Inference on random networks

Yann Issartel

université
PARIS-SACLAY

Inférence sur des graphes aléatoires

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale de Mathématique Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques Appliquées
Unité de recherche : Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue en visioconférence totale, le 24/11/2020, par

Yann ISSARTEL

Composition du jury :

Elisabeth Gassiat Professeure, Université Paris-Saclay	Présidente
Jing Lei Maître de conférences, Carnegie Mellon University	Rapporteur
Laurent Massoulié Chercheur, Directeur de Microsoft Research-INRIA	Rapporteur & examinateur
Stéphane Boucheron Professeur, Université de Paris	Examinateur
Cristina Butucea Professeure, ENSAE-CREST	Examinatrice
Emilie Kaufmann Chercheuse, CNRS	Examinatrice
Christophe Giraud Professeur, Université Paris-Saclay	Directeur de thèse
Nicolas Verzelen Chercheur, INRAE	Co-directeur de thèse

Thèse de doctorat

NNT: 2020UPASM019


université
PARIS-SACLAY
FACULTÉ
DES SCIENCES
D'ORSAY


MATHEMATIQUES
ORSAY


Fondation mathématique
FMJH
Jacques Hadamard


ÉCOLE
POLYTECHNIQUE


IP PARIS

Remerciements

Évidemment, les remerciements les plus sincères vont à Christophe et Nicolas. Incroyables pédagogues, ils m'ont fait découvrir l'apprentissage statistique en thèse, malgré mes études d'algèbre (ne sont-ils pas courageux?). Bienveillants aussi, ils ont consacré beaucoup de temps et d'énergie à ma formation d'apprenti chercheur. Surtout généreux dans les idées, ils m'ont donné des questions riches et originales qui ont guidé mes premiers travaux. Encore aujourd'hui, ces questions m'accompagnent.

I am very grateful to Jing Lei and Laurent Massoulié for reviewing my thesis in great detail. Their comments raised several future directions of research, which I hope will be fruitful. Je tiens à remercier chaque membre de mon jury, Stéphane Boucheron, Cristina Butucea, Elisabeth Gassiat et Emilie Kaufmann, qui ont très gentiment accepté de participer.

Cette thèse doit aussi beaucoup à mes co-auteurs, Luc Lehéricy et Matthieu Lerasle, ainsi qu'à de nombreux chercheurs de l'IMO, Ernesto Araya Valdivia, Yohann De Castro, Zacharie Naulet, Suzanne Varet et bien d'autres, pour de nombreuses discussions intéressantes.

C'est aussi l'occasion de remercier l'École Polytechnique et l'Université Paris-Saclay pour la grande liberté accordée tout au long de mes études. J'ai eu la chance d'avoir des enseignants de qualité exceptionnelle, ainsi qu'une administration remarquablement efficace. Je remercie notamment Vanessa Delaisse, Clotile D' Epenoux, Stéphane Nonnenmacher, Frédéric Paulin, et tous ceux qui m'ont aidé avec le sourire dans les choses administratives.

Ce fut un plaisir de venir travailler à Orsay. Un immense merci aux amis qui ont égayé l'ambiance du bureau, en particulier les inconditionnels de la "pause cafèt chez Jean-Marie". Merci aussi à tous mes proches qui n'ont aucun lien scientifique avec cette thèse, excepté ceux qui se demandent quand je chercherai enfin un "vrai boulot" ;)

Extended abstract

Many real-life data arise in the form of pairwise measurements $(A_{ij})_{1\leq i<j\leq n}$, where $A_{ij} \in \mathbb{R}$ is the outcome of some interaction between i and j. For instance, in online video games, A_{ij} may be the number of parties between the i^{th} and j^{th} players. In social networks, $A_{ij} \in \{0, 1\}$ encodes the presence/absence of a friendship link between two users. A major challenge is to extract useful information from these data sets which are not composed of i.i.d. observations but observations with complex dependencies. In that respect, latent space models are widely used to represent such relational information among interacting agents. They assume that the expected interaction between individuals depends on their positions in a latent (i.e. unobserved) space. Formally, there exist unobserved positions $x_1, \dots, x_n$ in a latent space X, and an affinity function $f : X \times X \mapsto \mathbb{R}$ such that the expected interaction between individuals i and j is the affinity between their attributes x_i and x_j, i.e. $\mathbb{E}[A_{ij}] = f(x_i, x_j)$. In latent space models, the latent positions associated with individuals can be random; for example, the $x_1, \dots, x_n$ may be i.i.d. random variables with distribution μ on X. This latent space formulation encompasses many models such as stochastic block models (SBM), random geometric graphs or graphon models. The triplet of parameters (X, μ, f) is often referred to as graphon in the literature.

In this thesis we study three inference problems in latent space models, from a non-asymptotic viewpoint. The first question concerns the identifiability issue inherent in these models. Neither the latent points $x_1, \dots, x_n$, nor the latent space X, nor the function f are identifiable from the data (A_{ij}). In other words, even if two latent structures are different, they can lead to a same data distribution. Due to this identifiability issue, latent space models may be difficult to interpret in practice. In order to make them more operational, one can define an identifiable functional which represents an interpretable property of the data. Following this direction in the general model of graphon, we define an identifiable notion of complexity for networks: Given a graphon (X, μ, f), we endow the latent space X with the so-called neighborhood distance r_f that measures the propensity $r_f(x_i, x_j)$ of two agents i and j to be connected with similar individuals; The complexity index is then based on the covering number and the Minkowski dimension of (a 'purified' version of) the metric space (X, r_f). In order to illustrate that this index is sound, we give several examples in classic models of random graphs. We also consider the problem of inferring this complexity from a single graph observation $(A_{ij})_{1\leq i<j\leq n}$. Optimal minimax estimators are proved for the neighbor-hood distance and the complexity index.

The second question is about optimal strategies in sequential learning. We introduce the pair-matching problem which appears in many applications where one wants to discover good matches between pairs of individuals. Formally, the set of individuals is represented by the nodes of a graph where the edges, unobserved

at first, represent the good matches. Then, the algorithm sequentially queries pairs of nodes and observes the presence/absence of edges. Its goal is to discover as many edges as possible with a fixed budget of queries. Pair-matching is a particular instance of multi-armed bandit problems in which the arms are pairs of individuals and the rewards are edges linking these pairs. This bandit problem is non-standard though, as each arm can only be played once. Given this last constraint, sub-linear regret can be expected only if the graph presents some underlying structure ('sub-linear regret' means that one perform substantially better than the trivial strategy sampling all pairs at random). We show that sub-linear regret is achievable in the case where the graph is generated according to a SBM with two communities. Optimal regret bounds are computed for this pair-matching problem.

The pair-matching problem is then investigated in more complex models such as SBM with K communities and random geometric graphs. Optimal strategies in these settings focus on a few nodes from a small region of the latent space, by making a local exploration of the graph. In online video games, for example, such a local strategy is bad since it matches a few players many times while keeping the others waiting. To avoid such undesirable features, the pair-matching problem is also considered in the case where strategies are constrained to explore many communities in SBM or a large portion of the latent space in geometric graphs. We study how optimal regrets depend on this constraint. Some lower-bounds on regrets are proved, and some detailed arguments are presented to explain how it should be possible to derive matching upper-bounds.

The final question is on optimal embedding of interaction data, with applications to random geometric graphs, pair-matching problem and statistical seriation. Given the observations $(A_{ij})_{1\leq i<j\leq n}$ that are modelled according to a latent space model, we are interested in building an estimator $(\hat{x}_1, \ldots, \hat{x}_n)$ of the latent positions $(x_1, \ldots, x_n)$. We restrict our attention to a one-dimensional space X endowed with a metric d. The affinity $f(x_i, x_j)$ is typically assumed to decrease as the metric distance $d(x_i, x_j)$ increases. In particular, close points x_i and x_j share a high affinity whereas distant points share a small affinity. Motivated by a conjecture in the pair-matching problem, the performance of the estimator $(\hat{x}_1, \ldots, \hat{x}_n)$ is assessed by the uniform error: $\max_{i=1,\ldots,n} d(x_i, \hat{x}_i)$. We establish the minimax rate for this localization problem, in the case where f is unknown and belongs to a class of (almost) bi-Lipschitz functions. Our estimation procedure takes a two-step approach, following a 'global to local' scheme. It first computes a good enough estimator of all the positions and then builds on this preliminary estimator to locally improve each position estimate. Since this procedure exhibits exponential-time complexity, we propose a computationally efficient alternative to it, which achieves the minimax rate in the particular case of random geometric graphs.

Contents

Chapter 0

Introduction

Dans cette thèse, nous étudions trois questions d'inférence sur les réseaux aléatoires, d'un point de vue non-asymptotique : la construction de quantité identifiable et interprétable pour des modèles à espace latent (chapitre 1), l'échantillonnage séquentiel et ses stratégies optimales (chapitres 2 et 3), et le plongement de données dans un espace métrique (chapitre 4).

Cette partie introductive est une présentation informelle du cadre d'étude ainsi que des résultats obtenus au cours de cette thèse. Une version détaillée de ces résultats peut être trouvée dans les chapitres 1-4.

Contents

0.1 Contexte actuel: la statistique mathématique à l'ère du 'Big data'

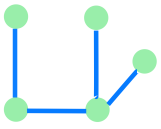
Depuis que la technologie permet de rassembler des quantités massives de données, de nombreux secteurs (finance, biotech, marketing, etc) ont choisi d'embrasser pleinement cette opportunité. Certaines entreprises n'hésitent pas à miser sur la valeur mercantile de la 'data', en investissant des sommes considérables ('data industry'). Ces grands volumes de données permettraient de dégager des connaissances nouvelles et utiles, auxquelles ne peuvent parvenir les moyens d'étude plus traditionnels. Pour réussir ce pari du '+ de données = + de valeurs', des efforts importants sont entrepris afin d'exploiter ces dépôts d'information.

Le besoin d'analyser ces données a engendré un domaine de recherche à part entière, l'analyse de données, qui s'est rapidement développé ces dernières décennies pour répondre à des questions fondamentales telles que la classification, la régression, le clustering, etc. Ce besoin est reconnu dans le monde académique, comme l'atteste la multiplication des conférences, des revues, ou encore des Masters 'Big data' et 'Machine Learning'. Ce développement se poursuit au rythme des technologies de l'information, qui apportent des données plus variées et complexes qu'avant.

