Optimal Partition Recovery in General Graphs

Yi Yu
University of Warwick

Oscar Hernan Madrid Padilla
University of California, Los Angeles

Alessandro Rinaldo
Carnegie Mellon University

Abstract

We consider a graph-structured change point problem in which we observe a random vector with piece-wise constant but otherwise unknown mean and whose independent, sub-Gaussian coordinates correspond to the n nodes of a fixed graph. We are interested in the localisation task of recovering the partition of the nodes associated to the constancy regions of the mean vector or, equivalently, of estimating the cut separating the sub-graphs over which the mean remains constant. Although graph-valued signals of this type have been previously studied in the literature for the different tasks of testing for the presence of an anomalous cluster and of estimating the mean vector, no localisation results are known outside the classical case of chain graphs. When the partition $\mathcal{S}$ consists of only two elements, we characterise the difficulty of the localisation problem in terms of four key parameters: the maximal noise variance σ^2, the size Δ of the smaller element of the partition, the magnitude κ of the difference in the signal values across contiguous elements of the partition and the sum of the effective resistance edge weights $|\partial_r(\mathcal{S})|$ of the corresponding cut – a graph theoretic quantity quantifying the size of the partition boundary. In particular, we demonstrate an information theoretical lower bound implying that, in the low signal-to-noise ratio regime $\kappa^2\Delta\sigma^{-2}|\partial_r(\mathcal{S})|^{-1} \lesssim 1$, no consistent estimator of the true partition exists. On the other hand, when $\kappa^2\Delta\sigma^{-2}|\partial_r(\mathcal{S})|^{-1} \gtrsim \zeta_n \log\{r(|E|)\}$, with $r(|E|)$ being the sum of effective resistance weighted edges and ζ_n being any diverging sequence in n, we show that a polynomial-time, approximate ℓ_0-penalised least squared estimator delivers a localisation error – measured by the symmetric difference between the true and estimated partition – of order $\kappa^{-2}\sigma^2|\partial_r(\mathcal{S})|\log\{r(|E|)\}$. Aside from the $\log\{r(|E|)\}$ term, this rate is minimax optimal. Finally, we provide discussions on the localisation error for more general partitions of unknown sizes.

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

1 INTRODUCTION

General graph-type data are ubiquitous in application areas, including social networks (e.g. Odeyomi et al., 2020), neuroscience (e.g. Khan et al., 2021), climatology (e.g. Mina et al., 2021), finance (e.g. Liu et al., 2021) , biology (e.g. Raimondi et al., 2021), epidemiology (e.g. Di Blasi et al., 2021), to name but a few.

In this paper, we consider a general graph-structured change point problem in which we observe a random vector in $\mathbb{R}^n$ with unknown, piece-wise constant mean and whose independent sub-Gaussian coordinates correspond to the nodes of a fixed and arbitrary graph. We are concerned with the localisation task of recovering the constancy regions of the mean vector in a manner that conforms to the topology of the underlying graph. This is motivated by the emerging of network-type data, where the graphs are not necessarily chain graphs or grid graphs. One example can be in epidemiology, people of study form a network and their interactions form edges. It would be vital to accurately identify an abnormal cluster of people.

To study this problem, we make the structural assumption that the partition associated with the mean vector specifies a multicut of the underlying graph of small weight, where the weight of each edge is its effective resistance, defined in (4). Informally, this assumption implies that the size of partition boundary is small or that the multicut is sparse relative to the topology of the graph.

The idea of weighting edges by the effective resistance in order to express the complexity of piece-wise con-

stant graph-valued signals was originally put forward by Fan and Guan (2018) for the related but different task of estimating a piece-wise constant signal over a graph in the squared error loss. In particular, the authors argue that such edge weighting scheme adapts to a varying degree of connectivity and spatial heterogeneity of the underlying graph. As a result, effective resistance provides a natural and effective quantification of the complexity of a piece-wise constant signal, more so than naively assigning a unit weight to each edge, which amounts to an ℓ_0 complexity. As our results reveal, the key insight of Fan and Guan (2018) extends to the localisation task of recovering the partition associated to the constancy regions of the mean vector, as the edge weighting by effective resistance is essential to provide nearly-optimal localisation guarantees over arbitrary graph topologies.

To the best of our knowledge, graph-structured change point localisation problems for general graphs have not been considered in the literature. Indeed, change point analysis for piece-wise constant signals traditionally assumes a total ordering of the coordinates to represent temporal changes, which can be trivially expressed using a chain graph (1-dimensional grid graph), or a d-dimensional grid graph. In contrast, when the signal is allowed to conform to an arbitrary graph in the manner described above, it is possible to obtain more sophisticated change point settings exhibiting a high degree of spatial complexity. However, due to the lack of a natural ordering of the coordinates in graph-structured mean change point problems, virtually all the existing methodologies for change point localisation, which are specifically designed to work in temporal settings, are inapplicable. To overcome this issue, we propose and analyse the properties of a change point estimator based on the approximate weighted ℓ_0-penalised least squares methodology of Fan and Guan (2018) – a polynomial-time procedure deploying the α-expansion algorithm of Boykov et al. (2001), to compute graph cuts. We make the following contributions:

• We characterise the difficulty of the partition recovery task in terms of four critical parameters, which are allowed to change with the size and topology of the graph: the size Δ of the smallest element of the partition $\mathcal{S}^*$ of the nodes induced by the piece-wise constant means, the sub-Gaussian variance factor of the noise σ^2, the smallest magnitude κ of the difference in the signal values across contiguous elements of the partition and the sum of the effective resistance edge weights for the multicut in the graph corresponding to $\mathcal{S}^*$. Specifically, we show that in the low signal-to-noise regime in which $\kappa^2\Delta\sigma^{-2}|\partial_r(\mathcal{S}^*)|^{-1} \lesssim 1$, no procedure is guaranteed to estimate the partition $\mathcal{S}^*$ with a localisation error smaller than the order of n.

• We then focus on the case in which the partition induced by the mean value is known to contain only two elements. This setting has been considered in testing the presence of an anomalous clusters (e.g. Arias-Castro et al., 2008; Sharpnack et al., 2013). We remark that, since we do not assume that the constancy regions of μ correspond to connected sub-graphs, even this simplified case allows for multiple clusters. We show that when

$$\kappa^2\Delta\sigma^{-2}|\partial_r(\mathcal{S}^*)|^{-1} \gtrsim \zeta_n \log\{r(|E|)\},$$

where $r(|E|)$ is the sum of effective resistance weighted edges and ζ_n is any diverging sequence in n, the polynomial-time Algorithm 1 delivers a localisation error upper bounded by

$$\kappa^{-2}\sigma^2|\partial_r(\mathcal{S}^*)| \log\{r(|E|)\}.$$

We further prove that, aside from the $\log\{r(|E|)\}$ term, this rate is minimax optimal.

• For partitions containing an unknown number of elements, we provide a general procedure for localisation and discuss localisation rates under a much stronger signal-to-noise ration condition.

• We illustrate the strength of our methodology in a variety of experiments.

We emphasise that the localisation task over general graph-structured signals turns out to be fundamentally different, in both its theoretical and computational aspects, from the detection task of testing for the presence of an anomalous cluster. See the discussions in Sections 2.1 and 3.3.

1.1 Problem Setup

We formalise our model and the localisation task.

Assumption 1 (Model). *Let $G = (V, E)$ be a fixed connected graph with vertex set $V = \{1, \ldots, n\}$ and edge set $E \subseteq V \times V$. We observe a random vector $Y = (Y_1, \ldots, Y_n)^\top$ whose coordinates correspond to the vertices of G and satisfy, for each $i \in V$,*

$$Y_i = \mu_i^* + \varepsilon_i, \tag{1}$$

where $\mu^* = (\mu_1^*, \ldots, \mu_n^*)^\top$ is a piece-wise constant mean vector and the errors $\{\varepsilon_i\}_{i\in V}$ are centred i.i.d. sub-Gaussian random variables with sub-Gaussian parameter $\sigma > 0$. We let $\mathcal{S}^* = \{S_1^*, \ldots, S_{K^*}^*\}$ be the partition of V supporting the constancy regions of μ^*. In detail, for some $(f_1^*, \ldots, f_{K^*}^*)^\top \in \mathbb{R}^{K^*}$, it holds that (i) for each $k \in \{1, \ldots, K^*\}$, $\mu_i^* = f_k^*$ for all $i \in S_k^*$; and (ii) for any $k, l \in \{1, \ldots, K^*\}$ with $k \neq l$, if $\partial(S_k^*, S_l^*) = \{(i, j) \in$

$E : i \in S_k^*, j \in S_l^*\} \neq \emptyset$, then $f_k^* \neq f_l^*$. The mean vector μ^* and its associated partition $\mathcal{S}^*$ are unknown.

We remark that the whole graph G is required to be connected, but each element of the partition is not required to induce a connected sub-graph, as it is customary in the literature on detection in graph-valued signals; see, e.g. Arias-Castro et al. (2011a). This is demonstrated numerically in Case 4 in Section 4.

Our goal is to recover the partition $\mathcal{S}^*$ accurately. Specifically, under the structural assumption detailed in Assumption 2, we seek an estimator of the partition $\widehat{\mathcal{S}}$ such that the Hausdorff distance between $\widehat{\mathcal{S}}$ and $\mathcal{S}^*$ normalised by the node size n vanishes, as the node size goes to infinity, i.e.

$$d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*) = \max\left\{ d_{\mathrm{H}_1}(\widehat{\mathcal{S}}, \mathcal{S}^*),\, d_{\mathrm{H}_1}(\mathcal{S}^*, \widehat{\mathcal{S}}) \right\} = o(n), \tag{2}$$

as $n \to \infty$, where for any two partitions $\mathcal{A}$ and $\mathcal{B}$ of V we set $d_{\mathrm{H}_1}(\mathcal{A}, \mathcal{B}) = \max_{A\in\mathcal{A}} \min_{B\in\mathcal{B}} |A \triangle B|$ with $A \triangle B = (A \cup B) \setminus (A \cap B)$.

1.2 Notation

For any $\delta > 0$, let $\delta\mathbb{Z} = \{m\delta\}_{m\in\mathbb{Z}}$. For any $x \in \mathbb{R}$ and any $\delta > 0$, let x^δ be a closest value to x in $\delta\mathbb{Z}$. For any $\delta > 0$ and any $\mu \in \mathbb{R}^n$, a $\delta\mathbb{Z}$-expansion of μ is any other vector $\widetilde{\mu} \in \mathbb{R}^n$ such that there exists a single value $c \in \delta\mathbb{Z}$ satisfying that for every $i \in \{1, \ldots, n\}$, either $\widetilde{\mu}_i = \mu_i$ or $\widetilde{\mu}_i = c$. For any partition $\mathcal{S} = \{S_1, \ldots, S_K\}$, with $K \geq 2$, let $\partial(\mathcal{S}) = \{(i,j) \in E : i \in S_k, j \in S_l, k \neq l\}$. For any edge weighting $w : E \to \mathbb{R}_+$ and any partition $\mathcal{S} = \{S_1, \ldots, S_K\}$, $K \geq 2$, let

$$|\partial_w(\mathcal{S})| = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{i\in S_k, j\in S_l} w(i,j)\mathbf{1}\{(i,j) \in E\} \tag{3}$$

and $w(|E|) = \sum_{(i,j)\in E} w(i,j)$. A spanning tree $\mathcal{T}$ of G is a sub-graph that is a tree which includes all the vertices of G. Let $r : E \to \mathbb{R}_+$ be the effective resistance edge weights, that is

$$r(i,j) = \frac{\#\text{ spanning trees that include } (i,j)}{\#\text{ spanning trees}}. \tag{4}$$

For any connected graph G with n nodes and any vector $v \in \mathbb{R}^n$, we say $\mathcal{S}_v = \{S_1, \ldots, S_K\}$ is the partition induced by v, if and only if $\mathcal{S}_v$ is the partition with the smallest size that v has constant values in each S_k, $k \in \{1, \ldots, K\}$.

