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 leads to 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 $B_1, \ldots, B_N$ 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

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 [25] 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:

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

(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

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

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. *Assume $\mathcal{G}$ is a stochastic block model with $\ell = 2$ clusters. Suppose $B_1, \ldots, B_N$ are independent and are distributed according to $\mathcal{G}$. If $\alpha = \min(|p_t -$*

$\frac{1}{2}|, |q - \frac{1}{2}|)$, $\delta^* = \frac{n\alpha}{4}$ and $a(\delta) := \frac{\delta\left(\alpha - \frac{\delta}{n}\right)^2}{n\left(1 + \frac{2\delta}{n}\right)}$ 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 4.3 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 these 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 [8, 15, 31], social data understanding [17, 24, 27] and other general machine learning tasks such as natural language processing [16, 29]. 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 [6, 14, 21, 23].
(2) SDP approaches based on the seminal work of Goemans and Williamson [19], [2, 20, 26].
(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 [9, 11, 10]. These methods have led to several convex optimization algorithms to find communities in graphs [4, 30, 13, 12, 28, 3].

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

(1) Our methods apply to the problem of learning a graph from a set consisting of *several* samples while, to the best of our knowledge, the above algorithms must be applied 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 4.3)

(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 4.3).

2. ROBUST LEARNING OF DETERMINISTIC SUMMARIES

Let $[n] := \{1, 2, \ldots, n\}$. 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 $A = A^*$. $\square$

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 $\mathcal{G}$ [18] 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

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

Next we move the supremum into the constraints obtaining

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

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

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

we therefore conclude that

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

Letting A vary in K we prove the Theorem. $\square$

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

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

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

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

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

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

where

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

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

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

which by duality is equivalent to

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

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

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

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. $\square$

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 [5].

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

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 $C_1, \ldots, C_\ell$ 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 3.2 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.
In this section we will study some basic properties of the subdifferential of $\Delta(A)$. Recall that if h is a real valued convex function on a real inner product space V then the subdifferential of h at a point $a \in \text{dom}(h)$ is the closed convex set given by

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

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

Recall that

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

And define functions f, g by

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

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_k with $k = 1, \ldots N$ and let Γ be the symmetric matrix with entries given by $\Gamma_{ij} = \frac{n_0(ij) - n_1(ij)}{N}$. Finally let $I := \{(i, j) : \exists t \in \{1, \ldots, l\} \, (\{i, j\} \subseteq C_t)\}$ and let $O = [n] \times [n] \setminus I$ so I contains edges which belong to some cluster while O contains all edges which connect distinct clusters. Henceforth we will adopt the convention that adjacency matrices have diagonal entries equal to one and in particular $\Gamma_{ii} = -1$ for all i.

The following Lemma summarizes some basic subdifferential properties of the functions f and g at A^*. Define the matrix $H^* := 2A^* - 11^t$ so H^* is an $\ell \times \ell$ block-matrix

whose blocks correspond to the clusters. The diagonal blocks of H^* have all entries equal to 1 and all the off-diagonal blocks have all entries equal to -1.

Lemma 3.1. *The following statements hold:*

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

$$\begin{cases} -1 \le C_{ij} \le 1, \text{ if } i = j, \\ \Gamma_{ij} \le C_{ij} \le 1, \text{ if } \{i, j\} \in I, i \ne j \\ -1 \le C_{ij} \le \Gamma_{ij}, \text{ if } \{i, j\} \in O. \end{cases}$$

(2) *The subdifferential of g at A^* is the set of symmetric matrices C which satisfy $\|C\| \le 2\delta$ and the equality $\langle C, H^* \rangle = 2\delta \|H^*\|_*$.*

Proof. (1) Since the subdifferential is additive it suffices to understand the subdifferential of the absolute value which is given by

$$\partial(|x - a|)(b) = \begin{cases} 1, \text{ if } b > a \\ [-1, 1], \text{ if } b = a \\ -1, \text{ if } b < a \end{cases}$$

If $i, j \in C_t$ with $i \ne j$ then $A_{ij}^* = 1$ and the entry ij of the subdifferential of the sum at A^* is $[-1, 1]$ for each B_k containing the edge and it is 1 for each B_k for which ij is not an edge. If $i = j$ then the entry ii of the subdifferential at A^* is $[-1, 1]$ leading to the inequality $-1 \le C_{ii} \le 1$. Finally 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_k which contains the edge ij and $[-1, 1]$ for each B_k which does not, proving the claim. (2) It follows from the definition of subdifferential that for any norm $\|\bullet\|$ and any point X the subdifferential of the norm at X is given by those C for which the dual norm $\|C\|_* \le 1$ and $\langle C, X \rangle = \|X\|$ from which claim (2) follows immediately. $\square$

To simplify the inequalities in the previous Theorem we define a linear operator $\widetilde{\bullet} : \mathcal{S}^{n \times n} \to \mathcal{S}^{n \times n}$ on symmetric matrices by the formula:

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

The operator $\widetilde{\bullet}$ is an involution which satisfies $\langle A, B \rangle = \langle \widetilde{A}, \widetilde{B} \rangle$ for any two symmetric matrices A and B. Moreover using this operator, the previous Lemma can ve restated by saying that C_{ij} belongs to the subdifferential of f at A^* if and only if $\widetilde{\Gamma}_{ij} \le \widetilde{C}_{ij} \le 1$ for all i, j. Note also that $\widetilde{H^*} = 11^t$

Our next Lemma gives sufficient conditions for the true cluster structure A^* to be a minimizer of $\Delta(A)$ in K. In order to describe it we introduce the following two quantities.