Parmi les nombreuses branches de l'analyse de données, la statistique mathématique se démarque par son formalisme: le cadre d'étude est défini par un modèle mathématique, et la procédure d'estimation est un algorithme précis, dont la performance théorique est prouvée dans un théorème. Cette validation par le 'théorème-preuve' n'est pas la plus répandue chez les chercheurs en analyse de données (loin derrière la simulation par ordinateur par exemple). Il y a en effet un véritable fossé aujourd'hui entre la théorie mathématique et les dernières avancées en analyse de données, ce qui d'une part offre un nouveau terrain de jeux aux mathématiciens, et d'autre part la perspective d'apporter des fondations théoriques à l'analyse de données, afin de mieux appréhender certaines questions abstraites.

0.2 Objectif général de thèse

On s'intéresse aux interactions ou connexions formées parmi un groupes d'entités; cet ensemble d'observations est appelé réseau dans la suite. L'étude statistique de telles données est délicate à cause des dépendances complexes entre observations, qui ne peuvent être vues comme de simples variables isolées et indépendantes. Une représentation standard de ces données est la forme matricielle $(A_{ij})_{1\leq i,j\leq n}$, où n est le nombre d'entités étudiées, et $A_{ij} \in \mathbb{R}$ est le résultat d'une interaction entre les entités i et j. Par exemple, dans un championnat sportif, A_{ij} peut être le nombre de matchs entre le $i^{\text{ème}}$ et $j^{\text{ème}}$ joueur.
On mettra l'emphase sur le cas particulier des réseaux, vus comme des graphes. Dans ce cas, chaque entité du réseau est représentée par un noeud du graphe, et la présence d'une interaction entre deux entités par une arête entre les noeuds correspondants. Ces informations sont commodément représentées par la matrice d'adjacence $(A_{ij})_{1\leq i,j\leq n}$ du graphe, où $A_{ij} \in \{0,1\}$ indique l'absence/présence d'une arête entre i et j. On suivra une démarche statistique, en supposant que les données $(A_{ij})_{1\leq i,j\leq n}$ sont simulées à partir de modèles probabilistes.



réseau=graphe non-orienté

Choisir des hypothèses de modélisation, c'est choisir un camp. Chez le praticien, un bon modèle est suffisamment riche pour rendre compte des spécificités des données qui ont été collectées sur le terrain. Chez le théoricien, un bon modèle est abstrait et surtout propice à l'analyse mathématique. Face à ce clivage entre la pratique et la théorie, un objectif important est d'amener l'analyse statistique sur des modèles plus généraux et proches de la pratique. Dans cette thèse, nous utiliserons des modèles non-paramétriques pour éviter une modélisation naïve des données, bien que des modèles paramétriques seront aussi utilisés pour amorcer l'étude de nouvelles questions. Nous travaillerons sans supposer une taille d'échantillon très grande (approche non-asymptotique), cette hypothèse étant irréaliste dans de nombreuses situations pratiques.

Les qualités que nous chercherons dans un estimateur sont les suivantes:

- Il est simple à comprendre et implémenter.
- Les garanties sur sa performance sont valables sous peu d'hypothèses.
- Sa vitesse de convergence est la "meilleure" possible, au sens minimax (ce terme est défini plus bas).
- On comprend comment sa performance dépend des paramètres du problème.
- Enfin, sa complexité en temps est faible.

0.3 Cadre d'étude: Modèles à espace latent pour les réseaux

Motivée par des applications en sciences sociales, génétiques ou marketing, l'analyse de réseaux est un domaine en plein essor. Une approche classique consiste à ajuster le réseau observé $(A_{ij})_{1\leq i,j\leq n}$ à un modèle de graphe aléatoire, ce afin de mettre en évidence des aspects importants de la structure des données. Parmi les modèles envisageables, ceux à espace latent sont largement utilisés depuis plus d'une vingtaine d'années [Hoff et al., 2002]. Ils offrent une description précise de la topologie du réseau, tout en restant facile à interpréter. Les deux sous-sections suivantes sont dédiées à l'introduction de quelques uns de ces modèles.

0.3.1 La modélisation par espace latent

Les premiers graphes aléatoires datent des années 1950, notamment avec le célèbre modèle d'Erdös-Rényi. Dans ce modèle, chaque arête a une même probabilité p d'être présente, indépendemment des autres. De façon surprenante, la simplicité du modèle va de paire avec une certaine richesse mathématique. Il est en effet possible de décrire la structure du graphe en fonction des valeurs de p [Bollobás, 1998]. Du point de vue du statisticien, cette probabilité de connexion constante est une hypothèse trop forte, qui ne permet pas de modéliser des données de la vie réelle.

Prenons un réseau social, par exemple, qui est donné par un graphe, où les noeuds représentent des individus et les arêtes des relations entre individus. La probabilité d'une relation entre deux personnes varie en fonction de leurs caractéristiques individuelles (lieu de vie, loisirs, travail, etc). Ces informations sont souvent privées et indisponibles pour le statisticien. Afin de modéliser ces réseaux au mieux, on utilise la modélisation par espace

latent (où 'latent' signifie 'non-observé'). Soit un espace latent X et une fonction $f : X \times X \to [0,1]$. A chaque noeud i est associé un point latent $x_i \in X$, et la probabilité d'avoir une arête entre i et j vaut $f(x_i, x_j)$. Autrement dit, chaque individu i possède un attribut individuel x_i, qui est un point de 'l'espace social' X.

Dans cette modélisation, il existe donc une réalité spatiale sous-jacente, qui permet d'expliquer de façon simple les dépendances complexes entre les observations du réseau. Cette approche recouvre plusieurs modèles de graphes aléatoires bien connus, tels que les modèles à blocs stochastiques, de graphon et de graphe aléatoire géométrique.

0.3.2 Des modèles populaires de graphes aléatoires à espace latent

Dans les modèles de graphes aléatoires ci-dessous, les positions x_i sont elles aussi aléatoires. Parmi les modèles paramétriques, le modèle à blocs stochastiques (SBM) est particulièrement populaire [Holland et al., 1983]. Pour tout entier l, la notation $[l]$ désigne l'ensemble discret $\{1, \dots, l\}$.

Definition 0.3.1 (SBM à K communautés/groupes)**.** *Soient un ensemble fini $X = [K]$ muni d'un vecteur de probabilité π de taille K, et d'une fonction $f : X \times X \to [0,1]$. A chaque noeud i du graphe est associée une variable aléatoire $x_i \in [K]$ tirée selon la loi π. Conditionnellement aux valeurs de $(x_i)_{i\in[n]}$, chaque arête est tirée indépendemment des autres, et la probabilité d'avoir une arête entre i et j vaut $f(x_i, x_j)$, c'est-à-dire, les observations A_{ij}, $i < j$, sont des variables aléatoires de Bernoulli indépendantes telles que $\mathbb{P}[A_{ij} = 1] = f(x_i, x_j)$.*

Dans un SBM, la probabilité de connexion entre deux noeuds i et j varie donc selon leur communauté d'appartenance x_i et x_j dans $[K]$. L'exemple le plus simple est sûrement le SBM assortatif à deux paramètres: il existe deux paramètres $0 \leq q < p \leq 1$ tels que $f(x_i, x_j) = p$ si i et j appartiennent à un même groupe, et $f(x_i, x_j) = q$ sinon. Dans cet exemple, les individus d'un même groupe ont une plus grande probabilité d'être connectés que des individus de groupes différents. Ce modèle permet ainsi de modéliser les réseaux avec une structure de communauté. Il est d'ailleurs très utilisé dans l'étude théorique du problème de clustering, qui consiste à estimer les communautés $(x_i)_{i\in[n]}$ à partir des observations $(A_{ij})_{1\leq i,j\leq n}$. Ce problème a fait l'objet d'une attention considérable [Abbe, 2017], dépassant les frontières entre statistiques, probabilités, informatique et physique.

Si l'utilisation du SBM permet de dégager des grands groupes de noeuds jouant un rôle similaire dans le réseau, ce modèle se révèle peu adapté à l'analyse de très grands graphes, pour lesquels des aspects plus fins de la structure pourraient être analysés (par exemple, les réseaux sociaux, ou d'interactions biologiques ou de co-citations académiques). Ces enjeux forts ont conduit à une vision non-paramétrique de l'analyse de réseaux [Bickel and Chen, 2009], notamment avec l'introduction de modèles plus généraux, tels que le W-graphe aléatoire (aussi appelé modèle de graphon) [Lovász, 2012, Diaconis and Janson, 2007].

Definition 0.3.2 (modèle de graphon)**.** *Un graphon est un triplet (X, μ, f), formé d'un espace X, d'une mesure de probabilité μ à support dans X, et d'une fonction mesurable $f : X \times X \to [0,1]$. A chaque noeud i est associé une étiquette $x_i \in X$ tirée selon la mesure*

μ. Conditionnellement aux $(x_i)_{i\in[n]}$, chaque arête est tirée indépendemment des autres, et la présence d'une arête entre i et j a pour probabilité $f(x_i, x_j)$.

On peut ainsi interpréter le modèle de graphon comme une généralisation à espace d'états continus des SBM. En théorie, l'utilisation de ce modèle non-paramétrique permettrait donc de capturer des propriétés plus fines du réseau. Ce point est soutenu par quelques propriétés mathématiques qui sont brièvement évoquées dans la remarque ci-dessous.

Un cas particulier du modèle de graphon est le modèle de graphe aléatoire géométrique [Penrose et al., 2003]. Dans ce dernier, l'espace latent est muni d'une métrique d, et la fonction f satisfait l'égalité suivante $f(x, x') = \widetilde{f}(d(x, x'))$ pour une fonction $\widetilde{f}$ d'une seule variable réelle. La probabilité d'une arête entre i et j dépend donc de la distance latente $d(x_i, x_j)$, et non plus des positions x_i et x_j. Cette modélisation est naturel pour les réseaux assortatifs (i.e. lorsque deux individus ont une grande affinité s'ils sont proches dans l'espace social, et vice versa).

Remarque: Sur le plan mathématique, il a été montré que le graphon satisfait deux propriétés universelles. La première stipule une étroite connexion avec la théorie des graphes échangeables, au sens suivant: toute distribution de graphe aléatoire, qui est invariante par permutation des noeuds, peut être exprimée à l'aide du modèle de graphons [Diaconis and Janson, 2007, Aldous, 1981, Kallenberg, 1989]. La seconde propriété énonce que le graphon permet d'encoder de nombreuses informations des grands graphes, en tant qu' "objet limite" de suites convergentes de graphes. Le lecteur intéressé pourra consulter la théorie des limites de graphes qui a été introduite par Lovász and Szegedy [2006] puis fait l'objet d'une monographie [Lovász, 2012].