2 INFORMATION THEORETIC BOUNDS

It is of our primary interest to understand the fundamental limits of localising the true partition of a general graph. We first characterise the hardness of the problem using the following model parameters.

Definition 1. *With the notation in Assumption 1, when $K^* \geq 2$, let $\Delta = \min_k |S_k^*|$ and $\kappa = \min_{\substack{k\neq l \\ \partial(S_k^*, S_l^*)\neq\emptyset}} |f_k^* - f_l^*|$ be the minimal piece size and the minimal jump size.*

Thus, for each distribution P specifying a change point model as described in Assumption 1, we can associate the quantities Δ, κ and $|\partial_r(\mathcal{S}^*)|$. For simplicity, in our notation we omit the dependence on P.

We will show that the problem of recovering the partition is completely characterised by: κ the minimal jump size, Δ the minimal size of constant signal, σ the fluctuation of the noise and $|\partial_r(\mathcal{S}^*)|$ the connectivity between the partition, where r consists of the effective resistance edge weights. Below, we firstly show in Proposition 1 that, if $\kappa^2\Delta\sigma^{-2} \leq c|\partial_r(\mathcal{S}^*)|$, where $c > 0$ is an absolute constant, then from an information-theoretic point of view, no algorithm is guaranteed to provide a consistent partition estimator, in the sense of (2).

Proposition 1. *Let $\mathcal{P}_1$ be the collection of joint distributions that*

$$\mathcal{P}_1 = \left\{ P : \Delta = \min\left\{ \lfloor c\kappa^{-2}\sigma^2 |\partial_r(\mathcal{S}^*)| \rfloor,\, \lfloor n/4 \rfloor \right\} \right\}.$$

It holds that $\inf_{\widehat{\mathcal{S}}} \sup_{P\in\mathcal{P}_1} \mathbb{E}_P\{d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*)\} \geq c_1 n$, where the infimum is taken over all possible partitions of V and $c_1 > 0$ is an absolute constant.

Next, we demonstrate that, provided that $\kappa^2\Delta\sigma^{-2} \geq C|\partial_r(\mathcal{S}^*)|$, where $C > 0$ is an absolute constant, the quantity $\kappa^{-2}\sigma^2|\partial_r(\mathcal{S}^*)|$ is a minimax lower bound on the localisation error.

Proposition 2. *Let $\mathcal{P}_2$ be the collection of joint distributions that $\mathcal{P}_2 = \left\{P : \kappa^2\Delta\sigma^{-2} > C|\partial_r(\mathcal{S}^*)|\right\}$, where $C > 0$ is an absolute constant. It holds that $\inf_{\widehat{\mathcal{S}}} \sup_{P\in\mathcal{P}_2} \mathbb{E}_P(d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*)) \geq C_1\kappa^{-2}\sigma^2|\partial_r(\mathcal{S}^*)|$, where the infimum is taken over all possible partitions of V and $C_1 > 0$ is an absolute constant.*

2.1 Connections with Other Literature

While our results for general graphs appear to be new, the special case of the chain graph has been thoroughly studied. Indeed, the partition recovery problem in the chain graph corresponds to the localisation task in the change point literature. In this case, it has been shown (e.g. Wang et al., 2020; Verzelen et al., 2020; Chan and Walther, 2013) that for $K^* = 2$ (i.e. $|\partial_r(\mathcal{S}^*)| = 1$), if $\kappa^2\Delta\sigma^{-2} \lesssim \log(n)$, then there is no algorithm guaranteed to be consistent; and provided that $\kappa^2\Delta\sigma^{-2} \gtrsim \zeta_n$, a minimax lower bound on the error is $\kappa^{-2}\sigma^2$. Ignoring the logarithmic factors, we see that the results

we have derived for general graphs in Propositions 1 and 2 include the known results on chain graphs with $K^* = 2$. Another special case of a general graph is the d-dimensional lattice. To the best of our knowledge, the only known result comes from Padilla et al. (2021), which studied localisation errors of regions constructed from unions of rectangles. Under some stronger model assumptions, Padilla et al. (2021) showed that the estimation error, up to a logarithmic factor, is of order $\kappa^{-2}\sigma^2$.

As for general graphs, we are not aware of any results in terms of the partition recovery accuracy in the existing work, but there has been a line of work on estimating the whole signal over the graphs, a.k.a. de-noising, (e.g. Jung et al., 2018; Kuthe and Rahmann, 2020; Fan and Guan, 2018; Han, 2019; Sharpnack et al., 2012; Hallac et al., 2015; Padilla et al., 2018). We see partition recovery and de-noising as two closely related but different topics. For instance, with chain graphs, it is well-understood that the fused lasso estimator (Tibshirani et al., 2005) is optimal for de-noising but sub-optimal in localising change points (Lin et al., 2017). Fan and Guan (2018) showed that the minimax optimal rate for de-noising a graph-structured signal with piece-wise constant mean μ^* is of order $\sigma^2|\partial_r(\mathcal{S}^*)|\log(n/|\partial_r(\mathcal{S}^*)|)$ and is achieved by iterating Algorithm 3. Proposition 2 suggests that the relationship between the minimax optimal rate for de-noising and localisation may be informally stated as

$$\text{localization rate} \asymp \kappa^{-2} \times \text{de-noising rate},$$

a connection that has been previously established in the change point literature. More discussions with Fan and Guan (2018) will be provided in Section 3.3 after we present all theoretical results.

Another relevant and closely-related problem is that of detecting an abnormal cluster in a general graph (e.g. Arias-Castro et al., 2008; Addario-Berry et al., 2010; Arias-Castro et al., 2011a,b; Hall and Jin, 2010). For illustration purposes, we use the results in Addario-Berry et al. (2010) to compare with our findings. Translated into our notation, the results in Addario-Berry et al. (2010) establish that, from a testing perspective, the detection boundary is

$$\kappa^2\Delta\sigma^{-2} \asymp \log(\#\text{ candidate sub-graphs}). \tag{5}$$

Note that, without any additional assumption on the graph or on the class of candidate clusters, there are 2^n possible sub-graphs and the result (5) reads as

$$\kappa^2\Delta\sigma^{-2} \asymp n, \tag{6}$$

which appears to be much stronger than the condition we require in Proposition 2. One would then argue that this conclusion seems to contradict the conventional wisdom that testing is easier than estimation. However, as what we will show later, this is in fact not the case, as the type of distributions and graph topologies for which (6) holds form a subset of $\mathcal{P}_1$ in Proposition 1.

Proposition 3. *Consider the family of distributions*

$$\mathcal{P}_3 = \left\{P : \kappa^2\Delta\sigma^{-2} = cn,\ E = \{(i,j), 1 \le i < j \le n\}\right\},$$

where $c > 0$ is an absolute constant. It holds that $\inf_{\widehat{\mathcal{S}}}\sup_{P\in\mathcal{P}_3}\mathbb{E}_P\{d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*)\} \ge c_1 n$, where the infimum is taken over all possible partitions of V and $c_1 > 0$ is an absolute constant.

Note that in Proposition 3, the definition $\mathcal{P}_3$ only includes complete graphs. It follows from Lemma 4 that $|\partial_r(\mathcal{S})|$ can be as large as of order n for complete graphs. Proposition 3, therefore, can be seen as a special case of Proposition 1.

Lemma 4. *For any complete graph G with n nodes and any nontrivial partition $\mathcal{S} = \{S_1, S_2\}$, $S_1, S_2 \neq \emptyset$, $|S_1| = n_1$ and $|S_2| = n_2$, it holds that $\partial_r(\mathcal{S}) = 2n_1n_2n^{-1}$.*

The requirements in (6) and in $\mathcal{P}_3$ are rather pessimistic. Since $\Delta \le n$ by definition, they in fact assume that the signal strength κ/σ should be at least of order $\sqrt{n}$, under which, off by a logarithmic factor, one may simply threshold the node values, completely ignore the graph structure and obtain a consistent detection and also partition. This highlights the role of "sparsity" in the models: for detection problems, (5) shows that a small class of candidate models requires a much smaller signal strength κ/σ; and for localisation problems, Proposition 2 shows that a small cut $|\partial_r(\mathcal{S})|$ requires a much smaller signal strength κ/σ.

3 CONSISTENT PARTITION RECOVERY

In this section, we introduce a polynomial-time algorithm for estimating the partition induced by the piece-wise constant mean vector as in Assumption 1. Towards that goal, we study the approximate solution to the least squared problem with weighted ℓ_0 penalty considered by Fan and Guan (2018). Given a graph $G = (V, E)$, observations $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, tuning parameters $\tau, \delta > 0$ and an edge weighting $w \ge r$, consider the objective function

$$F_w(\mu) = \frac{1}{2}\|Y - \mu\|^2 + \lambda \sum_{\{i,j\}\in E} w(i,j)\mathbf{1}\{\mu_i \neq \mu_j\}. \tag{7}$$

Throughout $\widehat{\mu}$ denotes a $(\tau, \delta\mathbb{Z})$-local-minimiser of (7), defined next.

Definition 2. *For $\delta > 0$ and $\tau \geq 0$, a $(\tau, \delta\mathbb{Z})$-local-minimiser of (7) is any $\mu \in \mathbb{R}^n$ such that for every $\delta\mathbb{Z}$-expansion $\widetilde{\mu}$ of μ, $F(\mu) - \tau \leq F(\widetilde{\mu})$.*

When the edge weights satisfy $w(i,j) = \mathbf{1}\{(i,j) \in E\}$, the function (7) is also known as the Potts functional (Potts, 1952). In chain graphs, the minimizer of (7) has been succesfully used for change point detection and is thoroughly studied (e.g. Wang et al., 2020; Chan and Walther, 2013; Verzelen et al., 2020). An exact minimiser of (7) in chain graphs can be calculated in polynomial time (Friedrich et al., 2008), but this is not the case for general graphs. Fan and Guan (2018) proposed a polynomial-time algorithm, built upon the α-expansion procedure (Boykov et al., 2001). Fan and Guan (2018) also showed that for the de-noising task, a $(\tau, \delta\mathbb{Z})$-local-minimiser of (7) is optimal, when the weights are the effective resistance.

As our goal is to recover the partition, we consider a simpler variant of Algorithm 1 in Fan and Guan (2018). We first focus on the case of K^* known and equal to 2 – a standard setting for detection (see, .e.g, Arias-Castro et al., 2011a) – for which we obtain nearly optimal minimax rates. We then extend our results to the case of a general, unknown K^*.

3.1 The Case of Known $K^* = 2$

The methodology we propose, detailed in Algorithm 1, is a simple modification of Algorithm 1 in Fan and Guan (2018), which in turn is centred on the α-expansion procedure (Boykov et al., 2001). We include it in Algorithm 2 for completeness.

Algorithm 1 Variant of Algorithm 1 in Fan and Guan (2018). Potts$(G, Y, \delta, \tau, \lambda, w)$

INPUT: $G = (V, E)$, $V = \{1, \ldots, n\}$, $\{Y_i, i \in V\}$, δ, τ, λ, w.
  $\overline{Y} \leftarrow n^{-1} \sum_{i \in V} Y_i$, FLAG $\leftarrow 0$.
  $(Y_{\min}, Y_{\max}) \leftarrow (\min_{i \in V} Y_i, \max_{i \in V} Y_i)$
  $(\widehat{\mu}, \mu_1) \leftarrow ((\overline{Y})^{\otimes n}, 0^{\otimes n}) \in \mathbb{R}^n \times \mathbb{R}^n$
  for each $c \in \delta\mathbb{Z} \cap [Y_{\min}, Y_{\max}]$ **do**
    $\widetilde{\mu} \leftarrow \alpha\mathrm{E}(\widehat{\mu}, Y, G, \lambda, w, c)$ ▷ Algorithm 2
    if $F_w(\widetilde{\mu}) \leq F_w(\widehat{\mu}) - \tau$ **then**
      FLAG $\leftarrow 1$.
      if $F_w(\widetilde{\mu}) \leq F_w(\mu_1)$ **then**
        $\mu_1 \leftarrow \widetilde{\mu}$
      end if
    end if
  end for
  if FLAG $= 0$ **then**
    $\mu_1 \leftarrow \widehat{\mu}$.
  end if
OUTPUT: μ_1

Algorithm 2 α-expansion from Boykov et al. (2001). $\alpha\mathrm{E}(\mu, Y, G, \lambda, w, c)$.