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

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

The quantity $b(\Gamma, \delta)$ (resp. $a(\Gamma, \delta)$) measures the total amount by which the matrix $-\frac{2\delta}{n}H^* = -\widetilde{\frac{2\delta}{n}11^t}$ fails (resp. succeeds) to be in the subdifferential of Lemma 3.1. This is because $b(\Gamma, \delta)$ (resp. $a(\Gamma, \delta)$) is the result of summing 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 using transportation matrices. In this sense the following Theorem explains the success of the spectral norm in cluster recovery algorithms.

We will begin with the case when our graph involves only two clusters. We will use that $\langle H^*, H^* \rangle = n^2$ and that the equalities $\|H^*\|_* = \|H\| = n$ hold when there are only two clusters.

Theorem 3.3. *Assume there are only two clusters. Let δ be a real number satisfying $0 < \delta < \frac{n}{4}$. If $b(\Gamma, \delta) < \min(\frac{\delta}{n}, a(\Gamma, \delta))$ then A^* is a minimizer of the optimization problem $\min_A \Delta(A)$.*

Proof. We will construct a matrix C such that $-C \in \partial(g)(A^*)$ and $C \in \partial(f)(A^*)$. As a result $0 = C - C$ belongs to the subdifferential $\partial(\Delta)(A^*)$ proving that A^* is a minimizer of $\Delta(A)$ as claimed. As a first candidate for C we could choose $C = -\frac{2\delta}{n}H^*$. Since

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

we conclude that $-C \in \partial g(A^*)$. Moreover $C \in \partial f(A^*)$ precisely if $\widetilde{\Gamma_{ij}} \leq \widetilde{C_{ij}} = -\frac{2\delta}{n} \leq 1$ for all ij. In general, the first inequality need not hold and we can split the entries ij into two sets U and V where this inequality is valid and where it (strictly) fails respectively. More precisely, define

$$U = \left\{ (i,j) : i \neq j \text{ and } \widetilde{\Gamma_{ij}} \leq -\frac{2\delta}{n} \right\} \ \text{and} \ V = \left\{ (i,j) : i \neq j \text{ and } \widetilde{\Gamma_{ij}} > -\frac{2\delta}{n} \right\}$$

and note that

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

If $b(\Gamma, \delta) < a(\Gamma, \delta)$ then we could modify the matrix $\widetilde{C}$ by increasing the entries of $\widetilde{C}$ in V and decreasing those in U by the same amount by adding transportation matrices so as to make the inequality valid in all entries. In the remainder of the proof we show that this process can be carried out while at the same time controlling the spectral norm of C.

Since $b(\Gamma, \delta) < a(\Gamma, \delta)$ there exist nonnegative scalars γ_{ij}^{st} for $ij \in U$ and $st \in V$ which tell us how much of the amount in entry $ij \in U$ has to be moved to entry $st \in V$. More preciselt, the numbers γ_{ij}^{st} satisfy the following inequalities:

$$(3) \quad \forall st \in V \left(\widetilde{\Gamma_{st}} \leq -\frac{2\delta}{n} + \sum_{ij \in U} \gamma_{ij}^{st} \right) \text{ and } \forall st \in U \left(\widetilde{\Gamma_{st}} \leq -\frac{2\delta}{n} - \sum_{ij \in V} \gamma_{st}^{ij} \right).$$

In order to transport the smallest possible total amount we can moreover assume that the γ_{st}^{ij} are chosen so the first inequality is an equality.

$$\forall st \in V \left(\widetilde{\Gamma_{st}} = -\frac{2\delta}{n} + \sum_{ij \in U} \gamma_{ij}^{st} \right)$$

For $st \in V$ define $W^{st} = \sum_{ij \in U} \gamma_{ij}^{st}(r_{st} - r_{ij})$ where r_{st} is the $n \times n$ symmetric matrix given by

$$(r_{ij})_{ab} = \begin{cases} -1, \text{ if } (a,b) = (i,i) \text{ or } (a,b) = (j,j) \\ 1, \text{ if } (a,b) = (i,j) \text{ or } (a,b) = (j,i) \\ 0, \text{ else.} \end{cases}$$

and let $C := -\frac{2\delta}{n}H^* + \sum_{st \in V} W^{st}$. We will show that $C \in \partial f(A^*)$ and that $-C \in \partial g(A^*)$ finishing the proof of the Theorem.