0.3.3 Inférence sur des graphes aléatoires, identifiabilité et vitesse minimax

L'inférence statistique sur les graphes aléatoires a donné lieu à une abondante littérature, tant théorique qu'appliquée. Voir par exemple [Matias and Robin, 2014, Rácz et al., 2017, Abbe, 2017] pour le premier point, et [Goldenberg et al., 2010, Sarkar et al., 2011] pour le second. Généralement, ce cadre d'étude suppose l'existence d'un élément inconnu dans le modèle sous-jacent et le but est d'estimer cet élément à partir d'une seule réalisation $(A_{ij})_{1\leq i,j\leq n}$ du graphe aléatoire. Nous suivons cette direction dans cette thèse, avec les modèles à espace latent. Sauf indication contraire, les paramètres X, f et μ, ainsi que les positions $(x_i)_{1\leq i\leq n}$ sont inconnues.

Si les modèles à espace latent offrent un cadre d'étude attractif, ils viennent aussi avec leur lot de défis à relever. Parmi ceux-là figurent l'absence d'identifiabilité de la structure latente. En effet, ni X ni f ni les positions (x_i) ne sont identifiables à partir des observations (A_{ij}). Ainsi, même si deux structures latentes sont différentes, elles peuvent engendrer la même loi de probabilité sur les graphes. Cette particularité des modèles à espace latent engendre des difficultés à la fois théoriques et pratiques. Ce point est abordé dans la première des trois questions de thèse. Avant d'entrer dans le vif du sujet, la définition de vitesse minimax est introduite. Elle donne une formalisation de 'meilleure vitesse' de convergence pour un estimateur. Cette vitesse est très utilisée en statistique mathématique comme point de référence pour évaluer la performance d'un estimateur.

Cadre d'étude minimax. Soit $\{\mathbb{P}_\theta : \theta \in \Theta\}$ un ensemble de lois de probabilité associées à des observations A de taille $n \times n$, tel que $\mathbb{P}_\theta$ est la loi de A sous le paramètre θ. Un estimateur $\widehat{\theta}$ basé sur les observations A, est une fonction à valeurs dans l'ensemble des paramètres Θ, qui est mesurable pour la tribu engendrée par A. L'estimateur est dit consistant s'il converge vers le paramètre sous-jacent θ, quand la taille n des observations tend vers l'infini. Cette convergence dépend du choix de la métrique sur Θ, ou plus généralement de la fonction de perte $L : \Theta \times \Theta \to \mathbb{R}_+$ (par exemple, la perte quadratique $L(\theta, \theta') = \|\theta - \theta'\|^2$). Parfois, il est plus commode d'étudier le risque de l'estimateur:

$$R_n(\theta, \widehat{\theta}) = \mathbb{E}_\theta \, L(\theta, \widehat{\theta})$$

où $\mathbb{E}_\theta$ est l'espérance sous le paramètre θ. Après avoir montré la consistance sur un grand ensemble Θ, on peut étudier la vitesse de convergence minimax associée à un sous-ensemble $\Theta_0 \subset \Theta$, qui est définie par :

$$\overline{R}_n(\Theta_0) := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_0} R_n(\theta, \widehat{\theta}) \ ,$$

où l'infimum est pris sur l'ensemble de tous les estimateurs. On peut observer que n'importe quel estimateur $\widehat{\theta}$ a un risque maximal $\sup_{\theta \in \Theta_0} R_n(\theta, \widehat{\theta})$ supérieur ou égal à la vitesse minimax $\overline{R}_n(\Theta_0)$. En ce sens, la vitesse minimax est la meilleure.

Dans cette thèse, on définira des estimateurs atteignant la vitesse minimax à constantes près, avec grande probabilité. Autrement dit, on cherchera $\widehat{\theta}$ satisfaisant

$$\overline{R}_n(\Theta_0) \lesssim \sup_{\theta \in \Theta_0} L(\theta, \widehat{\theta}) \lesssim \overline{R}_n(\Theta_0)$$

avec probabilité au moins $1 - 1/n$. La notation $a \lesssim b$ signifie qu'il existe une constante numérique $c > 0$ telle que $a \leq cb$.

0.4 Trois questions générales

0.4.1 Chapitre 1: Réconcilier interprétabilité et non-identifiabilité ?

La généralité et la versatilité du modèle de graphon semblent en faire un outil intéressant pour l'analyse de réseau. Malheureusement, le modèle souffre d'un problème majeur d'identifiabilité qui rend difficile son utilisation en pratique.

Problème d'identifiabilité et d'interprétabilité. La fonction f n'est pas identifiable à partir des observations A. En effet, pour toute bijection $\phi : \Omega \to \Omega$ préservant la mesure μ, la fonction $f^\phi(x, y) = f(\phi(x), \phi(y))$ laisse la distribution inchangée, c'est-à-dire

$$\mathbb{P}_{(X,\mu,f)} = \mathbb{P}_{(X,\mu,f^\phi)}.$$

En fait, même l'espace latent X n'est pas identifiable: pour toute bijection $\phi : X' \to X$ préservant la mesure μ, le triplet (X', μ, f^ϕ) définit la même loi que (X, μ, f). Une caractérisation plus complète de ce problème d'identifiabilité est donnée par la relation de faible isomorphisme [Lovász, 2012, chap.10]:

Des graphons (X, μ, f) et (X', μ', f') engendrent la même distribution pour tout n, si et seulement si, il existe des applications préservant la mesure $\phi : [0,1] \to X$ et $\psi : [0,1] \to X'$ telles que $f^{\phi}(x,y) = f'^{\psi}(x,y)$ presque partout.

Ici $[0,1]$ désigne un espace de probabilité muni de la mesure uniforme.

Formellement, le problème d'identifiabilité pourrait donc être résolu en quotientant l'ensemble des graphons par la relation de faible isomorphisme, puis en travaillant sur les classes d'équivalence induites (elles sont identifiables). Cependant, ces classes d'équivalence sont dures à décrire. Une classe d'équivalence donnée n'a à priori pas de représentant (X, μ, f) avec une forme simple à interpréter. L'utilisation du modèle du graphon est donc difficile en pratique.

Une direction de recherche. Un objectif général est donc de rendre le modèle de graphon plus opérationnel. Parmi les directions possibles, on pourrait définir puis estimer une fonctionnelle identifiable du graphon, qui correspond à une propriété interprétable du réseau.

Considérons une fonctionnelle $\Psi[(X, \mu, f)]$ qui est invariante par transformation préservant la mesure. Par définition, une telle fonctionnelle prend la même valeur pour tout représentant de la classe d'équivalence du graphon (X, μ, f). Ainsi, cette quantité est identifiable contrairement au triplet (X, μ, f).

On étudiera la construction puis l'estimation d'une fonctionnelle particulière caractérisant la complexité du graphon. Une attention particulière sera accordée à son interprétation sur des exemples classiques de graphes, ainsi qu'à la compréhension des vitesses minimax d'estimation.

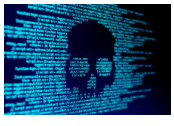
0.4.2 Chapitres 2 & 3: Stratégies d'échantillonnage séquentiel ?

Afin de motiver une question abstraite sur l'échantillonnage séquentiel, deux exemples sont présentés ci-dessous pour illustrer l'intérêt pratique de la question.

Situations pratiques

Tournoi. Soit une plate-forme de jeux vidéo en ligne regroupant deux types de joueurs, des forts des faibles. Idéalement, les joueurs aimeraient affronter des adversaires de leur niveau à chaque match, sans jamais rencontrer deux fois le même adversaire. Puisque les niveaux des joueurs sont inconnus au départ, on les découvre au fur et à mesure du tournoi, en observant les résultats des parties. Pour l'organisateur chargé de l'appariement des matchs, quelle stratégie permet de maximiser le plaisir des joueurs ?

Enquête. Un détective veut découvrir des liens cachés dans un réseau criminel, mais n'a qu'un budget limité pour son enquête. Chaque opération qui teste une connexion entre deux malfaiteurs, représente un coût. Quelle stratégie maximise le nombre de liens découverts, en respectant la limite de budget ?



Des problèmes similaires apparaissent dans d'autres applications pratiques, comme les problèmes d'appariements sur les sites de rencontre, ou l'exploration coûteuse de certains réseaux biologiques. Ces questions amènent naturellement au problème suivant, que l'on nomme 'pair-matching'.

Le problème de pair-matching. On suppose qu'il existe un graphe dont les noeuds représentent des entités, et les arêtes des liens entre ces entités. A l'instant $t = 0$, les noeuds sont connus du statisticien alors que les arêtes sont cachées. Un algorithme de matching (ou d'appariement) émet des requêtes sur des paires d'individus de façon séquentielle, c'est-à-dire, interroge une paire à chaque instant $t = 1, 2, 3, \ldots$, en essayant de découvrir autant d'arêtes que possible (voir figure page 13). Pour un réseau biologique comme un réseau protéine-protéine, les noeuds du graphe sont des protéines, une arête est une interaction entre deux protéines, et une requête de l'algorithme est une expérience du biologiste pour tester la présence d'interaction entre deux protéines.

L'algorithme de pair-matching est forcé d'explorer une nouvelle paire de noeuds à chaque instant t. Le choix de cette paire est basée sur les observations passées (aux temps $1, \ldots, t-1$). Sans hypothèses sur la structure sous-jacente du graphe, ces observations passées sont inutiles pour prendre une décision. Afin d'étudier des stratégies intéressantes qui peuvent se servir de l'information passée, on supposera donc une structure dans le graphe (par exemple, une structure de communauté (SBM)).