INPUT: $G = (V, E)$, $V = \{1, \ldots, n\}$, $Y, \mu \in \mathbb{R}^n$, $w : E \to \mathbb{R}_+$, $w \geq r$, $\lambda \geq 0$, c.
  Add two vertices `s` and `t`
  for each $i \in V$ **do**
    $\tilde{w}(s, i) \leftarrow (Y_i - c)^2/2$
    $\tilde{w}(t, i) \leftarrow (Y_i - \mu_i)^2 \mathbf{1}\{\mu_i \neq c\} + \infty \mathbf{1}\{\mu_i = c\}$
  end for
  for each $(i, j) \in E$ **do**
    if $\mu_i = \mu_j$ **then**
      $\tilde{w}(i, j) \leftarrow \lambda w(i, j) \mathbf{1}\{\mu_i \neq c\}$
    else
      Add a new vertex $\mathtt{a}_{i,j}$
      $E \leftarrow E \setminus \{(i, j)\} \cup \{(i, \mathtt{a}_{i,j}), (j, \mathtt{a}_{i,j}), (\mathtt{t}, \mathtt{a}_{i,j})\}$
      $\tilde{w}(i, \mathtt{a}_{i,j}) \leftarrow \lambda w(i, j) \mathbf{1}\{\mu_i \neq c\}$
      $\tilde{w}(j, \mathtt{a}_{i,j}) \leftarrow \lambda w(i, j) \mathbf{1}\{\mu_j \neq c\}$
      $\tilde{w}(\mathtt{t}, \mathtt{a}_{i,j}) \leftarrow \lambda w(i, j)$
    end if
  end for
  Find the minimum `s`-`t`-cut (S, T) of the newly constructed graph based on weights $\tilde{w}$ such that $\mathtt{s} \in S$ and $\mathtt{t} \in T$.
  for each $i \in V$ **do**
    $\widetilde{\mu}_i \leftarrow c \mathbf{1}\{i \in T\} + \mu_i \mathbf{1}\{i \in S\}$
  end for
OUTPUT: $\widetilde{\mu}$

In Fan and Guan (2018) Algorithm 2 is called iteratively until the objective function $F_w(\cdot)$ is not improved. This results in a partition of potentially more than two subsets of nodes, which is desirable for the purpose of signal estimation. Since our goal is to recover a partition with two elements, our strategy is different. In the language of Algorithm 1, with the initialiser $\widehat{\mu}_i = \overline{Y}$ for all i, Algorithm 2 is summoned repeatedly in order to find a value $c \in \delta\mathbb{Z}$ and a subset $V_1 \subset V$ such that $F_w(\widehat{\mu})$ reaches its minimum, where $\widehat{\mu}_i = c$, $i \in V_1$, and $\widehat{\mu}_i = \overline{Y}$, $i \in V \setminus V_1$.

To evaluate the computational cost of Algorithm 1, we adapt the proof of Proposition 2.2 in Fan and Guan (2018). Note that there are at most $(Y_{\max} - Y_{\min})/\delta$ different choices of c. For each value c, the augmented graph in Algorithm 2 has $O(|E|)$ vertices and edges. Solving minimum s-t cut using either the Edmonds–Karp or Dinic algorithm requires time $O(|E|^3)$. This implies that the computational cost of Algorithm 1 is of order $O\{|E|^3(Y_{\max} - Y_{\min})\delta^{-1}\}$. We acknowledge that the computational cost is high, despite being polynomial-time. We do not know whether there exist faster algorithms enjoying the same theoretical optimality. In chain graphs, faster algorithms exist, yet theoretically sub-optimal.

Note that edge-weights are used in the α-expansion subroutine and it means that Algorithm 1 is edge-weights dependent. It is naturally the case that the partition recovery performance of Algorithm 1 also depends on the choice of edge-weights.

We state our result in generality by assuming that the edge weights are point-wise larger than the effective resistance, as done in Fan and Guan (2018).

Assumption 2 (Signal-to-noise ratio). *For a certain choice of edge-weights $w : E \to \mathbb{R}_+$, with $w \geq r$, assume that there exists a constant $C_{\mathrm{SNR}} > 0$ such that $\kappa^2\Delta \geq C_{\mathrm{SNR}}\sigma^2|\partial_w(\mathcal{S}^*)|\log\{w(|E|)\}\zeta_n$, where ζ_n is any arbitrarily diverging sequence, as the number of node n grows unbounded.*

Theorem 5 ($K^* = 2$ and K^* is known). *Let $\delta \leq \sigma/\sqrt{n}$, $\tau \leq \sigma^2$ and $\lambda = C_\lambda\sigma^2\log\{w(|E|)\}$, where $C_\lambda > 0$ is an absolute constant. Assume that $K^* = 2$ and K^* is known. Under Assumptions 1 and 2, letting $\widehat{\mu}$ be an output of Algorithm 1, with $w \geq r$, it holds with probability at least $1 - \{w(|E|)\}^{-c}$ that*

$$d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \leq C\kappa^{-2}\sigma^2|\partial_w(\mathcal{S}^*)|\log\{w(|E|)\}, \quad (8)$$

where $c, C > 0$ are absolute constants and $\widehat{\mathcal{S}}$ is the partition induced by $\widehat{\mu}$.

Recalling the consistency definition (2) and in view of Assumption 2, we see that the estimation error satisfies

$$\begin{aligned} n^{-1}d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*) &\leq Cn^{-1}\kappa^{-2}\sigma^2|\partial_w(\mathcal{S}^*)|\log\{w(|E|)\} \\ &\leq CC_{\mathrm{SNR}}\zeta_n^{-1}n^{-1}\Delta \to 0, \end{aligned}$$

which implies that the output of Algorithm 1 provides a consistent partition recovery.

If we choose the effective resistance edge-weights in Algorithm 1, then Theorem 5 and Proposition 2 imply that the error rate we have obtained is nearly minimax optimal.

Furthermore, recall from Proposition 1 that no algorithm is guaranteed to be consistent in the low signal-to-noise regime $\kappa^2\Delta\sigma^{-2} \lesssim |\partial_r(\mathcal{S}^*)|$. On the other hand, letting the edge weights be the effective resistance weights, Theorem 5 shows that consistent estimation is possible when $\kappa^2\Delta\sigma^{-2} \gtrsim |\partial_r(\mathcal{S}^*)|$. Thus, Theorem 5 additionally reveals the existence of a phase transition in the space of model parameters, as in the high signal-to-noise ratio regime, consistent estimation is not only feasible but it can be done at a nearly minimax optimal rate.

Remark 1 (Edge weights). *The edge-weights play an important role in all the results we have shown so far. The two most commonly-used choices are: (1) the 0/1 weighting, i.e. assigning unit weight to each edge, and (2) the effective resistance weighting. Since $r(i,j) \leq 1 = \mathbf{1}\{(i,j) \in E\}$, for all $(i,j) \in E$, for the partition recovery task, our theory suggests that the effective resistance weighting should be preferred to 0/1 weighting. Having said this, when the sizes of V and E are moderately large, it is much more practical to directly adopt the 0/1 weighting rather than calculating the effective resistance. This is also what we do in Section 4.*

3.2 General K^*

In Section 3.1 we only consider the case when $K^* = 2$ and is known. There are ample applications where such cases are interesting and we refer to the Introduction of Arias-Castro et al. (2011a). Despite the popularity of the case $K^* = 2$, there are of course many interesting situations where $K^* \neq 2$. In chain graphs, an ℓ_0-penalised method is shown to be optimal in change point localisation for general K^*, where K^* is seen as the number of change points plus one, and is even allowed to diverge with the number of nodes n (e.g. Wang et al., 2020; Verzelen et al., 2020). This result has been further extended to d-dimensional lattice graphs in Padilla et al. (2021), where under stronger conditions, it is shown that a constrained ℓ_0-penalised estimator (dyadic classification and regression trees, DCART) is able to handle general K^*, despite possessing a gap regarding K^* in relation to a minimax lower bound. For general graphs, when the goal is de-noising, Fan and Guan (2018) showed that an ℓ_0-penalised estimator is optimal for general K^*. For us, with general graphs and partition recovery purpose, we show that the phenomenon is very different.

3.2.1 The Case of $K^* = 1$

The first result we show is when $K^* = 1$, with large probability, the output of Algorithm 1 is constant over the whole graph. This claim is formalised in Proposition 6, which is very similar to Theorem 3.5 in Fan and Guan (2018).

Proposition 6. *Let $\delta \leq \sigma/\sqrt{n}$, $\tau \leq \sigma^2$ and $\lambda = C_\lambda\sigma^2\log\{w(|E|)\}$, where $C_\lambda > 0$ is an absolute constant. Assume that $K^* = 1$. Under Assumption 1, letting $\widehat{\mu}$ be an output of Algorithm 1 with $w \geq r$, it holds with probability at least $1 - \{w(|E|)\}^{-c}$ that $|\widehat{\mathcal{S}}| = 1$, where $\widehat{\mathcal{S}}$ is the partition induced by $\widehat{\mu}$.*

3.2.2 The Case of Unknown $K^* > 1$

One may wish to say that an immediate consequence of Theorem 5 and Proposition 6 is that when $K^* = 2$ but K^* is unknown, Theorem 5 still holds by first conducting Algorithm 1 on the whole graph, then repeatedly and separately on the resulting two pieces. (For completeness, we formalise this procedure in Algorithm 3,

with the notation that for any two partitions $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \cap \mathcal{B}$ is their refinement.)

Algorithm 3 Variant of Algorithm 1

INPUT: $G = (V, E)$, $|V| = n$, $\{Y_i, i \in V\}$, δ, τ, λ, w.
  $\widetilde{\mu} \leftarrow 0^{\otimes n}$, FLAG $\leftarrow 0$
  while FLAG $= 0$ **do**
    $\widehat{\mathcal{S}}, \widetilde{\mathcal{S}} \leftarrow$ the induced partition of $\widetilde{\mu}$
    for each $S \in \widehat{\mathcal{S}}$ **do**
      $\mu_1 \leftarrow \text{Potts}(G_S, Y_{G_S}, \delta, \tau, \lambda, w)$ ▷ Algorithm 1
      $\mathcal{S}_1 \leftarrow$ the induced partition of μ_1
      $\widehat{\mathcal{S}} \leftarrow \widehat{\mathcal{S}} \cap \mathcal{S}_1$
      $\widetilde{\mu}_S \leftarrow \mu_1$ ▷ $\widetilde{\mu}_S$ is the sub-vector of $\widetilde{\mu}$ on S
    end for
    FLAG $= \mathbf{1}\{\widehat{\mathcal{S}} = \widetilde{\mathcal{S}}\}$
  end while
OUTPUT: $\widehat{\mathcal{S}}$

Unfortunately, such result can only hold under a much stronger assumption and with a much worse rate. A direct consequence of Theorem 5 is the following.