Since $\widetilde{C} = -\frac{2\delta}{n}11^t + \sum_{st \in V} W^{st}$ we know that

$$\widetilde{C}_{ab} = \begin{cases} -\frac{2\delta}{n} + \sum_{ij \in U} \gamma_{ij}^{ab} \text{ if } ab \in V \\ -\frac{2\delta}{n} - \sum_{ij \in V} \gamma_{ab}^{ij} \text{ if } ab \in U \\ -\frac{2\delta}{n} - \sum_{st \in V} \sum_{aj \in U} \gamma_{aj}^{st} + \sum_{ij \in U} \sum_{at \in V} \gamma_{ij}^{at} \text{ if } a = b \end{cases}$$

we conclude that $\widetilde{\Gamma}_{ab} \leq \widetilde{C}_{ab} \leq 1$ for $a \neq b$. The first inequality follows from those in (3). The second inequality follows since for $ab \in V$ we have $\widetilde{C}_{ab} = \widetilde{\Gamma}_{ab} \leq 1$ and for $ab \in U$ we have $\widetilde{C}_{ab} \leq -\frac{2\delta}{n} \leq 1$. Finally, for $a = b$ we see that

$$|\widetilde{C_{aa}}| \leq \frac{2\delta}{n} + \left| \sum_{st \in V} \sum_{aj \in U} \gamma_{aj}^{st} + \sum_{ij \in U} \sum_{at \in V} \gamma_{ij}^{at} \right| \leq \frac{2\delta}{n} + 2\beta(\Gamma, \delta) \leq 4\frac{\delta}{n} \leq 1$$

so $C \in \partial f(A^*)$ as claimed.

Moreover, we know that

$$\langle C, H^* \rangle = \langle \widetilde{C}, \widetilde{H^*} \rangle = \left\langle -\frac{2\delta}{n}11^t - \sum_{st \in V} W^{st}, 11^t \right\rangle = -2\delta n$$

where the last equality occurs since $\langle 11^t, r_{st} \rangle = 0$ for all st. For $i \neq j$ let $e^{ij} := \widetilde{r_{ij}}$ and note that the matrix e^{ij} satisfies the following properties:

(1) $H^* e^{ij} = 0 = e^{ij} H^*$ for all ij. This is immediate from the fact that e^{ij} is given by

$$(e^{ij})_{ab} = \begin{cases} 1, \text{ if } ij \in I \text{ and } \{i,j\} = \{a,b\}. \\ -1, \text{ if } ij \in O \text{ and } \{i,j\} = \{a,b\} \\ -1, \text{ if } a = b \text{ and } a \in \{i,j\} \\ 0, \text{ otherwise.} \end{cases}$$

(2) The inequality $\|e^{ij} - e^{st}\| \leq 2$ holds for all ij and st. This is immediate from the fact that the square of the spectral norm is bounded by the product of the induced 1-norm and the induced ∞-norm. The inequality is strict only if $|\{i,j\} \cap \{s,t\}| \geq 1$ and in this case the norm of the difference takes the values of $\sqrt{3}$ and 0 when the intersection has sizes one and two respectively.

It follows from (1) that the two summands in $C = -\frac{2\delta}{n}H^* + \sum_{st \in V} \widetilde{W^{st}}$ are symmetric matrices with product zero. We conclude that they are simultaneously diagonalizable and that the spectral norm of C is given by

$$\|C\| = \max\left(\left\| \frac{2\delta}{n} H^* \right\|, \left\| \sum_{st \in V} \widetilde{W^{st}} \right\| \right)$$

by the triangle inequality and (2) the second term is bounded by $2b(\Gamma, \delta)$ and thus the spectral norm equals 2δ since $b(\Gamma, \delta) < \delta$. We conclude that $-C \in \partial g(A^*)$ as claimed. $\qquad\square$

Remark 3.4. We believe that a similar argument should explain the very good experimental behavior of the WRNS in the case of several clusters but we have not yet been able to prove this is the case. A fundamental difficulty is that the dimension of the kernel of the matrix H^* for three or more clusters is much smaller restricting the space of possible transportation plans γ_{ij}^{st} which are obviously spectral-norm preserving.

Uniqueness of the minimizer. In this brief section we will show that A^* is in fact the unique minimizer of the optimization problem $min_A \Delta(A)$ for appropriate choices of δ. We begin by proving a well-known lemma. We could not find a specific reference for it so we include a proof for the reader's convenience.

Lemma 3.5. *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 definition of subdifferential we know that,

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

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

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

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

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

If $x \neq \hat{x}$ then $\epsilon \|x - \hat{x}\| > 0$. As a result $f(x) > f(\hat{x})$ for all $x \in D$ different of $\hat{x}$ as claimed.

$\square$

Remark 3.6. If $\widetilde{\Gamma_{ij}} \neq 1$ for all ij then the matrix C we constructed in Theorem 3.3 can be chosen to be an interior point of $\partial f(A^*)$ since it would satisfy all defining inequalities strictly. 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 automatically an interior point of $\partial(\Delta)(A^*)$ proving uniqueness of the minimizer A^*.

Corollary 3.7. *If $\widetilde{\Gamma_{ij}} \neq 1$ for all ij then under the conditions of Theorem 3.3, A^* is the unique minimizer of the optimization problem $\min_A \Delta(A)$.*

In the following section we estimate the probability of perfect recovery of the correct cluster structure by bounding above the probability of failure of the hypothesis of Theorem 3.3. Our main result is that this probability of failure decreases exponentially in N for a suitably chosen value of δ.