0.4.3 Chapitre 4: Localisation des points latents ?

Motivés par des questions non résolues en pair-matching (voir fin de la section 0.5.3), nous étudions un problème d'inférence des points latents dans un espace métrique (X, d). A partir des données $\{A_{ij}\}_{1 \le i,j \le n}$, peut-on estimer de façon uniforme les x_i dans le modèle suivant ?

$$\mathbb{E}\, A_{ij} = f(x_i, x_j) \ ,$$

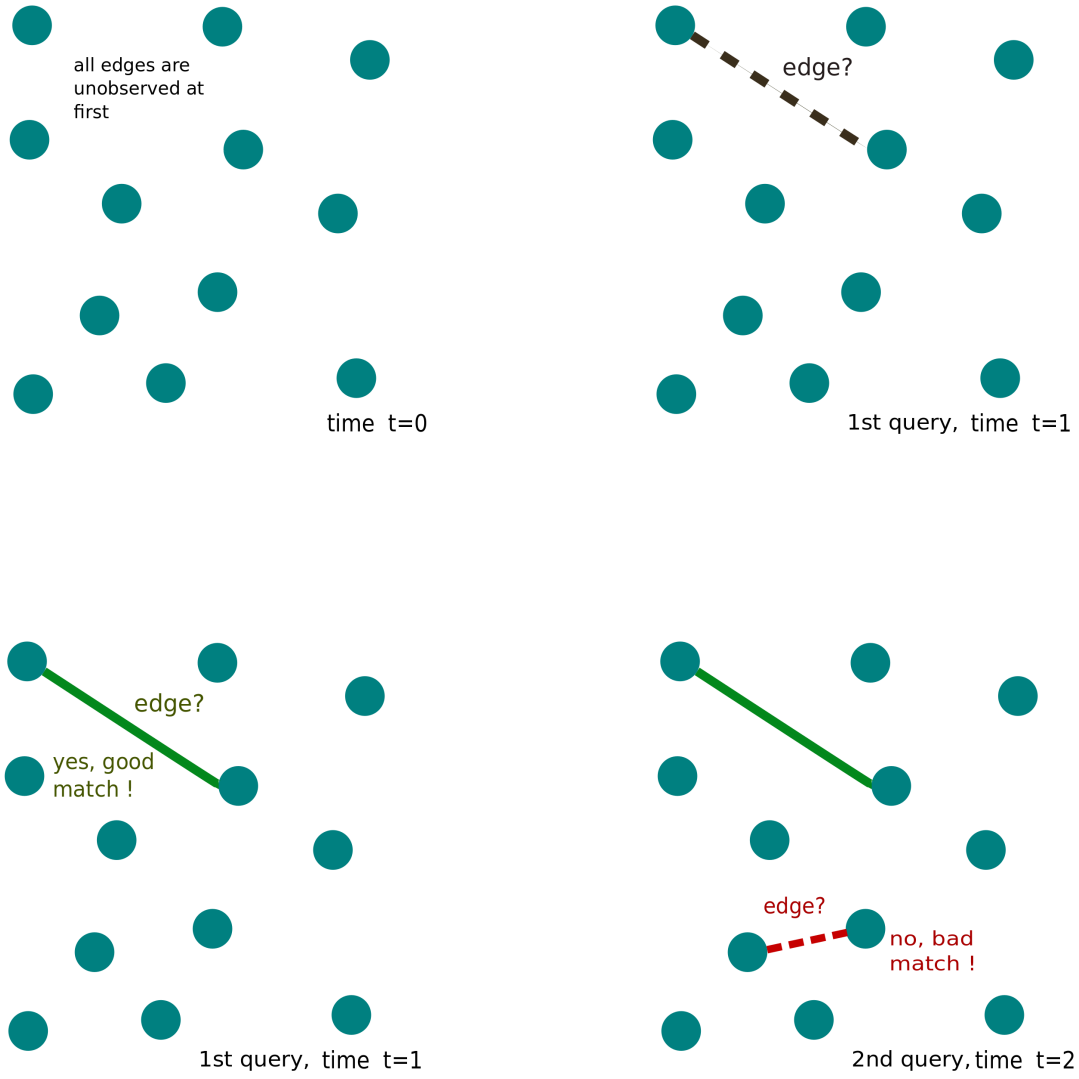
où la fonction f est inconnue mais prend des valeurs plus grandes si les points sont proches dans (X, d). Cette modélisation est raisonnable en pratique: l'affinité $f(x_i, x_j)$ des deux agents i et j est grande si $d(x_i, x_j)$ est petit, c'est-à-dire, si leurs attributs x_i et x_j sont similaires.

Cette question est formalisée dans un cadre simple pour faciliter l'analyse mathématique. On considère un espace latent uni-dimensionnel, et pour des raisons techniques, on choisit le cercle unité du plan (ci-après noté $\mathcal{C}$). On munit $\mathcal{C}$ de la distance géodésique d. Pour la contrainte de forme sur f, on suppose que f appartient à une certaine classe de fonctions bi-Lipschitz telle que $f(x, y)$ est grand si $d(x, y)$ est petit, et vice versa, $f(x, y)$ est petit si $d(x, y)$ est grand. On s'intéressera aussi au cas particulier des fonctions géométriques, qui sont définies par l'égalité suivante: $f(x, y) = \widetilde{f}(d(x, y))$, avec $\tilde{f}$ une fonction d'une variable réelle.

Étant données des observations (A_{ij}), nous construirons des estimateurs $\hat{x}_1, \ldots, \hat{x}_n$ des points latents $x_1, \ldots, x_n$. Pour résoudre une conjecture en pair-matching (présentée en section 0.5.3), nous avons besoin d'une garantie sur l'erreur uniforme des estimateurs. La fonction de perte est donc $d_\infty(\mathbf{x}, \hat{\mathbf{x}}) = \max_i d(x_i, \hat{x}_i)$.

Néanmoins, au vue des problèmes d'identifiabilité, il est impossible de retrouver les positions à partir des observations. L'absence d'identifiabilité persiste même si f est connue et égale à $f_0(x, y) = 1 - d(x, y)/(2\pi)$. Dans ce cas particulier, une solution simple est d'utiliser la pseudo-métrique suivante :

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}) = \min_{Q \in \mathcal{O}} \max_{i \in [n]} \ d\big(\hat{x}_i, Q x_i\big), \tag{0.1}$$



Algorithme de pair-matching jusqu'au temps t=2.
Pour la requête au temps $t = 1$, une arête est découverte, c'est un bon match.
Au temps $t = 2$, aucune arête n'est découverte, c'est un mauvais match.

où $\mathcal{O}$ est le groupe orthogonal de $\mathbb{R}^2$. La pseudo-métrique $d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x})$ mesure donc la performance d'un estimateur à transformations orthogonales près, ce qui résout le problème d'identifiabilité quand $f = f_0$. Le fait que f est inconnue dans le modèle d'intérêt, amène d'autres problèmes d'identifiabilité, mais nous les laissons de côté pour simplifier l'exposition ici. Nous utiliserons la pseudo-métrique $d_{\infty,\mathcal{O}}$ pour évaluer la performance d'un estimateur $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$.

0.5 Contributions

Cette section est un résumé informel de nos contributions. Des énoncés précis peuvent être trouvés aux chapitres 1-4.

0.5.1 Chapitre 1. Une fonctionnelle interprétable du graphon

Les résultats présentés dans cette section et détaillés au chapitre 1 sont issus de [Issartel, 2019].

Parmi les différentes caractérisations des réseaux, leur complexité est particulièrement populaire. Le lecteur intéressé pourra par exemple consulter les synthèses récentes de Dehmer and Mowshowitz [2011] et Zenil et al. [2018]. En pratique, la complexité des réseaux est utilisée dans de nombreuses applications, telles que l'étude de structures chimiques [Bonchev and Buck, 2005], de processus d'affaires [Latva-Koivisto] ou encore de librairies logicielles [Veldhuizen, 2005].

L'estimation de la complexité des réseaux est un domaine de recherche actif [Morzy et al., 2017, Zufiria and Barriales-Valbuena, 2017, Claussen, 2007]. Pourtant, aucune garantie statistique ne semble avoir été prouvée. Afin de couvrir ce besoin, nous introduisons un cadre statistique pour l'étude de la complexité des réseaux, en s'appuyant sur le modèle général de graphon.

Un indice de complexité. Notre premier objectif est la définition d'un indice de complexité dans le modèle de graphon. Un candidat intuitif pourrait être la dimension de l'espace latent, par exemple d si $X = [0,1]^d$. Cet indice est inadéquate puisque l'espace latent n'est pas identifiable. Pire encore, il a été montré que toute distribution de W-graphe aléatoire peut être représentée sur l'espace particulier $X = [0,1]$ [Lovász, 2012]. Il n'est donc pas convenable de penser la complexité du graphe purement en terme d'espace latent. De même, la régularité de la fonction de lien (par exemple, α si f est α-Hölder) n'est pas un indice complexité approprié, puisque la fonction f n'est pas identifiable.

Ces problèmes d'identifiabilité nous poussent à considérer un indice plus abstrait. Étant donné un graphon (X, μ, f), on munit l'espace latent X de la *distance de voisinage*:

$$r_f(x, x') = \left(\int_X \left| f(x, x'') - f(x', x'') \right|^2 \, \mu(dx'') \right)^{1/2}. \tag{0.2}$$

A partir de la description d'un W-graphe aléatoire vue précédemment, on peut voir que la quantité $r_f(x_i, x_j)$ mesure la propension des noeuds i et j à être connectés avec des noeuds similaires. Notre indice de complexité est alors défini comme le covering number

et la dimension de Minkowski d'une version 'purifiée' de l'espace pseudométrique (X, r_f). Le processus de purification de (X, r_f) est détaillé au chapitre 1. On rappelle la définition de ces mesures standards : le ϵ-covering number $N_X^{(c)}(\epsilon)$ est le nombre minimal de boules de rayon ϵ requis pour couvrir entièrement l'espace (X, r_f); Et la dimension de Minkowski est définie par la limite suivante sur le covering number:

$$dim\, X := \lim_{\epsilon \to 0} \frac{\log\, N_X^{(c)}(\epsilon)}{-\log \epsilon} \tag{0.3}$$

lorsque la limite existe. En particulier, la dimension de Minkowski n'est pas toujours entière.

Bien que les paramètres X, μ et f ne soient pas identifiables, nous prouvons l'identifiabilité de l'indice de complexité. De plus, ce dernier est facile à interpréter. Dans un SBM, nous verrons que le covering number est égal au nombre de communautés bien espacées; Dans des graphes géométriques, la dimension de Minkowski correspond à la dimension euclidienne de l'espace latent; Et dans des modèles de graphons Hölder, la dimension de Minkowski est égale à la régularité de la fonction f.