Assume that there exists a constant $C_{\text{SNR}} > 0$ such that

$$\kappa^2 \Delta \geq C_{\text{SNR}} \sigma^2 n \log(n) \zeta_n, \tag{9}$$

where ζ_n is any arbitrarily diverging sequence, as n grows unbounded. Let $\delta \leq \sigma/\sqrt{n}$, $\tau \leq \sigma^2$ and $\lambda = C_\lambda \sigma^2 \log\{w(|E|)\}$, where $C_\lambda > 0$ is an absolute constant. Assume that $K^* = 2$ but K^* is unknown, and assume that Assumption 1 holds. Letting $\widehat{\mathcal{S}}$ be an output of Algorithm 3, with $w \geq r$, it holds with probability at least $1 - \{w(|E|)\}^{-c}$ that

$$d_{\text{H}}(\widehat{\mathcal{S}}, \mathcal{S}^*) \leq C \kappa^2 \sigma^{-2} n \log(n) \zeta_n. \tag{10}$$

Comparing Theorem 5 and the aforementioned, we see that when K^* is unknown, the corresponding rates in both the signal-to-noise ratio condition and the estimation error bound, jump from $|\partial_w(\mathcal{S})|$ to n. Essentially, this is due to the complexity of the graphs and we elaborate as follows.

The proof of Theorem 5 is built upon the large probability event defined in (16), and (10) is due to the event

$$\left\{ \left| |A|^{-1/2} \sum_{i \in A} \varepsilon_i \right| \leq Cn\sigma, \quad \forall A \in 2^V \right\},$$

where 2^V is the power set of V. On both events, the difference between the sample and population quantities are upper bounded. The large probability events are constructed based on sub-Gaussian concentration inequalities and a union bound argument. In Theorem 5, since $K^* = 2$ and K^* is assumed to be known, and due to the assumption that G is a connected graph, the union bound argument is based on a certain spanning tree. This is shown in Lemma B.2 in Fan and Guan (2018). In Algorithm 3, due to the different layers of splits, at each layer, one works based on a random sub-graph yielded by the previous layer optimisation. To take this randomness into consideration, one therefore has to consider the complexity of the whole graph.

To further understand how the complexity kicks in, we note that in the change point localisation literature, i.e. when G is a chain graph, there is no such dramatic change in rates for general K^* cases (e.g. Wang et al., 2020). This is due to the fact that there is one and only one spanning tree in a chain graph, and any spanning tree of any sub-graph is a sub-tree of a spanning tree of the whole graph. For a general graph, it is possible to partition a sub-graph without cutting through a spanning tree of the whole graph. Therefore, when applying a union bound argument over random sub-graphs, one has to consider all possible spanning trees. The number of all spanning trees is exponential in n, which holds even when the graph is a k-regular graph, with k a fixed constant (e.g. Alon, 1990), or when the graph is a lattice (e.g. Shrock and Wu, 2000).

Based on the above discussions, one may be led to conjecture that for general K^*, (9) and (10) are in fact optimal. If true, this finding would be rather interesting – if the signal-to-noise ratio has to be as large as required in (9), then for any edge $(i, j) \in E$, one can simply cut this edge if and only if $|Y_i - Y_j| \gtrsim \sigma \log(n)$ and optimally recover the partition.

3.3 Comparisons with Fan and Guan (2018)

Fan and Guan (2018) is the closest-related literature and has heavily inspired our work. We therefore provide a thorough comparison with Fan and Guan (2018). Though the set-up is identical, the contributions of the present paper are markedly different in both the goals and technical features than those in Fan and Guan (2018). Below, we highlight how our work relates to Fan and Guan (2018). (a) Task. Fan and Guan (2018) investigate de-noising or signal estimation of a piece-wise constant signal over a graph while the present paper aims to estimate the boundary of the partition supporting the constancy regions of the signal - a problem that, somewhat surprisingly, has never been thoroughly studied in the literature. Despite their apparent similarities, de-noising and partition recovery are fundamentally different tasks, with different metrics and challenges. For the specific case of the chain graph, the two tasks have led to distinct lines of research and very different results. In particular, the partition recovery problem is effectively equivalent to change point localisation in a time se-

ries. (b) Optimality and assumptions. Both Fan and Guan (2018) and ours provided minimax optimal results, supported by minimax lower bounds. The lower bound proof in Fan and Guan (2018) is an application of Theorem 1(b) in Raskutti et al. (2011), which is based on the metric entropy of ℓ_q balls. We have three lower bound results, which rely on graph theory. (c) Algorithms. Algorithm 1 and that in Fan and Guan (2018) share the same core procedure, which is the α-expansion algorithm Boykov et al. (2001). Our Algorithm 1 is indeed a straightforward adaptation of that in Fan and Guan (2018) by simply imposing that only one split is performed. (d) The proofs of the theoretical guarantees of Algorithm 1. The aforementioned differences between the problems imply that the structure and key technical aspects of the proofs are completely different. Both sets of proofs rely on a large probability event put forward in Fan and Guan (2018), which we directly cite. (e) Other aspects. Due to the different nature of the problem, Fan and Guan (2018) is able to handle a growing number of constancy pieces, while such case becomes challenging and yet to be completely understood in the partition recovery case. However, through comparisons with change point detection in chain graphs and abnormal cluster detection in general graphs, we provide more insights on the role of graph sparsity in such problems.

4 NUMERICAL EXPERIMENTS

We now proceed to evaluate the empirical performance of the proposed method, with four cases described below. In each case we consider data in the 2D grid graph G with $\sqrt{n} \times \sqrt{n}$ nodes, where $K^* = 2$ is known, $n \in \{64^2, 128^2\}$, $\sigma = 1$ and $\kappa \in \{1, 2\}$. We consider three competing methods: Algorithm 1, a variant of the dyadic CART (DCART, Donoho, 1997) and the edge lasso (EL, Sharpnack et al., 2012). For each method and case we report the median Hausdorff distance based on 50 repetitions.

For the implementation of Algorithm 1, we set $\delta = 1/60$ and choose λ from the set of candidates $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$ via the Bayesian information criterion (BIC). Specifically, we first estimate σ^2 with $\hat{\sigma}^2$ given as $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n-1} (Y_{p(i)} - Y_{p(i+1)})^2$, where $p(1), \ldots, p(n)$ form a connected path of the grid graph. With this in hand, for any λ and its corresponding $\tilde{\mu}(\lambda)$ we calculate the score

$$\mathrm{BIC}_\lambda = \sum_{i=1}^{n} \{Y_i - \tilde{\mu}_i(\lambda)\}^2 + \hat{\sigma}^2 v_\lambda \log(n), \qquad (11)$$

where v_λ is the number of connected components induced by $\tilde{\mu}(\lambda)$ in G. We then choose the value of λ that minimizes BIC_λ and report the performance of Algorithm 1 based on this choice. The implementation of Algorithm 1 is done in Matlab using the package `gco-v3.0` (Boykov et al., 2001).

Table 1: Median Hausdorff distances of 50 repetitions under different cases. DCART: dyadic CART; EL: edge Lasso; κ: jump size; and n: number of nodes.

CASE	κ	$\sqrt{n}$	ALG. 1	DCART	EL
1	1.0	128	**56**	115	922
1	2.0	128	33	86	**21**
1	1.0	64	**29**	152	1630
1	2.0	64	12	34	**10**
2	1.0	128	**122**	230	516
2	2.0	128	42	160	**32**
2	1.0	64	**89**	294	1601
2	2.0	64	24	63	**15**
3	1.0	128	36	63	**7**
3	2.0	128	**1**	15	**1**
3	1.0	64	**20**	47	1392
3	2.0	64	10	15	**1**
4	1.0	128	**514**	604	6669
4	2.0	128	**83**	199	671
4	1.0	64	**387**	424	1836
4	2.0	64	**49**	108	226

For DCART, we first compute their values with the penalty $\lambda \in \{10^{-1+4j/19} : j = 0, 1, \ldots, 19\}$. We then select λ that minimises the BIC in (11), replacing $\tilde{\mu}(\lambda)$ with the corresponding DCART estimator. This produces an estimator $\hat{\mu}$. However, assuming that $K^* = 2$ is known, we let Λ be the set of unique values of $\hat{\mu}$ and perform k-means clustering on the elements of Λ setting the number of clusters to two. Let C_1 and C_2 be the centres obtained from the k-means clustering. Define $\tilde{\mu}_i = C_1$, if $\hat{\mu}_i$ is assigned to the centre C_1; and $\tilde{\mu}_i = C_2$, otherwise. The final estimator is the partition induced by $\tilde{\mu}$. The computation for the resulting procedure was done in R.

For EL, we proceed as we do with DCART employing the k-means post-processing and BIC model selection. The choices of the penalty parameter are $\{10^{-2+6j/19} : j \in \{1, \ldots, 19\}\}$.

The specific cases considered as as follows. In each case, we generate data as $Y_{i,j} = \mu^*_{i,j} + \varepsilon_{i,j}$, where $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $i, j = 1, \ldots, \sqrt{n}$ and μ^* is specified below.

Case 1. Let

$$\mu^*_{i,j} = \begin{cases} \kappa, & (i - \sqrt{n}/4)^2 + (j - \sqrt{n}/4)^2 < (\sqrt{n}/5)^2, \\ 0, & \text{otherwise.} \end{cases}$$

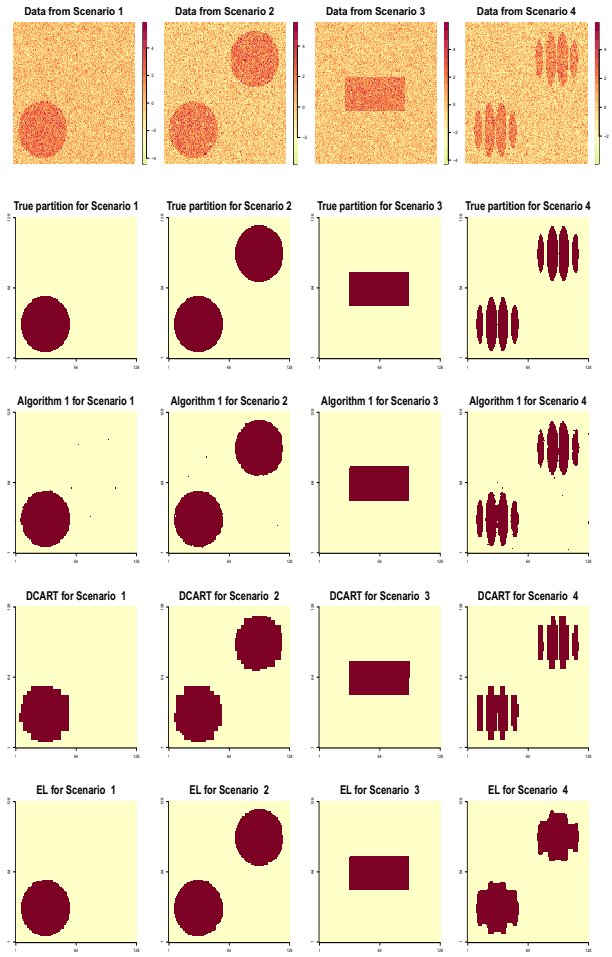

Figure 1: From top to bottom: instances of data, signal patterns, estimators of Algorithm 1, dyadic CART (DCART) and edge lasso (EL). From left to right: Cases 1, 2, 3 and 4, with $n = 128^2$ and $\kappa = 2$.