3.2. **A bound for recovery probabilities.** In this section we will prove Theorem 1.5 which gives an upper bound on the probability that the correct A^* is not the unique optimal solution of our optimization problem. Our main tools will be concentration inequalities and in particular the ideas used in the proof of Hoeffding's inequality. Recall that Hoeffding's inequality says that 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 holds for all $t \geq 0$

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

proof of Theorem 1.5. Assume $0 < \delta < \frac{n}{4}$. By Theorem 3.3, Corollary 3.7 and the union bound the probability that A^* is not an optimal solution of problem (1.2) is bounded above by

$$(4) \qquad \mathbb{P}\{b(\Gamma, \delta) \geq a(\Gamma, \delta)\} + \mathbb{P}\left\{b(\Gamma, \delta) \geq \frac{\delta}{n}\right\}.$$

and we will find upper bounds for the individual terms in (4). For $t = 1, \ldots, N$ and $i, j \in [n]$ define the random variable

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

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

$$\mathbb{E}[\widetilde{Z_{ij}^{(t)}}] = \begin{cases} 1 - 2p_t, & \text{if } ij \in C_t \\ 2q - 1, & \text{if } ij \in O \end{cases}$$

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

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

If $\alpha := \min(|p_t - \frac{1}{2}|, |q - \frac{1}{2}|)$ then $\mathbb{E}[M] \leq -2\alpha n(n-1)$ and we can therefore bound the probability in the first term in (4) with

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

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

by using 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 in (4) 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 on the second term 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 (4). For every $t > 0$ we have

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

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

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

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

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

and we let δ^* be a maximizer of this function in the interval $[0, \min(\alpha n, \frac{n}{4})]$. Equivalently we want to find a maximizer of the function $\frac{\delta}{n}\left(\alpha - \frac{\delta}{n}\right)^2$ which occurs when $\frac{\delta}{n} = \frac{\alpha}{4}$ by elementary calculus.

$\square$

4. A NUMERICAL 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 tend to not scale well. A better alternative, especially well suited for WRNS 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 a specialization of the general ADMM in the same spirit as [22].

4.1. **Global Variable consensus with regularization.** Let N be a positive integer. For $i = 1, \ldots, N$ let f_i, g be convex real-valued functions in $\mathbb{R}^m$. The general form of Global Variable consensus with regularization is given by

$$(5) \qquad \text{Minimize } \sum_{i=1}^{N} f_i(x_i) + g(z)$$
$$\text{Subject to: } x_i - z = 0, \quad i = 1, .., N.$$

with Lagrangian given by

$$L_\rho(x_1, \ldots, x_N, y_1, \ldots, y_N, z) = g(z) + \sum_{i=1}^{N} \left(f_i(x_i) + y_i^t(x_i - z) + \frac{\rho}{2} \|x_i - z\|_2 \right)$$

To solve this optimization problem, ADMM does iterative rounds of minimization of the primal and dual variables. For certain functions (for instance the ℓ_1-norm and crucially for us, the nuclear norm) these minimization problems admit analytic closed form expressions. The general algorithm (see [7, pg. 52]) to solve this optimization problem begins by initializing x_i^0, y_i^0 for $i = 1, .., N$, $\rho > 0$ and z^0 arbitrarily and carrying out the following steps for $k \geq 1$ assuming x_i^k, y_i^k and z^k have been defined,

(1) Set $\bar{x}^k = \frac{1}{N} \sum_{i=1}^{N} x_i$ and $\bar{y}^k = \frac{1}{N} \sum_{i=1}^{N} y_i$.
(2) For $i = 1, \ldots, N$ set

$$x_i^{k+1} = \arg\min_{x_i} \left(f_i(x_i) + \langle y_i^k, x_i - z^k \rangle + \frac{\rho}{2} \|x_i - z^k\|_2^2 \right).$$

(3) $z^{k+1} = \arg\min_z \left(g(z) + \frac{N\rho}{2} \|z^k - \bar{x}^{k+1} - \frac{1}{\rho} \bar{y}^k\|_2^2 \right).$
(4) $y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1}).$

The primal residual on the k-th step is defined as:

$$r^k := (x_1^k - z^k, x_2^k - z^k, ..., x_N^k - z^k)$$

and the dual residual in the k-th step as

$$s^k := -\rho(z^k - z^{k-1})$$

We use as stopping criteria that the norms of both residuals are below a given tolerance parameter. This choice leads to a computable upper bound on suboptimality of the k-th iterate as shown in [7, Section 3.3].

4.2. Specialization to the Wasserstein nuclear norm summarization. If f is a convex function and $\rho > 0$ then the proximal operator of f at w is defined as:

$$\text{prox}_{f,\rho}(w) = \arg\min_z \left(f(z) + \frac{\rho}{2}\|z - w\|_2^2 \right)$$

Note steps (3) and (4) of ADMM are computations of certain proximal operators. The following Lemma shows that these operations can be computed via closed-form expressions for some functions of interest. To describe them we will use the soft-thresholding operator which on numbers $x \in \mathbb{R}$ is defined by

$$P_\epsilon(x) = \begin{cases} x - \epsilon, \text{ if } x > \epsilon \\ x + \epsilon, \text{ if } x < -\epsilon \\ 0, \text{ otherwise} \end{cases}$$

and on vectors or matrices by applying it element-wise.