Idéalement, cet indice pourrait servir à l'ajustement de méthodes analytiques sur les réseaux, par exemple pour l'apprentissage de représentation [Hoff et al., 2002, Perozzi et al., 2014, Grover and Leskovec, 2016]. On s'appuierait alors sur l'estimation de la complexité pour choisir un espace latent dans lequel la forme du graphon est interprétable.

Estimation statistique de l'indice. A partir de la matrice d'adjacence (A_{ij}) d'un W-graphe aléatoire, nous estimons la distance de voisinage (0.2) sur les points latents $x_1, \dots, x_n$. Pour l'estimateur de distance $\widehat{r}$ défini au chapitre 1, le théorème suivant donne des bornes non-asymptotiques et universelles. On note $x_{m(i)} \in \{x_1, \dots, x_n\} \setminus \{x_i\}$ un plus proche voisin de x_i par rapport à la pseudo-distance r_f.

Théorème 0.5.1 *Pour tout graphon (X, μ, f), il y a une probabilité supérieure à $1 - 2/n$ d'avoir les bornes suivantes.*
$\forall i, j \in [n]$,

$$\left| r_f^2(x_i, x_j) - \widehat{r}^2(i,j) \right| \lesssim r_f(x_j, x_{m(j)}) + r_f(x_i, x_{m(i)}) + \sqrt{\log(n)/n} \ .$$

Le terme $r_f(x_j, x_{m(j)})$ dans la borne ci-dessus est un terme de biais, qui correspond à la distance de voisinage entre le point x_j et un de ses plus proches voisins $x_{m(i)}$. Ce biais dépend de la forme du graphon sous-jacent: pour un SBM, par exemple, il est égal à zéro avec grande probabilité.

Réciproquement, pour n'importe quel estimateur, l'erreur d'estimation des distances est supérieure ou égale à la borne du Théorème 0.5.1 (à constante multiplicative près). En d'autres termes, nous montrons que l'estimateur $\widehat{r}$ est optimal (au sens minimax), sur l'ensemble de tous les graphons.

Les estimées $\widehat{r}(i,j)$ permettent ensuite de calculer le covering number par plug-in. Nous obtenons des bornes d'erreur non-asymptotiques et universelles pour cet estimateur du covering number. Les résultats sur la distance et le covering number sont ainsi valides sur tous les graphons, contrairement à la majorité des résultats dans la littérature.

En combinant l'estimateur du covering number avec la formule (0.3), on déduit un estimateur de la dimension de Minkowski:

$$\widehat{dim}_D := \frac{\log \widehat{N}_X^{(c)}(\epsilon_D)}{-\log \epsilon_D} \ .$$

Cet estimateur satisfait la borne d'erreur du Théorème 0.5.2 avec grande probabilité. Ce résultat requiert certaines hypothèses sur le graphon. La dimension de Minkowski est bornée supérieurement par une constante D. Le rayon ϵ_D du covering number est alors choisi en fonction de cette constante. On suppose aussi des hypothèses faibles sur la géométrie du graphon, qui sont assez similaires à celles de la littérature sur l'estimation de la dimension d'une variété. Réciproquement, nous montrons que cet ensemble d'hypothèses est minimal, c'est-à-dire: si l'une des hypothèses est retirée, tout estimateur de la dimension est forcément non-consistant.

Théorème 0.5.2 *Sous de faibles hypothèses, et si $\dim X$ appartient à $[0, D]$, l'estimateur $\widehat{dim}_D$ satisfait l'inégalité suivante*

$$\left| \widehat{dim}_D - dim\, X \right| \lesssim \frac{1}{\log n}$$

avec probabilité supérieure à $1 - C'/n$, pour une constante C' indépendante de n.

Notre estimateur de la dimension de Minkowski converge donc à la vitesse $\log^{-1} n$. Nous prouvons qu'il s'agit de la vitesse minimax pour le problème considéré.

Nous généralisons ce travail au cas significatif des graphes creux, qui a été considéré à plusieures reprises dans la littérature [see Bickel et al., 2011, Wolfe and Olhede, 2013, Klopp et al., 2017, Xu et al., 2014].

Puisque la complexité en temps du covering number est exponentiel, nous proposons un estimateur alternatif avec un temps de calcul polynomial, et quelques garanties théoriques sur sa performance.

Finalement, nous testons si le packing number de (X, r_f) est plus petit qu'un entier K, avec un soin particulier pour contrôler la probabilité d'erreur de type I. Nous prouvons que cette erreur est plus petite que $2/n$, uniformément sur tous les graphons. Pour des raisons techniques ici, le packing number remplace le covering number, mais ce sont essentiellement les mêmes mesures de complexité.

0.5.2 Chapitre 2. Pair-matching dans un SBM à 2 groupes

Les résultats présentés dans cette section et détaillés au chapitre 2 sont issus de [Giraud et al., 2019] et ont été obtenus en collaboration avec Christophe Giraud, Luc Lehéricy et Matthieu Lerasle.

On étudie le problème de pair-matching dans une situation simple où le graphe est généré par un SBM à deux communautés. Soient p et q les probabilité de connexion intra-communautés et inter-communautés. Deux noeuds issus de la même communauté ont une plus grande probabilité d'être connectés que deux noeuds issus de communautés différentes,

c'est-à-dire, $p > q$. Pour simplifier l'analyse, on considère un SBM équilibré et conditionnel, signifiant que les communautés sont de même taille et que les étiquettes sont déterministes. On supposera aussi que le ratio p/q est majoré par une constante numérique.

Rappelons l'objectif du statisticien dans le problème de pair-matching: trouver le plus grand nombre d'arêtes possible avec un nombre limité d'action. Une stratégie peut donc être évaluée par le nombre d'arêtes qu'elle découvre, en moyenne, après T requêtes. Cette "moyenne" est prise par rapport à l'aléa du graphe aléatoire et celui de l'algorithme. Au vu de cet objectif, on distingue deux types de paires dans le SBM: les "bonnes paires" de noeuds, qui ont une plus grande probabilité d'avoir une arête, et ont donc des noeuds issus de la même communauté, puis les "mauvaises paires", composées par des noeuds de communautés différentes. Un algorithme de pair-matching devrait donc visiter autant de bonnes paires que possible. De façon équivalente, nous évaluerons la qualité d'une stratégie par le nombre moyen de mauvaises paires qu'elle a tirées. Cette quantité est appelée *regret d'échantillonnage* et notée $\mathbb{E}\left[N^{bad}(T)\right]$ dans la suite. La difficulté à minimiser ce regret d'échantillonnage vient de la partition entre bonnes et mauvaises paires qui est inconnue.

Un paramètre clé dans notre analyse est le ratio

$$s = \frac{(p-q)^2}{p+q} \ .$$

Ce paramètre apparaît dans divers résultats de la littérature sur le SBM. Par exemple, la propriété suivante, prouvée entre autre dans [Yun and Proutière, 2014a, Chin et al., 2015, Fei and Chen, 2019, Giraud and Verzelen, 2019], est utile dans notre algorithme de pair-matching. Pour un graphe généré selon le modèle de SBM décrit ci-dessus, il existe un algorithme de clustering avec une complexité en temps polynomiale qui retourne une partition des noeuds telle que, avec grande probabilité, la proportion de noeuds mal-classifiés décroît exponentiellement:

$$\text{Proportion de noeuds mal classifiés} \leq \exp(-cns), \quad \text{dès que } ns \geq c' \ , \tag{0.4}$$

où c, c' sont des constantes numériques. Le taux ns de décroissance exponentielle est optimal sous certaines hypothèses sur p et q. Le paramètre s mesure ainsi la difficulté de clustering.

Regret d'échantillonnage optimal. Notre principale contribution dans [Giraud et al., 2019] est de montrer que le regret d'échantillonnage est optimal s'il est de l'ordre de

$$\mathbb{E}\left[N^{bad}(T)\right] \asymp T \wedge \frac{\sqrt{T}}{s} \ . \tag{0.5}$$

En effet, nous montrons que, pour n'importe quel algorithme de pair-matching, le regret d'échantillonnage est supérieur à $\mathbb{E}\left[N^{bad}(T)\right] \gtrsim T \wedge \frac{\sqrt{T}}{s}$, à constante multiplicative près. Et réciproquement, on décrit un algorithme polynomial (en temps de calcul) dont le regret est borné par une constante fois $T \wedge (\sqrt{T}/s)$. Ces résultats montrent qu'aucune stratégie ne peut réaliser un regret sous-linéaire avant $T = O(1/s^2)$ paires tirées, et d'autre part que, il existe des stratégies avec regret sous-linéaire $\sqrt{T}/s$ dès que $T \gtrsim 1/s^2$.

Ce résultat peut être compris intuitivement. Tant que les communautés ne peuvent être retrouvées mieux qu'au hasard, il n'y a pas d'espoir d'avoir un regret d'échantillonnage

meilleur qu'avec un tirage au hasard des paires. Dans ce régime, le regret (d'échantillonnage) croît linéairement avec T. Pour identifier quand cela se produit, on peut considérer la situation où toutes les arêtes sont tirées entre N noeuds et $T \asymp N^2/2$. Il est connu d'après [Decelle et al., 2011, Massoulié, 2014, Mossel et al., 2015] que la détection de communautés est possible si et seulement si $N(p-q)^2 \geq 2(p+q)$, ce qui donne $\sqrt{T}s \gtrsim 1$, ou encore $T \gtrsim 1/s^2$. Ainsi, aucune information sur les communautés ne peut être retrouvée tant que $T \leq c/s^2$, où c est une constante numérique suffisamment petite. On s'attend donc à ce que le regret grandisse linéairement avec T, pour $T = O(1/s^2)$. Cette intuition est confirmée par (0.5).

Quand $T \gg 1/s^2$, la situation est différente. D'après la décroissance exponentielle (0.4), les communautés d'un graphe à N noeuds peuvent être retrouvées presque parfaitement si $N \gg 1/s$. Donc, pour $1/s \ll N \leq (\sqrt{T}/s)^{1/2} \ll \sqrt{T}$, on peut tirer toutes les paires entre N noeuds et retrouver leur communauté avec un regret plus petit que $\sqrt{T}/s$.