Case 2. Let

$$\mu^*_{i,j} = \begin{cases} \kappa, & (i - \sqrt{n}/4)^2 + (j - \sqrt{n}/4)^2 < (\sqrt{n}/5)^2 \\ & \text{or } (i - 3\sqrt{n}/4)^2 + (j - 3\sqrt{n}/4)^2 \\ & \quad < (\sqrt{n}/5)^2, \\ 0, & \text{otherwise.} \end{cases}$$

Case 3. Let

$$\mu^*_{i,j} = \begin{cases} \kappa, & |i - \sqrt{n}/2| < \sqrt{n}/4 \\ & \text{and } |j - \sqrt{n}/2| < \sqrt{n}/4, \\ 0, & \text{otherwise.} \end{cases}$$

Case 4. Let

$$\mu^*_{i,j} = \begin{cases} \kappa, & (i - \sqrt{n}/4)^2 + (j - \sqrt{n}/4)^2 \\ & \quad < |\cos(10\,\pi\, i/n)|(\sqrt{n}/5)^2 \\ & \text{or } (i - 3\sqrt{n}/4)^2 + (j - 3\sqrt{n}/4)^2 \\ & \quad < |\cos(10\,\pi\, i/n)|(\sqrt{n}/5)^2, \\ 0, & \text{otherwise.} \end{cases}$$

Visualisations of instances of generated data, true signals and different estimators are given in Figure 1. We can see that the signal in Case 1 has two pieces, a circle and its complement. Case 2 is based on two separated circles and their complement, but the signals on the two circles are the same. The partition in Case 3 is a rectangle and its complement, making it the most attractive case for using the variant of DCART. The final case is perhaps the most challenging since the true signal has multiple pieces with boundaries that rapidly change in some regions. Recall in Assumption 1, S_k's are assumed to possess constant signal values, but are not necessarily connected. Under this assumption, all four cases have $K^* = 2$. Cases 2 and 4 show that although our method and theory are designed for $K^* = 2$, we enjoy great model complexity that we can handle cases with more than two connected constancy regions.

The results measured by the mean Hausdorff distances of 50 repetitions are shown in Table 1, where we can see that Algorithm 1 performs better than the variant of DCART across all cases considered. The EL can be competitive in cases where there is large signal-to-noise ratio but suffer greatly otherwise.

5 CONCLUSION

In this paper, we study the partition recovery problem in general graphs, where nodes are associated with independent random variables. We have shown that an ℓ_0-penalised estimator is optimal when it is known that $K^* = 2$. We have further derived a phase transition phenomenon, establishing the fundamental limits in the partition recovery problem in general graphs. For general K^*, we have provided some seemingly unsatisfactory results, which we conjecture to be optimal and which prompt us discuss some crucial difference between the de-noising and partition recovery problems in general graphs.

Acknowledgements

All the authors thanks to the reviewers for constructive comments. Yu is partially funded by EPSRC EP/V013432/1. Madrid Padilla and Rinaldo are partially funded by NSF DMS 2015489.

References

Addario-Berry, L., Broutin, N., Devroye, L., and Lugosi, G. (2010). On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092.

Alon, N. (1990). The number of spanning trees in regular graphs. *Random Structures & Algorithms*, 1(2):175–181.

Arias-Castro, E., Candes, E. J., and Durand, A. (2011a). Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304.

Arias-Castro, E., Candes, E. J., Helgason, H., and Zeitouni, O. (2008). Searching for a trail of evidence in a maze. *The Annals of Statistics*, pages 1726–1757.

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011b). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, pages 2533–2556.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239.

Chaiken, S. and Kleitman, D. J. (1978). Matrix tree theorems. *Journal of combinatorial theory, Series A*, 24(3):377–381.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica*, pages 409–428.

Di Blasi, M., Gullo, S., Mancinelli, E., Freda, M. F., Esposito, G., Gelo, O. C. G., Lagetto, G., Giordano, C., Mazzeschi, C., Pazzagli, C., et al. (2021). Psychological distress associated with the covid-19 lockdown: A two-wave network analysis. *Journal of affective disorders*, 284:18–26.

Donoho, D. L. (1997). Cart and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5):1870–1911.

Fan, Z. and Guan, L. (2018). Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *Annals of Statistics*, 46(6B):3217–3245.

Friedrich, F., Kempe, A., Liebscher, V., and Winkler, G. (2008). Complexity penalized m-estimation: fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732.

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396.

Han, Q. (2019). Set structured global empirical risk minimizers are rate optimal in general dimensions. *arXiv preprint arXiv:1905.12823*.

Jung, A., Tran, N., and Mara, A. (2018). When is network lasso accurate? *Frontiers in Applied Mathematics and Statistics*, 3:28.

Khan, D. M., Kamel, N., Muzaimi, M., and Hill, T. (2021). Effective connectivity for default mode network analysis of alcoholism. *Brain Connectivity*, 11(1):12–29.

Kuthe, E. and Rahmann, S. (2020). Engineering fused lasso solvers on trees. In *18th International Symposium on Experimental Algorithms (SEA 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Lin, K., Sharpnack, J. L., Rinaldo, A., and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. *Advances in neural information processing systems*, 30.

Liu, S., Caporin, M., and Paterlini, S. (2021). Dynamic network analysis of north american financial institutions. *Finance Research Letters*, page 101921.

Mina, M., Messier, C., Duveneck, M., Fortin, M.-J., and Aquilué, N. (2021). Network analysis can guide resilience-based management in forest landscapes under global change. *Ecological Applications*, 31(1):e2221.

Odeyomi, O. T., Kwon, H. M., and Murrell, D. A. (2020). Time-varying truth prediction in social networks using online learning. In *2020 International Conference on Computing, Networking and Communications (ICNC)*, pages 171–175. IEEE.

Padilla, O. H. M., Sharpnack, J., Scott, J. G., and Tibshirani, R. J. (2018). The dfs fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:176–1.

Padilla, O. H. M., Yu, Y., and Rinaldo, A. (2021). Lattice partition recovery with dyadic cart. *arXiv preprint arXiv:2105.13504*.

Potts, R. B. (1952). Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pages 106–109. Cambridge University Press.

Raimondi, D., Simm, J., Arany, A., and Moreau, Y. (2021). A novel method for data fusion over entity-relation graphs and its application to protein–protein interaction prediction. *Bioinformatics*.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994.

Sharpnack, J., Singh, A., and Krishnamurthy, A. (2013). Detecting activations over graphs using spanning tree wavelet bases. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 536–544.

Sharpnack, J., Singh, A., and Rinaldo, A. (2012). Sparsistency of the edge lasso over graphs. In *Artificial Intelligence and Statistics*, pages 1028–1036. PMLR.

Shrock, R. and Wu, F. Y. (2000). Spanning trees on graphs and lattices in d dimensions. *Journal of Physics A: Mathematical and General*, 33(21):3881.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Verzelen, N., Fromont, M., Lerasle, M., and Reynaud-Bouret, P. (2020). Optimal change-point detection and localization. *arXiv preprint arXiv:2010.11470*.

Wang, D., Yu, Y., and Rinaldo, A. (2020). Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.

Supplementary Material: Optimal Partition Recovery in General Graphs

A PROOFS OF RESULTS IN SECTION 2

Propositions 1 and 2 can be directly shown by noticing that chain graphs are special cases of general graphs. In chain graphs with $K^* = 2$, it holds that $|\partial_r(\mathcal{S}^*)| = 1$. The results of Propositions 1 and 2 follow directly from Lemmas 1 and 2 in Wang et al. (2020). In this section, we provide different proofs allowing $|\partial_r(\mathcal{S}^*)|$ to depend on the sample size n.

Proof of Proposition 1. We prove the result by the Le Cam lemma (e.g. Lemma 1 in Yu, 1997).

Constructing distributions We consider a complete n-node graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ and $E = \{(i, j), 1 \leq i < j \leq n\}$.

Given an absolute constant $c_1 > 0$, without loss of generality, we assume that $n/4$, $M = n/\{4c_1 \log(n)\}$ and $c_1 \log(n)$ are all positive integers. For $l \in \{1, \ldots, M\}$, let $\widetilde{u}_l \in \mathbb{R}^n$ be such that the $i$the coordinate of $\widetilde{u}_l$, $i = 1, \ldots, n$, satisfies that

$$\widetilde{u}_l(i) = \begin{cases} \sigma, & i \in \{(l-1)c_1 \log(n) + 1, \ldots, lc_1 \log(n)\}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\widetilde{v}_l \in \mathbb{R}^n$ be such that $\widetilde{v}_l(i) = \widetilde{u}_l(n - i + 1)$, $i = 1, \ldots, n$. Let $\widetilde{P}_l$ and $\widetilde{Q}_l$ be multivariate Gaussian distributions $\mathcal{N}(\widetilde{u}_l, \sigma^2 I_n)$ and $\mathcal{N}(\widetilde{v}_l, \sigma^2 I_n)$, respectively and set

$$\widetilde{P} = \frac{1}{M} \sum_{l=1}^{M} \widetilde{P}_l \quad \text{and} \quad \widetilde{Q} = \frac{1}{M} \sum_{l=1}^{M} \widetilde{Q}_l.$$

Note that for each $l \in \{1, \ldots, M\}$, $\widetilde{P}_l$ has two pieces, with $\Delta = c_1 \log(n)$ and the smaller piece is contained in $\{1, \ldots, n/4\}$. As a result, it holds that

$$\kappa \sqrt{\Delta}/\sigma = c_1 \log(n). \tag{12}$$

In addition, since we are considering complete graphs, each edge (i, j) has effective resistance weight

$$r(i, j) = \frac{(n-1)n^{n-2}}{n(n-1)/2} = 2n^{n-3},$$

following from the proof of Lemma 4. Then

$$|\partial_r(\mathcal{S})| = \frac{2n^{n-3}}{n^{n-2}} c_1 \log(n) \{n - c_1 \log(n)\} = \frac{2}{n} c_1 \log(n) \{n - c_1 \log(n)\}.$$

Provided that $n \geq 2c_1 \log(n)$, it holds that

$$c_1 \log(n) \leq |\partial_r(\mathcal{S})| \leq 2c_1 \log(n),$$

which implies that there exists an absolute constant $c_2 > 0$ such that

$$|\partial_r(\mathcal{S})| = c_2 \log(n). \tag{13}$$

Combining (12) and (13), we have that

$$\kappa \sqrt{\Delta}/\sigma = c_1 c_2^{-1} |\partial_r(\mathcal{S})|$$

and therefore $\widetilde{P}_l \in \mathcal{P}$. The same arguments show that $\widetilde{Q}_l \in \mathcal{P}$, for all $l \in \{1, \ldots, M\}$. By construction, we also note that

$$d_{\mathrm{H}}(\mathcal{S}_{\widetilde{P}_l}, \mathcal{S}_{\widetilde{Q}_m}) \geq n/2, \quad l, m \in \{1, \ldots, M\}.$$

It then follows from Le Cam's lemma (e.g. Lemma 1 in Yu, 1997) that

$$\inf_{\widehat{\mathcal{S}}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left\{ d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}(P)) \right\} \geq \frac{n}{4} \left\{ 1 - d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q}) \right\},$$

where $d_{\mathrm{TV}}(\cdot, \cdot)$ is the total variation distance between two probability measures and the infimum is over all estimators $\widehat{\mathcal{S}}$.

Upper bounding the total variation distance Let $u_l \in \mathbb{R}^{n/2}$ be a sub-vector of $\widetilde{u}_l$ containing only the first $n/2$ entries. Let P_l and P_0 be the multivariate Gaussian distributions $\mathcal{N}(u_l, \sigma^2 I_{n/2})$ and $\mathcal{N}(0, \sigma^2 I_{n/2})$, respectively. Due to the symmetry between $\widetilde{u}_l$ and $\widetilde{v}_l$, it holds that

$$d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q}) \leq 2 d_{\mathrm{TV}}(P, P_0), \quad \text{where } P = \frac{1}{M} \sum_{m=1}^{M} P_l.$$

Since $d_{\mathrm{TV}}(P, P_0) \leq \sqrt{\chi^2(P, P_0))}$ (e.g. Equation 2.27 in Tsybakov, 2008), it suffices to provide an upper bound on $\chi^2(P, P_0)$. We have that

$$\begin{aligned} \chi^2(P, P_0) &= \frac{1}{M^2} \sum_{l,m=1}^{M} \mathbb{E}_{P_0} \left(\frac{dP_l dP_m}{dP_0 dP_0} \right) - 1 = \frac{1}{M^2} \sum_{l,m=1}^{M} \exp\left(\frac{u_l^\top u_m}{\sigma^2} \right) - 1 \\ &= \frac{1}{M^2} \left[\sum_{l=1}^{M} \exp\{c_1 \log(n)\} + M(M-1) \right] - 1 = M^{-1}(n^{c_1} - 1). \end{aligned}$$

Recall that $M = n/\{4c_1 \log(n)\}$. Therefore, for $c_1 \in (0, 1)$, there exists a sufficiently large $n_0 = n(c_1)$ such that for any $n \geq n_0$, $M^{-1}(n^{c_1} - 1) \leq 1/16$. We therefore complete the proof. □

Proof of Proposition 2. We prove the result by the Le Cam Lemma (e.g. Lemma 1 in Yu, 1997).