Lemma 4.1. *The following statements hold* [22]:

(1) If $\| \cdot \|_*$ denotes the nuclear norm, and if w is a matrix with singular value decomposition USV^t then,

$$\text{prox}_{\|\cdot\|_*,\frac{1}{\epsilon}}(w) = \arg\min_x \left(\epsilon\|x\|_* + \frac{1}{2}\|x - w\|_2^2 \right) = UP_\epsilon(S)V^t.$$

(2) If $\| \cdot \|_1$ denotes the l_1 norm, then

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

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 (5) to solve the problem we need the following change of variables:

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

Our optimization problem reduces to one as in (5). Using Lemma 4.1 the specialized ADMM algorithm for WRNS is given by:

(1) initialize $\rho > 0$, $x_i^0 = 0$, $y_i^0 = 0$, for $i = 1, .., N$ and $z_0 = \frac{1}{N}\sum_{i=1}^N C_i$ and for $k \geq 1$ do:

(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) $y^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1})$.
(6) Until the stopping criteria are met.

4.3. **Numerical simulations.** In this Section we empirically test the recovery quality of Wasserstein robust nuclear norm summarization for community detection on graphs generated by the stochastic block model. Our Julia [**?**] implementation of the algorithm is available in `Zlinki`.

For the tests, we first implemented a Julia script which creates random graphs that are distributed according to the stochastic block model. It's input parameters are the following:

(1) n : the number of vertices of the graph.
(2) k : the number of clusters of the graph.
(3) q : the probability of connecting two vertices that do not belong to the same cluster.
(4) p_i for $i = 1, .., k$: the probability of connecting two vertices that belong to cluster i.

The output of this algorithm is the adjacency matrix of the generated graph.

We begin testing our algorithm by generating 3 observations of a random graph ♣♣♣ Mauricio: [De cuantos vértices? con qué α (ver teorema 1.5 para la definición)]. We then ran our algorithm for different values of δ. Recovery success is evaluated by using the *recovery error* given by the formula

$$\|A^* - \bar{A}\|_1$$

where A^* is the true adjacency matrix of the underlying cluster structure and $\bar{A}$ is the optimum of the optimization problem. We present one such plot in 1. Notice that the true graph structure is exactly recovered for certain values of δ. This is intuitively clear from the Wasserstein robust formulation. If the size of the ball is too small, we won't capture the true distribution of the random graph. If it is too big, we will capture too many distributions and the worst case error will be high. Choosing a correct δ is therefore of the utmost importance. Our simulations suggest that the region where one can choose δ to obtain perfect recovery grows as more observations are added: see figures on 4.

As δ varies, the recovered graph changes as well. Figure 2 shows the recovered graphs for different values of δ. Note that the optimum of our algorithm does not necessarily have entries in $\{0, 1\}$ (except in the region with perfect recovery) and one might be tempted to round down the entries below certain threshold (say 0.4) and round up those above another threshold (say 0.6). However, it is not clear what to do with entries between both thresholds.

Theorem 3.3 assures that as long as $p_i > 0.5$ and $q < 0.5$ the probability of perfect recovery converges to 1. The following experiments corroborate this result: We fix $n = 50$, $k = 2$, and let the first cluster have 15 vertices and the second one 35. We let $p = p_1 = p_2$ vary between 0.5 and 1 and q vary between 0 and 0.5. For each pair (p, q), we generate N observations, run our algorithm and measure the recovery error. We repeat this process 10 times and store the mean error of recovery in a
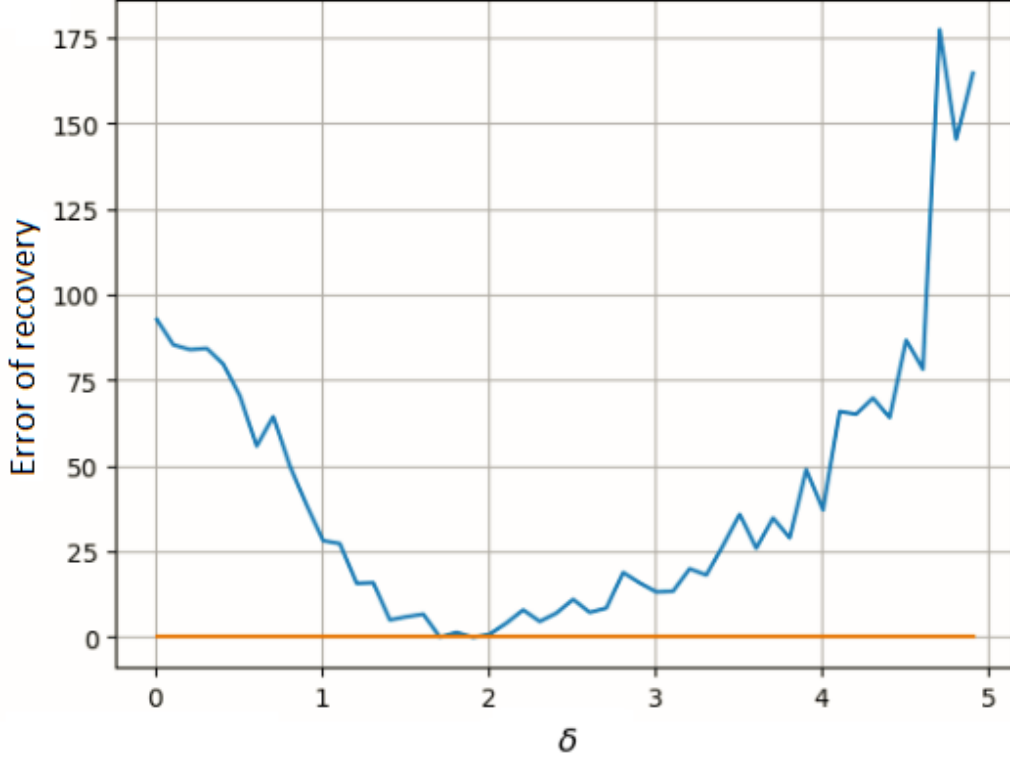
FIGURE 1. Recovery of the true cluster structure using 3 observations on a graph with generated with the stochastic block model with 2 clusters and 50 vertices.