Étant donnés ces noeuds classifiés (qu'on utilise maintenant comme noeuds de référence), il est possible d'identifier la communauté d'un nouveau noeud, avec un regret de l'ordre de $O(1/s)$. En procédant récursivement, $\Theta(\sqrt{T})$ nouveaux noeuds peuvent être identifiés avec un regret de l'ordre de $O(\sqrt{T}/s)$. Quant au budget restant, il est dépensé en tirant les paires entre ces $O(\sqrt{T})$ noeuds, en fonction de leur communauté d'appartenance (donc sans affecter le regret). Ce raisonnement informel suggère que le regret optimal croît comme $\sqrt{T}/s$ quand $T \gg 1/s^2$. Une nouvelle fois cette intuition est confirmée par (0.5).

Quand il est interdit d'échantillonner les mêmes noeuds de nombreuses fois. Ci-dessus, les stratégies sont autorisées à échantillonner un même noeud autant de fois que voulu. En particulier, notre algorithme (celui qui atteint le regret minimal (0.5)) utilise cette possibilité de façon excessive: on dit que ses requêtes sont localisées dans le graphe. Plus précisément, dans un petit groupe de $\Theta(\sqrt{T})$ noeuds, chaque noeud a été sondé $\Theta(\sqrt{T})$ fois. Cette caractéristique peut être problématique dans certaines applications de la vie réelle où les individus ne peuvent être sollicités trop de fois. Pour modéliser ces situations, nous imposons un nombre maximal de requêtes par noeud : les stratégies ne peuvent interroger un individu plus de B_T fois au cours des T requêtes. Sous cette nouvelle règle, nous montrons que le regret optimal est de l'ordre de

$$\mathbb{E}\left[N^{bad}(T)\right] \asymp T \wedge \frac{\sqrt{T} \vee (T/B_T)}{s} \quad .$$

En comparant avec (0.5), on observe que la contrainte d'échantillonnage revient à remplacer $\sqrt{T}$ par $\sqrt{T} \vee (T/B_T)$ dans le regret optimal. Pour une contrainte (trop) forte comme $B_T = O(1/s)$, il n'y a pas assez d'observations par noeud pour inférer leur communauté (mieux qu'au hasard), ce qui induit inévitablement un regret linéaire pour toute stratégie. A l'opposé, la formule ci-dessus montre qu'une contrainte faible $B_T \gtrsim \sqrt{T}$ n'a pas d'effet sur le regret optimal. Cela n'est pas surprenant puisque notre algorithme (celui avec la performance (0.5)) n'émet pas plus de $O(\sqrt{T})$ requêtes par noeud. Finalement, on peut observere que la contrainte d'échantillonnage détériore le regret optimal dès que $B_T \lesssim \sqrt{T}$.

0.5.3 Chapitre 3. Pair-matching dans des modèles plus généraux (bornes inférieures et conjectures)

Les résultats présentés dans cette section et détaillés au chapitre 3 sont issus de travaux en cours [Issartel, 2020b] et [Issartel, 2020c].

Nous continuons l'étude du problème de pair-matching dans des modèles plus complexes: dans un SBM à K groupes et dans un graphe géométrique. Des contraintes seront aussi ajoutées sur les stratégies pour les forcer à explorer l'espace latent.

Dans un SBM à K communautés. Le graphe est généré par un SBM conditionnel à K communautés de même taille. Les probabilités intra et inter-communauté sont notées p et q et satisfont $p > q$ avec $p < 1/2$. On supposera aussi que le ratio p/q est majoré par une constante numérique (ce qui exclut en particulier le régime trivial $p = 1$ et $q = 0$). Dans ce modèle, Giraud et al. [2019] conjecturent les regrets optimaux suivants.

Sur l'ensemble de tous les algorithmes, sans contrainte de temps de calcul:

$$\mathbb{E}\left[N^{bad}(T)\right] \asymp \left(\left(\frac{K\log(K)}{s}\right)^2 \vee \frac{K\sqrt{T}}{s}\right) \wedge T \ . \tag{0.6}$$

Pour les algorithmes dont la complexité en temps est au plus polynomiale:

$$\mathbb{E}\left[N^{bad}(T)\right] \asymp \left(\left(\frac{K^2}{s}\right)^2 \vee \frac{K\sqrt{T}}{s}\right) \wedge T \ . \tag{0.7}$$

La conjecture ci-dessus donne des bornes supérieures et inférieures sur les regrets d'échantillonnage, caractérisant ainsi les regrets optimaux. Pour une discussion détaillée de cette conjecture, le lecteur pourra consulter le chapitre 2. Nous donnons ci-dessous un résultat partiel sur les bornes inférieures.

On peut voir que les deux regrets conjecturés (0.6) et (0.7) ont un terme en commun $(K\sqrt{T}/s)\wedge T$. En généralisant les techniques de preuve de [Giraud et al., 2019], nous montrons une borne inférieure pour ce terme (Théorème 0.5.3).

Théorème 0.5.3 *Sous certaines hypothèses sur les paramètres p et q, on a*

$$\mathbb{E}\left[N^{bad}(T)\right] \gtrsim \frac{K\sqrt{T}}{s} \wedge T.$$

Ainsi, pour les bornes inférieures correspondant à (0.6) et (0.7), il ne reste plus que les termes $(K\log(K)/s)^2$ et $(K^2/s)^2$ à prouver. Malheureusement, nous n'avons pas de preuve pour eux. Ces termes sont purement conjecturels et basés sur les arguments informels ci-après.

Tant que les communautés ne peuvent être détectées, le regret croît linéairement avec T. Comme dans la section précédente, on identifie la terminaison de cette phase en tirant toutes les arêtes entre N noeuds, avec $T \asymp N^2/2$. Il a été prouvé dans [Banks et al., 2016] que la détection des communautés est impossible en dessous du seuil d'information $Ns \lesssim K\log(K)$, ce qui donne $T \lesssim (K\log(K)/s)^2$. Cette intuition mène au terme dans (0.6). Il a été conjecturé dans [Decelle et al., 2011] qu'aucun algorithme en temps polynomial ne réussit

une classification non-triviale des communautés, dès que $Ns \lesssim K^2$, soit $T \lesssim (K^2/s)^2$. Ce raisonnement informel donne le premier terme dans (0.7).

Quand les stratégies sont contraintes d'explorer de nombreuses communautés. Dans un jeux vidéo en ligne, par exemple, le mauvais algorithme de pair-matching fait jouer peu de personnes de nombreuses fois, et beaucoup de personnes peu de fois. C'est ce que font les stratégies optimales dans le SBM à K classes, en concentrant leurs requêtes sur une seule et même communauté. Afin d'éviter cet inconvénient, nous allons ajouter une règle qui force les stratégies à explorer un grand nombre de communautés différentes.

La contrainte d'exploration se définit ainsi. Dans le SBM à K communautés, il existe une partition inconnue des noeuds en K groupes $G_1, G_2, \ldots, G_K$. Parmi les T paires interrogées, on note $N_{G_i}(T)$ le nombre d'entre elles appartenant à $G_i \times G_i$, c'est-à-dire, composées par deux noeuds du groupe G_i. Soient c_G, c_P et h_{BS} des réels dans $(0,1)$. Alors, avec une probabilité plus grande que c_P,

$$N_{G_i}(T) \geq h_{BS} \frac{T}{K} \tag{0.8}$$

pour au moins $c_G K$ groupes différents. Le paramètre h_{BS} représente la force de la contrainte d'exploration.

Pour les stratégies respectant la contrainte (0.8), nous conjecturons les regrets optimaux suivants.

Sur l'ensemble de tous les algorithmes, sans contrainte de temps de calcul:

$$\mathbb{E}\left[N^{bad}(T)\right] \asymp \left(\left(\frac{K\log(K)}{s}\right)^2 \vee \frac{K\sqrt{(h_{BS}K \vee 1)T}}{s}\right) \wedge T \ .$$

Pour les algorithmes dont la complexité en temps est au plus polynomiale:

$$\mathbb{E}\left[N^{bad}(T)\right] \asymp \left(\left(\frac{K^2}{s}\right)^2 \vee \frac{K\sqrt{(h_{BS}K \vee 1)T}}{s}\right) \wedge T \ .$$

Comparés aux regrets (0.6) et (0.7) du cas non-contraint, le terme $(K\sqrt{T})/s$ est le seul à changer: il augmente d'un facteur $\sqrt{(h_{BS}K \vee 1)}$ sous l'effet de la contrainte (0.8). En particulier, nous prouvons (Théorème 0.5.4) une borne inférieure pour ce nouveau terme $K\sqrt{(h_{BS}K \vee 1)T})/s$ qui apparaît dans les deux regrets ci-dessus.

Théorème 0.5.4 *Pour n'importe quelle stratégie satisfaisant la contrainte (0.8), on a*

$$\mathbb{E}\left[N^{bad}(T)\right] \gtrsim \frac{\sqrt{((h_{BS}K) \vee 1)}\, K\sqrt{T}}{s} \wedge T \ .$$

De façon identique au cas non-contraint, les termes $(K\log(K)/s)^2$ et $(K^2/s)^2$ restent à prouver pour l'obtention de bornes inférieures complètes. A notre avis, la démonstration des bornes supérieures devrait (elle aussi) être similaire à celle des conjectures (0.6) et (0.7) sans contrainte. Le lien étroit entre ces deux cas (non-contraint - contraint) suggère que le dernier suivra une fois le premier résolu. Nous étayerons ces conjectures au chapitre 3.

Sur un graphe géométrique. Nous continuons l'étude du problème de pair-matching dans un modèle de graphe "continu", qui est un cas particulier de graphe géométrique. Sur le tore $[0,1)$, muni de la distance de plus court chemin d, on définit une fonction f, affine en la distance, par $f(x,y) = (3/4) - d(x,y)/4$. Les points latents $x_1, \dots, x_n \in [0,1)$ sont de la forme spécifique $\sigma(1)/n, \dots, \sigma(n)/n$, avec σ une permutation inconnue de $[n]$. La probabilité de connexion entre des noeuds i et j vaut alors $f(x_i, x_i)$. Dans la suite, cette probabilité sera simplement notée P_e pour n'importe quelle paire $e = \{i,j\}$.

