Finding dense clusters via" low rank+ sparse" decomposition. *arXiv preprint arXiv:1104.5186*, 2011.

[27] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[28] R. K. Vinayak, S. Oymak, and B. Hassibi. Sharp performance bounds for graph clustering via convex optimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8297–8301. IEEE, 2014.

[29] Y. Xu, V. Olman, and D. Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, 2002.

Departamento de matemáticas, Universidad de los Andes, Carrera $1^{\text{ra}}\#18A - 12$, Bogotá, Colombia

E-mail address: d.de1033@uniandes.edu.co

Departamento de matemáticas, Universidad de los Andes, Carrera $1^{\text{ra}}\#18A - 12$, Bogotá, Colombia

E-mail address: mvelasco@uniandes.edu.co