matrix. We present the heat maps of such matrices in Figure 3 for different values of N.

In the proof of theorem 3.3, the use of the spectral norm was crucial. Still, one might ask if other norms (say the 1-norm, 2-norm or ∞-norm) can also recover the underlying cluster structure.

To check this, We fix $n = 50$, $k = 2$ where the first cluster has 15 vertices and the second one has 35. We let $p_1 = p_2 = 0.70$ and $q = 0.2$. We compute the error of recovery for different values of δ. For recovery using the nuclear norm, we use the algorithm given in the previous section. For the 1-norm, 2-norm and ∞-norm we use the *Mosek* [**?**] implementation on Julia. We also include the recovery error of the graph obtained by simply averaging the observations, which we call the mean graph. We repeat this experiment twice. First with 3 observations and then with 7. Results are shown in figure 4.

Note that the recovery error of the mean graph decreases as more observations are considered. However it is still far off the true cluster structure even for 7 observations. By contrast, even with 3 observations our nuclear norm method is capable of recovering the true cluster structure in a wide range of δs.

It is important to ask how fast does the algorithm converge to the true cluster structure as the number of samples N grows. The following figures show that this convergence occurs very quickly (at least for the stochastic block model), and thus that the Wasserstein summarization algorithm is a potentially useful approach.

All Julia code used for building the simulations in this section is available at ♣♣♣ Mauricio: [link de Daniel en GitHub].

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] J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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`
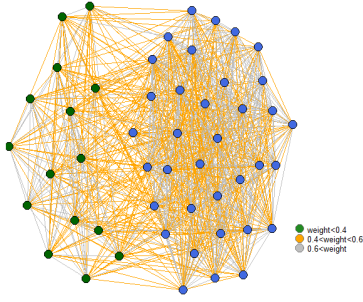
(A) Recovery with $\delta = 0$
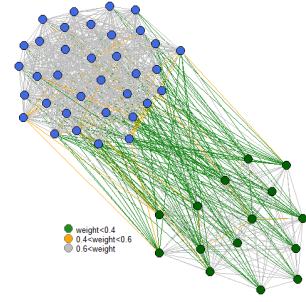
(B) Recovery with $\delta = 0.1$
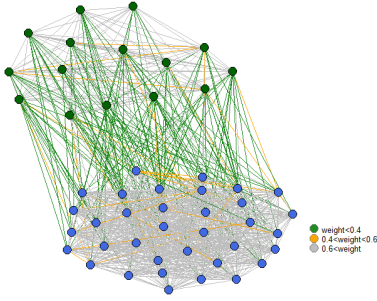
(C) Recovery with $\delta = 0.5$
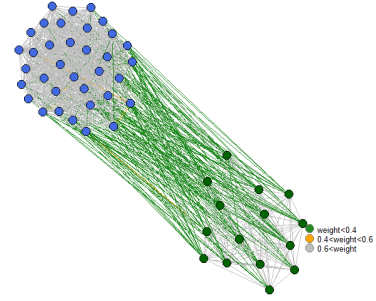
(D) Recovery with $\delta = 1.2$
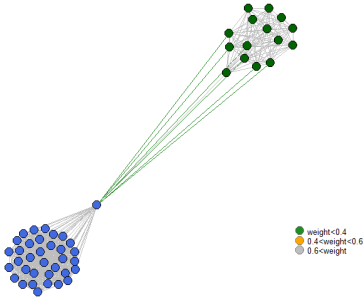
(E) Recovery with $\delta = 1.5$
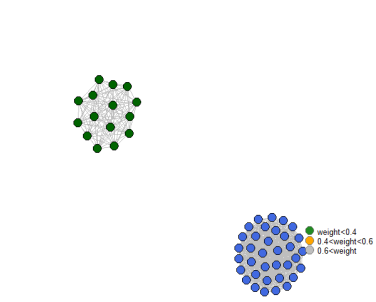
(F) Recovery with $\delta = 2$
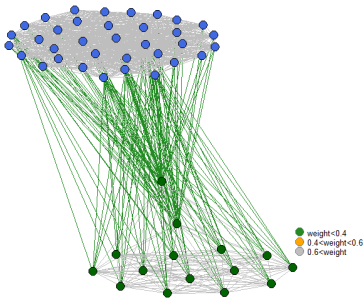
(G) Recovery with $\delta = 11$
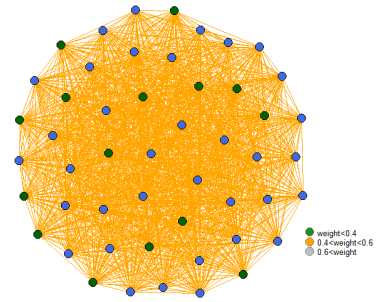
(H) Recovery with $\delta = 14$