Rappelons que l'objectif est de maximiser le nombre moyen d'arêtes découvertes. En notant $(\widehat{e}_1, \dots, \widehat{e}_T)$ la séquence de paires interrogées, ce nombre moyen vaut $\mathbb{E}\left[\sum_{t=1}^T A_{\widehat{e}_t}\right] = \sum_{t=1}^T \mathbb{E}\left[P_{\widehat{e}_t}\right]$. De façon équivalente, on cherchera à minimiser le regret associé, c'est-à-dire, la différence entre la meilleure performance possible et celle de la stratégie $(\widehat{e}_1, \dots, \widehat{e}_T)$. Ce regret s'écrit:

$$\mathbb{E}\left[R_T\right] = \sum_{t=1}^T P_{e_t^*} - \sum_{t=1}^T \mathbb{E}\left[P_{\widehat{e}_t}\right] \ ,$$

où $P_{e_1^*} \geq \ldots \geq P_{e_T^*}$ désignent les T plus grandes probabilités de connexion dans le graphe.

La contrainte suivante est similaire à celle écrite dans le SBM. Elle impose (avec une certaine probabilité) une exploration d'une portion de l'espace latent, de façon linéaire en T. En d'autres mots, il existe une portion du tore $[0,1)$ dans laquelle tout intervalle I (de longueur $|I|$ pas trop petite) est échantillonné plus de $|I|T$ fois, c'est-à-dire :

$$\sum_{t=1}^T \mathbf{1}_{\widehat{e}_t \in I \times I} \gtrsim |I|\, T \ . \tag{0.9}$$

Nous démontrons dans le théorème suivant une borne inférieure sur le regret des stratégies satisfaisant (0.9). La démonstration est une généralisation des théorèmes 0.5.4 et 0.5.3.

Théorème 0.5.5 *Pour toute stratégie respectant la contrainte d'exploration (0.9) on a*

$$\mathbb{E}\left[R_T\right] \gtrsim T^{4/5} \ .$$

Réciproquement, nous décrirons un algorithme dont la performance conjecturelle est $T^{4/5}$ (à un possible facteur logarithmique près). Si nos arguments sont solides, la preuve de cette borne supérieure est toutefois incomplète. La pièce manquante est un algorithme de localisation des points latents, qui devrait satisfaire la propriété suivante (ou une version faible dans laquelle les x_i sont 'bien répartis' en un certain sens).

Existence Conjecturelle d'un Algorithme de Localisation Uniforme.
Pour n'importe quels points latents $x_1, \dots, x_N$ dans $[0,1)$, il existe des estimateurs $\hat{x}_1, \dots, \hat{x}_N$ tels que, pour une certaine transformation $\widehat{Q}$ préservant les distances sur le tore,

$$\max_{i \in [N]} d(\widehat{Q}x_i, \hat{x}_i) \lesssim \sqrt{\frac{\log(N)}{N}} \ , \tag{0.10}$$

avec grande probabilité.

0.5.4 Chapitre 4. Plongement optimal des données dans un espace métrique

Les résultats présentés dans cette section et détaillés au chapitre 4 sont issus du travail [Issartel, 2020a] en préparation.

Au chapitre 4, nous prouvons l'existence d'un algorithme satisfaisant (0.10), quand les $x_1, \dots, x_n$ sont répartis de manière régulière sur le cercle unité $\mathcal{C}$, et la probabilité de connexion vaut $P_{ij} = f(x_i, x_j)$, avec f une fonction bi-Lipschitz (inconnue). Malheureusement, cette hypothèse de répartition sur les x_i est assez restrictive et elle ne répond pas au besoin du pair-matching dans le graphe géométrique.

Pour des raisons de clarté au chapitre 4, on note les 'vrais' points latents avec une étoile, par $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$. On appelle 'vecteur régulier' tout vecteur qui s'écrit sous la forme $(e^{i2\pi\sigma(j)/n})_{j\in[n]}$, avec σ une permutation de $[n]$. Soit $\mathcal{P}_n$ l'ensemble des $n!$ vecteurs réguliers. Notre algorithme de localisation ci-dessous est performant si le vecteur latent $\mathbf{x}^*$ est bien approché par un vecteur régulier. C'est en particulier vrai si $\mathbf{x}^*$ est un échantillon uniforme de l'espace latent, ce qui est une modélisation courante dans la littérature statistique sur les graphes [Klopp et al., 2017, De Castro et al., 2017]. On simplifie l'exposition ici en mettant les problèmes d'identifiabilité de côté; On suppose aussi avoir deux copies i.i.d. $A^{(1)}$ et $A^{(2)}$ (au lieu d'une seule A).

Algorithme 1. L'algorithme se décompose en deux étapes. Soit d_1 la distance définie définie sur $\mathcal{C}^n$ par $d_1(\mathbf{x}, \mathbf{x}') = \sum_i d(x_i, x_i')$. L'étape d'initialisation estime les positions latentes en distance d_1, en utilisant seulement les données observées. Ensuite, l'étape de raffinement s'appuie sur ces positions estimées pour les améliorer et obtenir des garanties en distance d_∞.

Étape 1: initialisation en distance d_1. Pour tout vecteur de points $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{C}^n$, on définit la matrice des distances associées $D(\mathbf{x}) = \big(d(x_i, x_j)\big)_{1\le i,j\le n}$. Puis, on considère l'estimateur suivant:

$$\hat{\mathbf{x}}' = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{P}_n} \langle A^{(1)}, D(\mathbf{x})\rangle \ . \tag{0.11}$$

On cherche ainsi un vecteur $\hat{\mathbf{x}}' = (\hat{x}_1', \dots, \hat{x}_n')$ tel que les distances $d(\hat{x}_i', \hat{x}_j')$ sont petites lorsque le signal $\mathbb{E}\, A_{ij}^{(1)} = f(x_i^*, x_j^*)$ est grand. D'après la contrainte de forme sur f, le signal $f(x_i^*, x_j^*)$ est grand si les distances sous-jacentes $d(x_i^*, x_j^*)$ sont petites. On a donc bien $d(\hat{x}_i', \hat{x}_j')$ petit quand les vraies distances $d(x_i^*, x_j^*)$ sont petites. A partir de cette estimation des distances, on arrive à obtenir des garanties sur l'erreur d'estimation des positions en distance d_1. Avec un tel cheminement, on peut remarquer que ces garanties sont au mieux valables à transformations orthogonales près (ces transformations étant celles qui préservent les distances dans $\mathcal{C}$).

Étape 2 : raffinement en distance d_∞. Soit $\mathcal{C}_k = \{e^{i2\pi(j-1)/k};\ j = 1, \dots, n\}$ la grille régulière de taille k du cercle. Pour tout $x \in \mathcal{C}$, on définit le vecteur distance entre $x \in \mathcal{C}$ et $\mathbf{x} \in \mathcal{C}^n$ par $D(x, \mathbf{x}) = \big(d(x, x_1), \dots, d(x, x_n)\big)$. Alors, pour $i = 1, \dots, n$, on calcule

$$\hat{x}_i = \operatorname*{argmin}_{x\in\mathcal{C}_n} \langle A_i^{(2)}, D(x, \hat{\mathbf{x}}')\rangle, \tag{0.12}$$

pour obtenir l'estimateur $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$. Chaque calcul (0.12) a une complexité linéaire en temps, et ces n calculs sont parallélisables.

Garanties théoriques. On considère le modèle suivant $A_{ij}^{(t)} = f(x_i^*, x_j^*) + E_{ij}^{(t)}$, où $E^{(t)}$ est une matrice sous-Gaussienne, $t \in [2]$, et où f est une fonction inconnue dans une

certaine classe de fonctions bi-Lipschitz (cette classe est définie formellement au chapitre 4). Ce modèle recouvre notamment le cas des graphes aléatoires. Le théorème suivant donne une borne uniforme sur l'erreur d'estimation de $\hat{\mathbf{x}}$.

Théorème 0.5.6 *Avec une probabilité plus grande que* $1 - 1/n$, *on a*

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim \min_{\mathbf{x} \in \mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}} \ . \tag{0.13}$$

L'inégalité oracle (0.13) est vraie pour n'importe quel vecteur de positions $\mathbf{x}^* \in \mathcal{C}^n$, même s'il n'est pas un vecteur régulier de $\mathcal{P}_n$. La borne d'erreur contient deux termes. Le premier correspond à l'erreur d'approximation entre $\mathbf{x}^*$ et l'ensemble cible $\mathcal{P}_n$. Le second terme est de l'ordre de $\sqrt{\log(n)/n}$ et s'avère être optimal (voir ci-dessous).

En particulier, le terme de biais dans Théorème 0.5.6 devient négligeable si le vecteur de points latents est bien approché par un vecteur régulier, c'est-à-dire quand $\min_{\mathbf{x} \in \mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}^*) \lesssim \sqrt{\log(n)/n}$. On obtient alors

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim \sqrt{\frac{\log(n)}{n}} \ ,$$

avec grande probabilité. Réciproquement, nous montrons que la vitesse $\sqrt{\log(n)/n}$ est minimax sur la classe de points considérée. Par exemple, dans la modélisation courante où $\mathbf{x}^*$ est tiré aléatoirement dans l'espace latent selon la loi uniforme, la condition $\min_{\mathbf{x} \in \mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}^*) \lesssim \sqrt{\log(n)/n}$ est satisfaite avec grande probabilité, donc la borne ci-dessus suit.

Un défaut majeur de l'estimateur (0.11) est son temps de calcul exponentiel. Nous proposons donc une alternative avec une complexité en temps polynomial, qui consiste à remplacer l'étape 1 de Algorithme 1 par un algorithme spectral, tout en conservant l'étape 2 comme avant. Cet algorithme spectral a déjà été étudié dans un cadre sans bruit pour des problèmes de seriation [Recanati et al., 2018]. Nous analysons cet Algorithme 1 modifié dans un cas particulier du modèle de cette section: on suppose en plus que f appartient à l'ensemble des fonctions géométriques, et que le vecteur des x_i est bien approché par un des vecteurs réguliers de $\mathcal{P}_n$. Dans ce cas, nous avons des garanties sur l'estimation des positions en distance $d_{\infty,\mathcal{O}}$. Si de plus f est affine en la distance d, c'est-à-dire $f(x, y) = c - c' d(x, y)/(2\pi)$, alors l'estimateur des positions converge à vitesse $\sqrt{\log(n)/n}$, ce qui est la vitesse minimax sur cette classe.

Chapter 1

On the Estimation of Network Complexity: the dimension of graphon

Network complexity has been studied for over half a century and has found a wide range of applications. Many methods have been developed to characterize and estimate the complexity of networks. However, there has been little research with statistical guarantees. In this paper, we develop a statistical theory of graph complexity in a general model of random graphs, the so-called graphon model.