Constructing distributions We consider a complete n-node graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ and $E = \{(i, j), 1 \leq i < j \leq n\}$.

Given an absolute constant $c_1 > 0$, without loss of generality, we assume that $c_1 \log(n)$ and M are positive integers, where M satisfies that $2M + 1 = n/\{c_1 \log(n)\}$. For $l \in \{1, \ldots, M\}$, let $\widetilde{u}_l \in \mathbb{R}^n$ be such that the $i$the coordinate of $\widetilde{u}_l(i)$, $i = 1, \ldots, n$, satisfies that

$$\widetilde{u}_l(i) = \begin{cases} \sigma, & i \in \{(l-1)c_1 \log(n) + 1, \ldots, l c_1 \log(n)\}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\widetilde{v}_l \in \mathbb{R}^n$ be such that $\widetilde{v}_l(i) = \widetilde{u}(n - i + 1)$, $i = 1, \ldots, n$. Let $\widetilde{P}_l$ and $\widetilde{Q}_l$ be multivariate Gaussian distributions $\mathcal{N}(\widetilde{u}_l, \sigma^2 I_n)$ and $\mathcal{N}(\widetilde{v}_l, \sigma^2 I_n)$, respectively and set

$$\widetilde{P} = \frac{1}{M} \sum_{l=1}^{M} \widetilde{P}_l \quad \text{and} \quad \widetilde{Q} = \frac{1}{M} \sum_{l=1}^{M} \widetilde{Q}_l.$$

Following the identical arguments as those in the proof of Proposition 1, we have that $\widetilde{P}_l, \widetilde{Q}_l \in \mathcal{P}$. By construction, we also note that

$$d_{\mathrm{H}}(\mathcal{S}_{\widetilde{P}_l}, \mathcal{S}_{\widetilde{Q}_m}) \geq c_1 \log(n), \quad l, m \in \{1, \ldots, M\}.$$

It then follows from Le Cam's lemma (e.g. Lemma 1 in Yu, 1997) that

$$\inf_{\widehat{\mathcal{S}}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left\{ d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S}(P)) \right\} \geq c_1 \log(n)/2 \left\{ 1 - d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q}) \right\}.$$

Upper bounding the total variation distance Let $u_l \in \mathbb{R}^{n/2}$ be a sub-vector of $\widetilde{u}_l$ containing only the first $n/2$ entries. Let P_l and P_0 be the multivariate Gaussian distributions $\mathcal{N}(u_l, \sigma^2 I_{n/2})$ and $\mathcal{N}(0, \sigma^2 I_{n/2})$, respectively. It follows from the same arguments as those in the proof of Proposition 1 that

$$\begin{aligned} d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q}) \leq 2 d_{\mathrm{TV}}(P, P_0) \leq 2\sqrt{\chi^2(P, P_0)} \\ = 2\sqrt{\frac{1}{M^2}\sum_{l,m=1}^{M} \mathbb{E}_{P_0}\left(\frac{dP_l dP_m}{dP_0 dP_0}\right) - 1} = 2\sqrt{M^{-1} n^{c_1} - 1}, \end{aligned}$$

where the equation follows from the identical arguments as those in the proof of Proposition 1. Recall that $M = [n/\{c_1 \log(n)\} - 1]/2$. Therefore, for any $c_1 \in (0, 1)$, there exists a large enough n such that $d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q}) < 1/2$ and we conclude the proof. □

Proof of Proposition 3. Recall that $E = \emptyset$. We therefore only need to construct nodes in the proof.

$0.2n$-packing number of 2^V. We first construct a collection of distributions. Let

$$\mathcal{W} = \left\{A \in 2^V : |A| \in [0.49n, 0.51n]\right\}.$$

We introduce independent Rademacher random variables X_i associated with each node $i \in V$. Therefore

$$\mathbb{P}\left\{\left|\sum_{i \in V} X_i - n/2\right| \geq 0.01n\right\} \leq 2\exp(-0.0002n).$$

This means there exists an absolute constant $c_1 \in [1, 2)$ and a sufficiently large $n_0 \in \mathbb{N}^*$, such that for any $n \geq n_0$,

$$|\mathcal{W}| \geq \{1 - 2\exp(-0.0002n)\} 2^n \geq c_1^n.$$

For any $A \in \mathcal{W}$, we let

$$\mathcal{W}_A = \{B \in \mathcal{W} : |A \triangle B| \leq 0.2n\}.$$

Note that for any fixed A,

$$|\mathcal{W}_A| \leq \left|\left\{B \in 2^V : |B| \leq 0.2n\right\}\right|.$$

Using the Rademacher random variables X_i's, we have that

$$\mathbb{P}\left\{\sum_{i \in V} X_i - n/2 \leq -0.3n\right\} \leq \exp(-0.18n).$$

This means for any fixed $A \in \mathcal{W}$ and a sufficiently large n, there exists an absolute constant $c_2 \in (0, c_1)$ that

$$|\mathcal{W}_A| \leq \exp(-0.18n) 2^n \leq c_2^n.$$

We let $\mathcal{M} \subset \mathcal{W}$ and any $A, B \in \mathcal{M}$ satisfy that $|A \triangle B| > 0.2n$. Then the $0.2n$-packing number of 2^V with respect to the Hamming distance satisfies

$$M(0.2n, 2^V, d_{\mathrm{Hamm}}) \geq |\mathcal{M}| \geq \frac{|\mathcal{W}|}{|\mathcal{W}_A| + 1} \geq c_3^n, \tag{14}$$

where $c_3 \in (1, 2)$ is an absolute constant.

Construction of nodes. We now construct a class of distributions $\{P_A, A \in \mathcal{M}\}$. Each P_A is the joint distribution of independent random variables

$$Y_i \sim \begin{cases} \mathcal{N}(\kappa, \sigma^2), & i \in A, \\ \mathcal{N}(0, \sigma^2), & i \notin A. \end{cases}$$

We further associate P_A with a connected graph $G_A = (V, E_A)$.

Fano's method. To use Fano's method, we adopt the version in Yu (1997). Note that for any $A, B \in \mathcal{M}$, we have that

$$d_{\mathrm{H}}(A, B) \geq 0.2n,$$

and

$$\mathrm{KL}(P_A, P_B) = 0.2n\kappa^2/\sigma^2.$$

Then provided that

$$\Delta\kappa^2\sigma^{-2} \lesssim n,$$

it holds that

$$\begin{aligned}\inf_{\widehat{\mathcal{S}}} \sup_{P\in\mathcal{S}} \mathbb{E}_P\{d_{\mathrm{H}}(\widehat{\mathcal{S}}, \mathcal{S})\} \geq & \frac{0.2n}{2}\left(1 - \frac{0.2n\kappa^2/\sigma^2 + \log(2)}{\log(M(0.2n, 2^V, d_{\mathrm{Hamm}}))}\right) \\ \geq & 0.1n\left(1 - \frac{0.2n\kappa^2/\sigma^2 + \log(2)}{n\log(c_3)}\right) \geq cn.\end{aligned}$$

□

Proof of Lemma 4. It follows from Kirchhoff's maximum tree theorem (e.g. Theorem 1 in Chaiken and Kleitman, 1978) that the number of spanning trees of G equals $\det(L)/n$, where L is the Laplacian of G. Since G is a complete graph, the eigenvalues of G's Laplacian are n^{n-1} and 0. Hence, Kirchhoff's maximum-tree theorem implies that there are n^{n-2} spanning trees. Due to the definition of spanning trees, each spanning tree consists of $n-1$ edges. Then there are in total $(n-1)n^{n-2}$ edges contained in all the spanning trees.

On the other hand, there are $n(n-1)/2$ edges in the complete graph G. Since this is a complete graph, all edges are equivalent. This implies that each edge (i, j) appears in

$$\frac{(n-1)n^{n-2}}{n(n-1)/2} = 2n^{n-3}$$

spanning trees.

Due to the definition of the effective resistance weights, we have that

$$\partial_r(\mathcal{S}) = \frac{2n^{n-3}}{n^{n-2}} n_1 n_2 = 2n_1 n_2/n,$$

which concludes the proof. □

B PROOFS OF RESULTS IN SECTION 3

Proof of Theorem 5. Without loss of generality, in this proof, we assume $f_1^* = 0$, $f_2^* = \kappa$ and $\Delta = |S_1^*| \leq |S_2^*|$. For any vector $v \in \mathbb{R}^n$ and any subset $A \subset V$, define

$$\overline{v}_A = |A|^{-1} \sum_{i\in A} v_i.$$

For any partition $\mathcal{C}$, it is associated with a $|\mathcal{C}|$-dimensional subspace $K \subset \mathbb{R}^n$, such that $v \in K$ if and only if v takes a constant value on each element of $\mathcal{C}$. We denote the orthogonal projection onto K by $P^{\mathcal{C}} : \mathbb{R}^n \to K$. For the estimator $\{\widehat{S}_1, \widehat{S}_2\}$, we let $A_{kl} = \widehat{S}_k \cap S_l^*$, $k, l = 1, 2$. Based on this notation, for the uniqueness of the definition, we let

$$\frac{|A_{11}|}{|\widehat{S}_1|} \geq \frac{|A_{21}|}{|\widehat{S}_2|}. \tag{15}$$

Step 1. Let

$$\mathcal{E} = \left\{\|P^{\mathcal{C}}\varepsilon\|^2 \leq C|\partial_w(\mathcal{C})|\log\{w(E)\}, \forall \text{ partition } \mathcal{C} \text{ of } G\right\}. \tag{16}$$

It follows from Lemma B.2 in Fan and Guan (2018) that

$$\mathbb{P}\{\mathcal{E}\} \geq 1 - w(E)^{-c},$$

where $C, c > 0$ are absolute constants. The rest of the proof is conducted on the event $\mathcal{E}$.