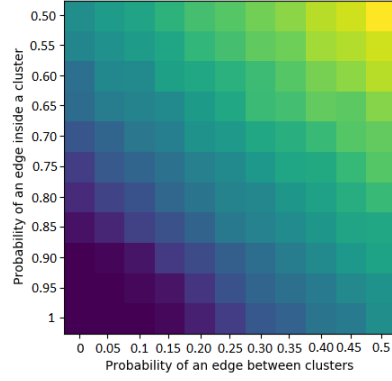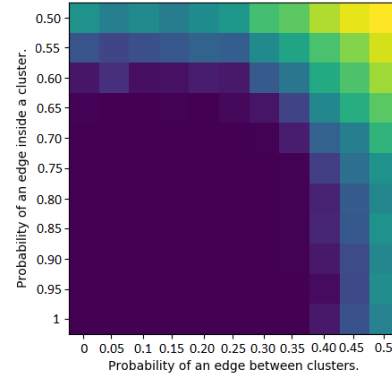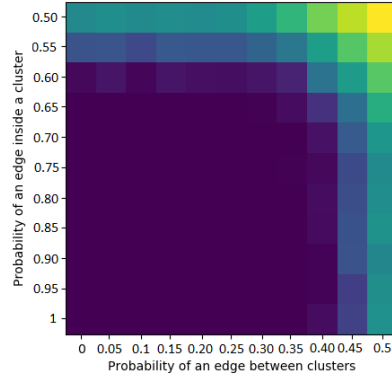FIGURE 2. Recovered graph from the same observations for different values of δ
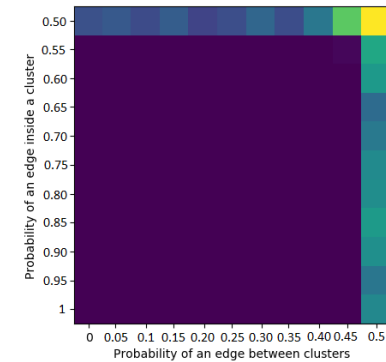
(A) $N = 1$