Step 2. Let $\{\widehat{S}_1, \widehat{S}_2\}$ be the output of Algorithm 1. It must hold that

$$\begin{aligned}
&\sum_{i\in\widehat{S}_1}(Y_i - \overline{Y}^{\delta}_{\widehat{S}_1})^2 + \sum_{i\in\widehat{S}_2}(Y_i - \overline{Y})^2 + 2\lambda|\partial_w(\widehat{S}_1, \widehat{S}_2)| \\
\leq &\sum_{i\in S_1^*}(Y_i - \overline{Y}^{\delta}_1)^2 + \sum_{i\in S_2^*}(Y_i - \overline{Y})^2 + 2\lambda|\partial_w(S_1^*, S_2^*)| + 4\tau,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sum_{i\in\widehat{S}_1}(Y_i - \overline{Y}_{\widehat{S}_1})^2 + \sum_{i\in\widehat{S}_2}(Y_i - \overline{Y})^2 + 2\lambda|\partial_w(\widehat{S}_1, \widehat{S}_2)| \\
\leq &\sum_{i\in S_1^*}(Y_i - \overline{Y}_1)^2 + \sum_{i\in S_2^*}(Y_i - \overline{Y})^2 + 2\lambda|\partial_w(S_1^*, S_2^*)| + 2n\delta^2 + 4\tau. \qquad (17)
\end{aligned}$$

Let

$$(\mu_1)_i = \begin{cases} \overline{Y}_{\widehat{S}_1}, & i \in \widehat{S}_1, \\ \overline{Y}, & i \in \widehat{S}_2, \end{cases} \quad (f_1)_i = \begin{cases} \frac{|A_{12}|}{|\widehat{S}_1|}\kappa, & i \in \widehat{S}_1, \\ \frac{|S_2^*|}{n}\kappa, & i \in \widehat{S}_2, \end{cases} \quad (\varepsilon_1)_i = \begin{cases} \frac{1}{|\widehat{S}_1|}\sum_{i\in\widehat{S}_1}\varepsilon_i, & i \in \widehat{S}_1, \\ \frac{1}{n}\sum_{i\in V}\varepsilon_i, & i \in \widehat{S}_2, \end{cases}$$

$$(\mu_2)_i = \begin{cases} \overline{Y}_{S_1^*}, & i \in S_1^*, \\ \overline{Y}, & i \in S_2^*, \end{cases} \quad (f_2)_i = \begin{cases} 0, & i \in S_1^*, \\ \frac{|S_2^*|}{n}\kappa, & i \in S_2^*, \end{cases} \quad \varepsilon_2 = \begin{cases} \frac{1}{|S_1^*|}\sum_{i\in S_1^*}\varepsilon_i, & i \in S_1^*, \\ \frac{1}{n}\sum_{i\in V}\varepsilon_i, & i \in S_2^*, \end{cases}$$

$$(\mu_0)_i = \begin{cases} 0, & i \in S_1^*, \\ \kappa, & i \in S_2^* \end{cases} \quad \text{and} \quad Y = \mu_0 + \varepsilon,$$

where, we have assumed that $\widehat{S}_1 \neq \emptyset$. We will come back to prove this claim in **Step 5**.

Based on the notation above, we note that

$$(f_1 + f_2 - 2\mu_0)_i = \begin{cases} \frac{|A_{12}|}{|\widehat{S}_1|}\kappa, & i \in A_{11}, \\ -\left(\frac{|A_{11}|}{|\widehat{S}_1|} + \frac{|S_1^*|}{n}\right)\kappa, & i \in A_{12}, \\ \frac{|S_2^*|}{n}\kappa, & i \in A_{21}, \\ -2\frac{|S_1^*|}{n}\kappa, & i \in A_{22}, \end{cases} \quad (f_1 - f_2)_i = \begin{cases} \frac{|A_{12}|}{|\widehat{S}_1|}\kappa, & i \in A_{11}, \\ \left(\frac{|A_{12}|}{|\widehat{S}_1|} - \frac{|S_2^*|}{n}\right)\kappa, & i \in A_{12}, \\ \frac{|S_2^*|}{n}\kappa, & i \in A_{21}, \\ 0, & i \in A_{22}. \end{cases}$$

and

$$(\varepsilon_1 - \varepsilon_2)_i = \begin{cases} \frac{1}{|\widehat{S}_1|}\sum_{j\in\widehat{S}_1}\varepsilon_j - \frac{1}{|S_1^*|}\sum_{j\in S_1^*}\varepsilon_j, & i \in A_{11}, \\ \frac{|\widehat{S}_2|}{n|\widehat{S}_1|}\sum_{j\in\widehat{S}_1}\varepsilon_j - \frac{1}{n}\sum_{j\in\widehat{S}_2}\varepsilon_j, & i \in A_{12}, \\ \frac{1}{n}\sum_{j\in S_2^*}\varepsilon_j - \frac{|S_2^*|}{n|S_1^*|}\sum_{j\in S_1^*}\varepsilon_j, & i \in A_{21}, \\ 0, & i \in A_{22}. \end{cases}$$

With the above notation, we have that $\mu_1 = f_1 + \varepsilon_1$, $\mu_2 = f_2 + \varepsilon_2$,

$$\sum_{i\in\widehat{S}_1}(Y_i - \overline{Y}_{\widehat{S}_1})^2 + \sum_{i\in\widehat{S}_2}(Y_i - \overline{Y})^2 = \|Y - \mu_1\|^2 \quad \text{and} \quad \sum_{i\in S_1^*}(Y_i - \overline{Y}_1)^2 + \sum_{i\in S_2^*}(Y_i - \overline{Y})^2 = \|Y - \mu_2\|^2.$$

Step 3. In this step, we are to exploit (15), which directly implies that

$$|A_{11}||A_{22}| > |A_{12}||A_{21}|.$$

This means

$$|A_{11}||A_{22}| + |A_{11}||A_{12}| + |A_{12}|^2 + |A_{12}||A_{22}| > |A_{12}||A_{21}| + |A_{11}||A_{12}| + |A_{12}|^2 + |A_{12}||A_{22}|,$$

which is equivalent to

$$(|A_{11}| + |A_{12}|)(|A_{12}| + |A_{22}|) > |A_{12}|(|A_{11}| + |A_{12}| + |A_{21}| + |A_{22}|).$$

Simplifying the above leads to

$$\frac{|S_2^*|}{n} > \frac{|A_{12}|}{|\widehat{S}_1|}. \tag{18}$$

Another consequence of (15) is that

$$\frac{|A_{11}|}{|\widehat{S}_1|} \geq \frac{|S_1^*|}{n}. \tag{19}$$

If (19) does not hold, then we have that

$$\frac{|S_1^*|}{n} = \frac{|A_{11}| + |A_{21}|}{|\widehat{S}_1| + |\widehat{S}_2|} \leq \frac{|A_{11}|}{|\widehat{S}_1|} < \frac{|S_1^*|}{n},$$

which is a contradiction

Step 4. In this step, we are to lower bound

$$Q = \|Y - \mu_1\|^2 - \|Y - \mu_2\|^2.$$

Note that, with an absolute constant $c \in (0, 1)$, it holds that

$$\begin{aligned}
Q &= \|\mu_0 - \mu_1 + \varepsilon\|^2 - \|\mu_0 - \mu_2 + \varepsilon\|^2 = \|\mu_0 - \mu_1\|^2 - \|\mu_0 - \mu_2\|^2 + 2(\mu_2 - \mu_1)^\top \varepsilon \\
&\geq \|\mu_0 - f_1 - \varepsilon_1\|^2 - \|\mu_0 - f_2 - \varepsilon_2\|^2 - c\|f_1 - f_2 + \varepsilon_1 - \varepsilon_2\|^2 - \frac{1}{c}\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*}\varepsilon\|^2 \\
&= (f_1 + f_2 - 2\mu_0)^\top (f_1 - f_2) + \langle f_1 + f_2 - 2\mu_0, \varepsilon_1 - \varepsilon_2\rangle + \langle f_1 - f_2, \varepsilon_1 + \varepsilon_2\rangle + \langle \varepsilon_1 \\
&\qquad + \varepsilon_2, \varepsilon_1 - \varepsilon_2\rangle - c\|f_1 - f_2 + \varepsilon_1 - \varepsilon_2\|^2 - \frac{1}{c}\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*}\varepsilon\|^2 \\
&\geq (f_1 + f_2 - 2\mu_0)^\top (f_1 - f_2) + \langle f_1 + f_2 - 2\mu_0, \varepsilon_1 - \varepsilon_2\rangle - c_1\|f_1 - f_2\|^2 - \frac{1}{4c_1}\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*}(\varepsilon_1 + \varepsilon_2)\|^2 \\
&\qquad - c_1\|\varepsilon_1 - \varepsilon_2\|^2 - \frac{1}{4c_1}\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*}(\varepsilon_1 + \varepsilon_2)\|^2 - 2c\|f_1 - f_2\|^2 - 2c\|\varepsilon_1 - \varepsilon_2\|^2 - \frac{1}{c}\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*}\varepsilon\|^2 \\
&= (f_1 + f_2 - 2\mu_0)^\top (f_1 - f_2) + \langle f_1 + f_2 - 2\mu_0, \varepsilon_1 - \varepsilon_2\rangle - (c_1 + 2c)\|f_1 - f_2\|^2 \\
&\qquad - \left(\frac{2}{c_1} + \frac{1}{c}\right)\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*}\varepsilon\|^2 - (c_1 + 2c)\|\varepsilon_1 - \varepsilon_2\|^2,
\end{aligned} \tag{20}$$

where facts that $\mu_1 = f_1 + \varepsilon_1$, $\mu_2 = f_2 + \varepsilon_2$ and $2ab \geq ca^2 + b^2/c$, with $c > 0$, are repeatedly used above.

For the first term in (20), we have that

$$\begin{aligned}
&(f_1 + f_2 - 2\mu_0)^\top (f_1 - f_2) \\
=&\frac{|A_{11}||A_{12}|^2}{|\widehat{S}_1|^2}\kappa^2 - \left(\frac{|A_{11}||A_{12}|^2}{|\widehat{S}_1|^2}\kappa^2 - \frac{|A_{11}||A_{12}||S_2^*|}{|\widehat{S}_1|n}\kappa^2 + \frac{|A_{12}|^2|S_1^*|}{|\widehat{S}_1|n}\kappa^2 - \frac{|A_{12}||S_1^*||S_2^*|}{n^2}\kappa^2\right) \qquad (21) \\
&\qquad + \frac{|A_{21}||S_2^*|^2}{n^2}\kappa^2 \qquad (22) \\
=&\frac{|A_{11}||A_{12}||S_2^*|}{|\widehat{S}_1|n}\kappa^2 - \frac{|A_{12}|^2|S_1^*|}{|\widehat{S}_1|n}\kappa^2 + \frac{|A_{12}||S_1^*||S_2^*|}{n^2}\kappa^2 + \frac{|A_{21}||S_2^*|^2}{n^2}\kappa^2 \\
\geq&\frac{|A_{11}|}{2|\widehat{S}_1|}|A_{12}|\kappa^2 + \frac{|A_{12}||S_1^*|}{n}\left(\frac{|S_2^*|}{n} - \frac{|A_{12}|}{|\widehat{S}_1|}\right)\kappa^2 + \frac{|A_{21}|}{4}\kappa^2
\end{aligned}$$

$$\geq \left(\frac{|A_{11}|}{2|\widehat{S}_1|}|A_{12}| + \frac{1}{4}|A_{21}| \right) \kappa^2, \tag{23}$$

where the first term in (21) is from the products indexed in A_{11}, the four terms in the brackets in (21) are from the products indexed in A_{12}, the term in (22) is from the products indexed in A_{21}, the third inequality is due to the fact that $|S_2^*| \geq |S_1^*|$ and the final inequality is due to (18).