(B) $N = 3$

(C) $N = 7$

(D) $N = 25$

FIGURE 3. Simulation results for the average error of recovery (over 10 trials) for different values of p and q with N observations. $\delta = 1.5$ for the first two figures and 0.8 for the other two. Dark purple indicates
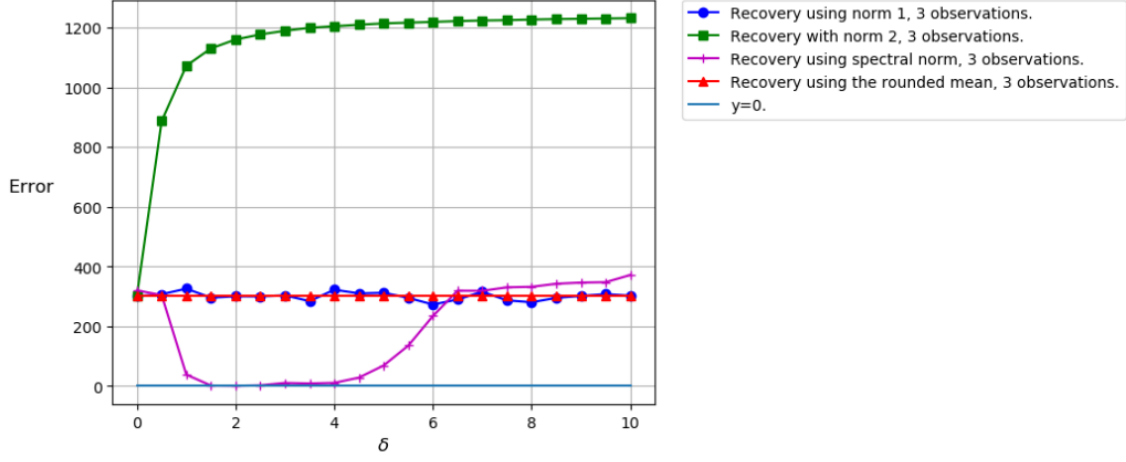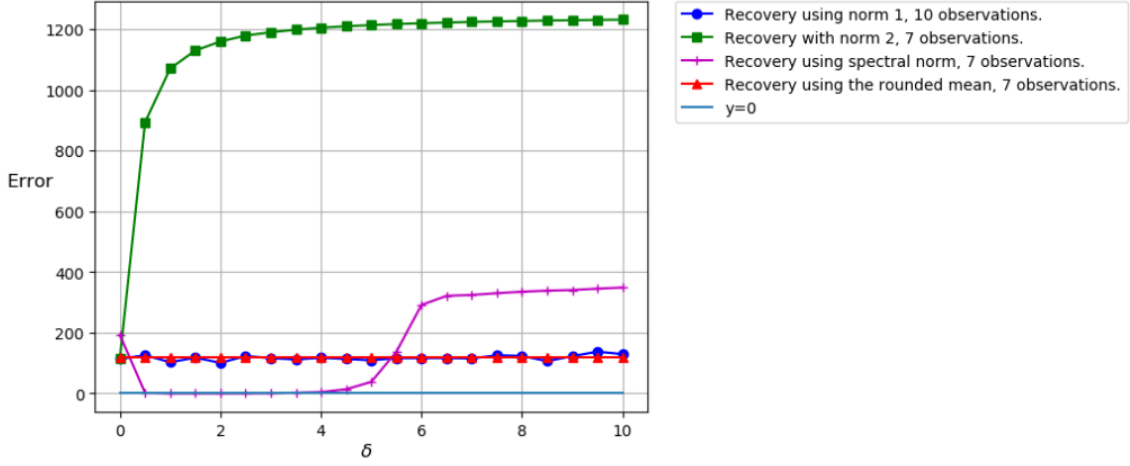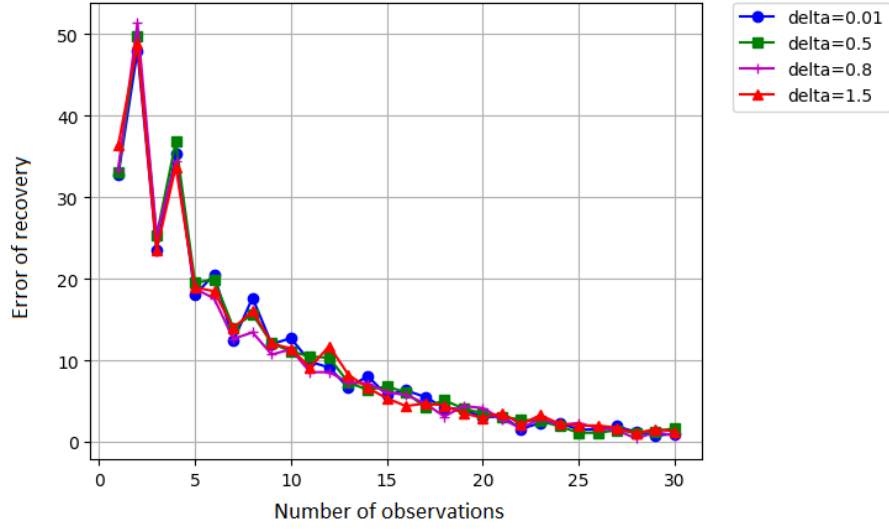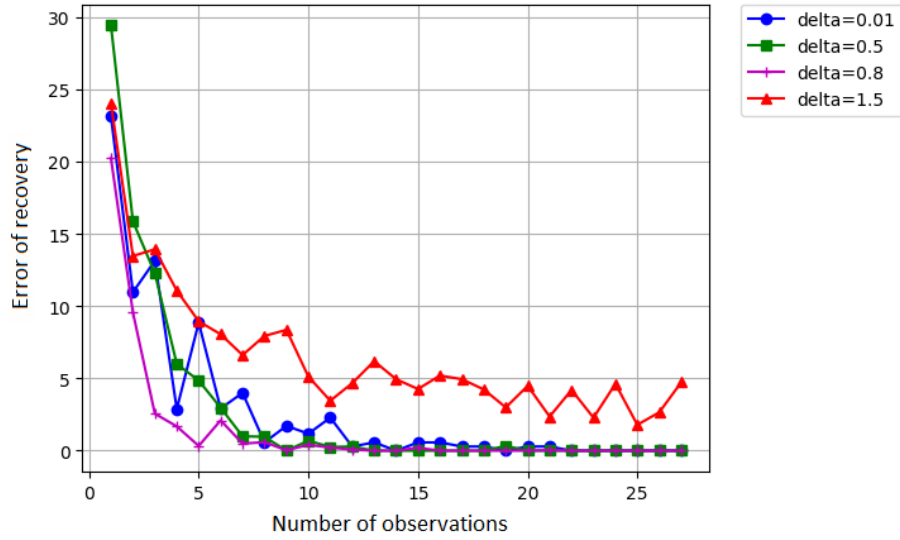
(A) $N = 3$

(B) $N = 7$

FIGURE 4. Error of recovery for different values of δ using the Wasserstein robust nuclear norm graph summarization and the l_1, l_2 norms as regularizers. The recovery error of the mean graph is also included.

(A) Error of recovery using the mean graph.

(B) Error of recovery using the Wasserstein robust nuclear norm graph summarization.

FIGURE 5. Simulation results for the speed of convergence to the true cluster structure of the Wasserstein robust nuclear norm graph summarization versus the mean graph summarization .