For the second term in (20), with the notation that

$$\overline{\varepsilon}_{kl} = |A_{kl}|^{-1} \sum_{i \in A_{kl}} \varepsilon_i, \quad k, l = 1, 2,$$

we have that

$$\begin{aligned}
&\langle f_1 + f_2 - 2\mu_0, \varepsilon_1 - \varepsilon_2 \rangle \\
=&\frac{|A_{11}||A_{12}|\kappa}{|\widehat{S}_1|} \left(\frac{1}{|\widehat{S}_1|} \sum_{j \in \widehat{S}_1} \varepsilon_i - \frac{1}{|S_1^*|} \sum_{j \in S_1^*} \varepsilon_i \right) - |A_{12}| \left(\frac{|A_{11}|}{|\widehat{S}_1|} + \frac{|S_1^*|}{n} \right) \kappa \left(\frac{|\widehat{S}_2|}{n|\widehat{S}_1|} \sum_{j \in \widehat{S}_1} \varepsilon_j - \frac{1}{n} \sum_{j \in \widehat{S}_2} \varepsilon_j \right) \\
&+ \frac{|S_2^*||A_{21}|}{n} \kappa \left(\frac{1}{n} \sum_{j \in S_2^*} \varepsilon_j - \frac{|S_2^*|}{n|S_1^*|} \sum_{j \in S_1^*} \varepsilon_j \right) \\
\leq& \left| \frac{|A_{11}|^2|A_{12}|^2}{|\widehat{S}_1|^2|S_1^*|} - \frac{|A_{11}|^2|A_{21}||A_{12}|}{|\widehat{S}_1|^2|S_1^*|} - \frac{|A_{12}||A_{11}|^2|\widehat{S}_2|}{n|\widehat{S}_1|^2} - \frac{|A_{12}||S_1^*||\widehat{S}_2||A_{11}|}{n^2|\widehat{S}_1|} - \frac{|S_2^*|^2|A_{21}||A_{11}|}{n^2|S_1^*|} \right| \kappa |\overline{\varepsilon}_{11}| \\
&+ \left| \frac{|A_{12}||A_{11}||A_{22}|}{n|\widehat{S}_1|} + \frac{|A_{12}||S_1^*||A_{22}|}{n^2} + \frac{|S_2^*||A_{21}||A_{22}|}{n^2} \right| \kappa |\overline{\varepsilon}_{22}| \\
&+ \left| \frac{|A_{11}||A_{12}|^2}{|\widehat{S}_1|^2} - \frac{|A_{12}|^2|A_{11}||\widehat{S}_2|}{n|\widehat{S}_1|^2} - \frac{|A_{12}|^2|S_1^*||\widehat{S}_2|}{n^2|\widehat{S}_1|} + \frac{|S_2^*||A_{21}||A_{12}|}{n^2} \right| \kappa |\overline{\varepsilon}_{12}| \\
&+ \left| -\frac{|A_{11}||A_{12}||A_{21}|}{|\widehat{S}_1||S_1^*|} + \frac{|A_{12}||A_{11}||A_{21}|}{n|\widehat{S}_1|} + \frac{|A_{12}||A_{21}||S_1^*|}{n^2} - \frac{|S_2^*|^2|A_{21}|^2}{n^2|S_1^*|} \right| \kappa |\overline{\varepsilon}_{21}| \\
\leq& cB\kappa^2 + C\|P^{\widehat{\mathcal{S}} \vee \mathcal{S}^*} \varepsilon\|^2,
\end{aligned} \tag{24}$$

where $c, C > 0$ are absolute constants and

$$\begin{aligned}
B =& \frac{|A_{11}|^4|A_{12}|^4}{|\widehat{S}_1|^4|S_1^*|^2|A_{11}|} + \frac{|A_{11}|^4|A_{21}|^2|A_{12}|^2}{|\widehat{S}_1|^4|S_1^*|^2|A_{11}|} + \frac{|A_{12}|^2|A_{11}|^4|\widehat{S}_2|^2}{n^2|\widehat{S}_1|^4|A_{11}|} + \frac{|A_{12}|^2|S_1^*|^2|\widehat{S}_2|^2|A_{11}|^2}{n^4|\widehat{S}_1|^2|A_{11}|} + \frac{|S_2^*|^4|A_{21}|^2|A_{11}|^2}{n^4|S_1^*|^2|A_{11}|} \\
&+ \frac{|A_{12}|^2|A_{11}|^2|A_{22}|^2}{n^2|\widehat{S}_1|^2|A_{22}|} + \frac{|A_{12}|^2|S_1^*|^2|A_{22}|^2}{n^4|A_{22}|} + \frac{|S_2^*|^2|A_{21}|^2|A_{22}|^2}{n^4|A_{22}|} \\
&+ \frac{|A_{11}|^2|A_{12}|^4}{|\widehat{S}_1|^4|A_{12}|} + \frac{|A_{12}|^4|A_{11}|^2|\widehat{S}_2|^2}{n^2|\widehat{S}_1|^4|A_{12}|} + \frac{|A_{12}|^4|S_1^*|^2|\widehat{S}_2|^2}{n^4|\widehat{S}_1|^2|A_{12}|} + \frac{|S_2^*|^2|A_{21}|^2|A_{12}|^2}{n^4|A_{12}|} \\
&+ \frac{|A_{11}|^2|A_{12}|^2|A_{21}|^2}{|\widehat{S}_1|^2|S_1^*|^2|A_{21}|} + \frac{|A_{12}|^2|A_{11}|^2|A_{21}|^2}{n^2|\widehat{S}_1|^2|A_{21}|} + \frac{|A_{12}|^2|A_{21}|^2|S_1^*|^2}{n^4|A_{21}|} + \frac{|S_2^*|^4|A_{21}|^4}{n^4|S_1^*|^2|A_{21}|} \\
\leq& \frac{|A_{11}|}{|\widehat{S}_1|}|A_{12}| + \frac{|A_{11}|^3}{|\widehat{S}_1|^3}|A_{12}| + \frac{|A_{11}|^3}{|\widehat{S}_1|^3}|A_{12}| + \frac{|A_{11}|}{|\widehat{S}_1|}|A_{12}| + |A_{21}| \\
&+ \frac{|A_{11}|^2}{|\widehat{S}_1|^2}|A_{12}| + \frac{|S_1^*|}{n}|A_{12}| + |A_{21}| \\
&+ \frac{|A_{11}|^3}{|\widehat{S}_1|^3}|A_{12}| + \frac{|A_{11}|^2}{|\widehat{S}_1|^2}|A_{12}| + \frac{|S_1^*|^2}{n^2}|A_{12}| + |A_{21}| \\
&+ |A_{21}| + \frac{|A_{11}|}{|\widehat{S}_1|}|A_{12}| + \frac{|S_1^*|^2}{n^2}|A_{12}| + |A_{21}|
\end{aligned}$$

$$\leq \frac{11|A_{11}|}{|\widehat{S}_1|}|A_{12}| + 5|A_{21}|, \tag{25}$$

where the last inequality is due to (19).

For the third term in (20), we have that

$$\begin{aligned}
\|f_1 - f_2\|^2 = &\frac{|A_{12}|^2|A_{11}|}{|\widehat{S}_1|^2}\kappa^2 + \left(\frac{|A_{12}|}{|\widehat{S}_1|} - \frac{|S_2^*|}{n}\right)^2 |A_{12}|\kappa^2 + \frac{|S_2^*|^2|A_{21}|}{n^2}\kappa^2 \\
\leq &\frac{|A_{11}|}{|\widehat{S}_1|}|A_{12}|\kappa^2 + \left(\frac{|S_1^*|}{n} - \frac{|A_{11}|}{|\widehat{S}_1|}\right)^2 |A_{12}|\kappa^2 + \frac{1}{4}|A_{21}|\kappa^2 \\
\leq &\frac{5|A_{11}|}{|\widehat{S}_1|}|A_{12}|\kappa^2 + \frac{1}{4}|A_{21}|\kappa^2,
\end{aligned} \tag{26}$$

where the last inequality is due to (19).

For the last term in (20), we have that

$$\begin{aligned}
&\|\varepsilon_1 - \varepsilon_2\|^2 \\
\leq &|A_{11}|\left(\frac{|A_{12}||A_{11}| - |A_{21}||A_{11}|}{|\widehat{S}_1||S_1^*|}\overline{\varepsilon}_{11} + \frac{|A_{12}|}{|\widehat{S}_1|}\overline{\varepsilon}_{12} - \frac{|A_{21}|}{|S_1^*|}\overline{\varepsilon}_{21}\right)^2 \\
&+ |A_{12}|\left(\frac{|\widehat{S}_2||A_{11}|}{n|\widehat{S}_1|}\overline{\varepsilon}_{11} + \frac{|\widehat{S}_2||A_{12}|}{n|\widehat{S}_1|}\overline{\varepsilon}_{12} - \frac{|A_{21}|}{n}\overline{\varepsilon}_{21} - \frac{|A_{22}|}{n}\overline{\varepsilon}_{22}\right)^2 \\
&+ |A_{21}|\left(-\frac{|S_2^*||A_{11}|}{n|S_1^*|}\overline{\varepsilon}_{11} + \frac{|A_{12}|}{n}\overline{\varepsilon}_{12} - \frac{|S_2^*||A_{21}|}{n|S_1^*|}\overline{\varepsilon}_{21} + \frac{|A_{22}|}{n}\overline{\varepsilon}_{22}\right)^2 \\
\leq &4\left(\frac{|A_{11}|^3|A_{12}|^2}{|\widehat{S}_1|^2|S_1^*|^2} + \frac{|A_{11}|^3|A_{21}|^2}{|\widehat{S}_1|^2|S_1^*|^2} + \frac{|A_{12}||\widehat{S}_2|^2|A_{11}|^2}{n^2|\widehat{S}_1|^2} + \frac{|A_{21}||S_2^*|^2|A_{11}|^2}{n^2|S_1^*|^2}\right)\overline{\varepsilon}_{11}^2 \\
&+ 4\left(\frac{|A_{11}||A_{12}|^2}{|\widehat{S}_1|^2} + \frac{|A_{12}||\widehat{S}_2|^2|A_{12}|^2}{n^2|\widehat{S}_1|^2} + \frac{|A_{21}||A_{12}|^2}{n^2}\right)\overline{\varepsilon}_{12}^2 \\
&+ 4\left(\frac{|A_{11}||A_{21}|^2}{|S_1^*|^2} + \frac{|A_{12}||A_{21}|^2}{n^2} + \frac{|A_{21}||S_2^*|^2|A_{21}|^2}{n^2|S_1^*|^2}\right)\overline{\varepsilon}_{21}^2 + 4\left(\frac{|A_{12}||A_{22}|^2}{n^2} + \frac{|A_{21}||A_{22}|^2}{n^2}\right)\overline{\varepsilon}_{22}^2 \\
\leq &16|A_{11}|\overline{\varepsilon}_{11}^2 + 12|A_{12}|\overline{\varepsilon}_{12}^2 + 12|A_{21}|\overline{\varepsilon}_{21}^2 + 8|A_{22}|\overline{\varepsilon}_{22}^2 \\
\leq &16\|P^{\widehat{\mathcal{S}}\vee\mathcal{S}^*}\varepsilon\|^2.
\end{aligned} \tag{27}$$

Combining (20), (23), (24), (25), (26) and (27), we have that, with absolute constants $c_*, C_1 > 0$,

$$Q \geq c_*\left(\frac{|A_{11}||A_{12}|}{|\widehat{S}_1|} + |A_{21}|\right)\kappa^2 - C_1\sigma^2|\partial_w(S_1^*, S_2^*)| - C_1\sigma^2|\partial_w(\widehat{S}_1, \widehat{S}_2)|, \tag{28}$$

where the second inequality is due to the choices of the constants.

Combining (17) and (28), with a sufficiently large $C_\lambda > 0$, we have that

$$C\sigma^2|\partial_w(\widehat{S}_1, \widehat{S}_2)| + c_*\left(\frac{|A_{11}||A_{12}|}{|\widehat{S}_1|} + |A_{21}|\right)\kappa^2 \leq 2\lambda|\partial_w(S_1^*, S_2^*)| + 2n\delta^2 + \tau,$$

which implies (8).

Step 5. To conclude the proof, the only remaining task is to show that $\widehat{S}_1 \neq \emptyset$. We prove by contradiction and therefore have that

$$\Delta(\overline{Y} - \overline{Y}_1)^2 \leq 2\lambda|\partial_w(\mathcal{S}^*)| + 2n\delta^2 + \tau.$$

On the event $\mathcal{E}$, it implies that

$$\Delta\kappa^2 \leq C\lambda|\partial_w(\mathcal{S})| + 2n\delta^2 + \tau,$$

which contradicts with Assumption 2. We thus conclude the proof. □

Proof of Proposition 6. Note that due to the design of the algorithms, the output of Algorithm 1 has fewer pieces than the output of Algorithm 1 in Fan and Guan (2018). Therefore, the claim holds directly follows from Theorem 3.5 in Fan and Guan (2018). □