Given a graphon, we endow the latent space of the nodes with the neighborhood distance that measures the propensity of two nodes to be connected with similar nodes. Our complexity index is then based on the covering number and the Minkowksi dimension of (a purified version of) this metric space. Although the latent space is not identifiable, these indices turn out to be identifiable. This notion of complexity has simple interpretations on popular examples of random graphs: it matches the number of communities in stochastic block models; the dimension of the Euclidean space in random geometric graphs; the regularity of the link function in Hölder graphon models.

From a single observation of the graph, we construct an estimator of the neighborhood-distance and show universal non-asymptotic bounds for its risk, matching minimax lower bounds. Based on this estimated distance, we compute the corresponding covering number and Minkowski dimension and we provide optimal non-asymptotic error bounds for these two plug-in estimators.

Contents

1.1 Introduction

Networks appear in many areas where data is a collection of objects interacting with each other. Examples include numerous phenomena in the fields of physics, biology, neuroscience and social sciences. A major issue is to extract information from these data repositories. This exciting challenge has led researchers to seek characterizations of networks, among which their complexity has received a lot of attention for more than half a century. See [Dehmer and Mowshowitz, 2011, Zenil et al., 2018] for two recent reviews. Indeed, network complexity is a key feature used in various applications, for example, to quantify the complexity of chemical structures [Bonchev and Buck, 2005], to describe business processes [Latva-Koivisto], to characterize software libraries [Veldhuizen, 2005], and to study general graphs [Constantine, 1990].

The definition and estimation of network complexity is an active line of research [Morzy et al., 2017, Zufiria and Barriales-Valbuena, 2017, Claussen, 2007]. However, there appear to be little (or no) mathematical results on the statistical side of the problem. In this paper, we develop a statistical theory of graph complexity in a universal model of random graphs. To the best of our knowledge, it is the first contribution on complexity estimation with statistical guarantees.

1.1.1 Modeling assumption

Statistical inference on random graphs is a fast-growing area of research [Matias and Robin, 2014, Rácz et al., 2017, Abbe, 2017] and has found a wide range of applications [Goldenberg et al., 2010, Sarkar et al., 2011]. Usually, it assumes there exists an unknown feature in the

underlying model and the goal is to recover this feature from a single realization of the random graph.

Here, we follow this direction with the *W-random graph model* (also known as graphon model). This general model falls into the category of non-parametric descriptions of networks [Bickel and Chen, 2009] and satisfies some forms of universality [Diaconis and Janson, 2007]. See section 1.1.3 for details. In this paper, we define a notion of complexity for this model and then consider the problem of inferring this complexity from a single graph observation.

W-random graphs allow to model many real-world networks, such as social networks where nodes represent different people and edges people's friendships. In this example, one may expect that the friendship probability p_{ij} between individual i and j depends on their personal attributes (like jobs, ages, leisure). To model such mechanism, one may assume the observed graph is generated according to the W-random graph model, i.e. 1/for each node i of the network, an attribute ω_i is drawn from a distribution μ on a space Ω (where Ω can be seen as the social space of all possible individual features: jobs, ages,...); 2/two people are friends, independently of the others, with probability $p_{ij} = W(\omega_i, \omega_j)$, where $W : \Omega \times \Omega \to [0,1]$ is a symmetric function. Thus, a W-random graph is specified by the triplet of parameters (Ω, μ, W), often called *graphon* in the literature [Lovász, 2012].

Such modeling falls into the popular "latent space approach" [Hoff et al., 2002]. Indeed, the personal attributes may not be observed in practice and accordingly, the W-random graph model assumes that the ω_i and Ω are latent (unobserved). In fact, all parameters of the graphon (Ω, μ, W) are unknown, and the only observation is the edges of the graph, i.e. the adjacency matrix A where $A_{ij} = 1$ stands for the presence of an edge between the i^{th} and j^{th} nodes, and $A_{ij} = 0$ otherwise. See Section 1.2 for a formal presentation of this model.

1.1.2 Contribution

1.1.2.1 Complexity index

Our first objective is the definition of a complexity index in the W-random graph model. As a natural candidate, one might think of the dimension of the latent space, like d if $\Omega = [0,1]^d$. However, this index is inadequate because of a major identifiability issue. Indeed, it is known that [see Lovász, 2012] the attribute space Ω is not identifiable from the observed adjacency matrix A. Even worse, it has been shown that all W-random graph distributions can be represented on the specific space $\Omega = [0,1]$ [Lovász, 2012]. It is therefore pointless to think about the graph complexity purely in terms of the latent space. Likewise, the regularity of the link function (like α if W is α-Hölder) is not suited due to the non-identifiability of W.

These issues motivate the introduction of a more abstract index. Given a graphon (Ω, μ, W), we endow the latent space Ω with the so-called *neighborhood distance*

$$r_W(\omega, \omega') = \left(\int_\Omega \left| W(\omega, \omega'') - W(\omega', \omega'') \right|^2 \mu(d\omega'') \right)^{1/2}. \tag{1.1}$$

From the above description of a W-random graph, we can see that the quantity $r_W(\omega_i, \omega_j)$ measures the propensity of the nodes i and j to be connected with similar nodes. Our complexity index is then defined as the covering number and the Minkowski dimension of a

purified version of the (pseudo-) metric space (Ω, r_W). The purification process is detailed in section 1.3.1. Recall the definitions of these two standard measures for metric spaces: the ϵ-covering number $N_\Omega^{(c)}(\epsilon)$ is the minimal number of balls of radius ϵ required to entirely cover the (pseudo-) metric space (Ω, r_W). And the Minkowski dimension is the following limit on the covering number

$$dim\,\Omega := \lim_{\epsilon \to 0} \frac{\log\, N_\Omega^{(c)}(\epsilon)}{-\log \epsilon} \tag{1.2}$$

when the limit exists. In particular, the Minkowski dimension does not have to be an integer.

Although none of the three parameters Ω, μ and W are identifiable in the W-ranom graph model, we prove that the covering number and the Minkowski dimension of a purified version of (Ω, r_W) are identifiable.

We also illustrate that this notion of complexity is sound on classic examples of random graphs. Specifically, we show that $N_\Omega^{(c)}(\epsilon)$ is equal to the number of well-spaced communities in the stochastic block model; that $dim\,\Omega$ matches the dimension of the Euclidean space in some random geometric graphs; and that $dim\,\Omega$ is equal to the regularity of the link function in some Hölder graphon models. See Section 1.3.2 for details.

In addition to all applications listed in the introduction, these complexity indices may also be useful to adjust analytical methods to particular networks, for example, when estimating the link function W (see section 1.1.3 and 1.3.2 for related comments) or in learning representation where the goal is to find an informative metric space to place/represent the nodes of the network [Hoff et al., 2002, Perozzi et al., 2014, Grover and Leskovec, 2016].

1.1.2.2 Statistical estimation

From the observed adjacency matrix A of a W-random graph, we estimate the neighborhood distance (1.1) on the sampled points $\omega_1, \dots, \omega_n$. The corresponding distance estimator $\widehat{r}$ is defined in Section 1.4.1. We show universal non-asymptotic bounds for its risk (Theorem 1.1.1). Let $\omega_{m(i)} \in \{\omega_1, \dots, \omega_n\} \setminus \{\omega_i\}$ denote a nearest neighbor of ω_i with respect to the distance r_W.

Theorem 1.1.1 *Consider the distance estimator $\widehat{r}$, defined in Section 1.4.1. Then, for any graphon (Ω, μ, W), we have*
$\forall i, j \in [n]$,

$$\left| r_W^2(\omega_i, \omega_j) - \widehat{r}^2(i,j) \right| \lesssim r_W(\omega_j, \omega_{m(j)}) + r_W(\omega_i, \omega_{m(i)}) + \sqrt{\log(n)/n}$$

with probability at least $1 - 2/n$.

In the upper bound, there is a bias term $r_W(\omega_j, \omega_{m(j)})$ which is the distance between the sampled point ω_j and its nearest neighbor $\omega_{m(i)}$ (w.r.t. the neighborhood distance). This bias depends on the form of the underlying graphon (Ω, μ, W), for example, it is equal to zero w.h.p. in the stochastic block model (i.e., when the link function W is piecewise constant on $\Omega = [0,1]$). We also derive a minimax lower bound that matches the upper bound of Theorem 1.1.1. See Section 1.4.2 for details on the distance estimation.

Based on the estimated distances $\widehat{r}(i,j)$, we estimate the covering number $N_\Omega^{(c)}(\epsilon)$ by plug-in and provide universal non-asymptotic error bounds for this estimator. See Section 1.4.3 for details. Our results on the distance and covering number are therefore valid for all graphons, unlike most results in the graphon literature.

Combining the above covering number estimator $\widehat{N}_\Omega^{(c)}$ with formula (1.2), we derive an estimator of the Minkowski dimension

$$\widehat{dim}_D := \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon_D)}{-\log \epsilon_D}$$

which satisfies a high probability convergence rate (Theorem 1.1.2). For this result, we assume the Mikowski dimension is upper bounded by some constant D and use a particular radius ϵ_D defined in Section 1.4.4. We also make some mild assumptions on the graphon geometry, which are inspired by the problem of estimation of manifold dimension (see section 1.1.3 for this related literature). Besides, we show that this set of assumptions is minimal, in the sense that, if any of these assumptions is removed, all dimension estimators make an estimation error of the order 1.

Theorem 1.1.2 *Under some mild assumptions, defined in Section 1.4.4, the following holds. If $dim\,\Omega$ is any real in $[0, D]$, then*

$$\left| \widehat{dim}_D - dim\,\Omega \right| \lesssim \frac{1}{\log n}$$

with probability at least $1 - C'/n$ for some constant C' independent of n.

Finally, we prove that the upper bound $\log^{-1} n$ is optimal, which means that no estimator can improve on this error. For detailed results, see Section 1.4.4.

As extensions, we show that the above results also cover the important setting of sparse networks, which has been considered several times in the literature [see Bickel et al., 2011, Wolfe and Olhede, 2013, Klopp et al., 2017, Xu et al., 2014]. In addition, we describe a polynomial-time algorithm to approximate the covering number estimator; we do so by using a classic greedy algorithm that is known to satisfy some theoretical guarantees. See Section 1.6 for these two extensions.

Finally, we test if the packing number (of a purified version of (Ω, r_W)) is smaller than K, with a specific care for controlling the type I error probability uniformly over all graphons. We prove this error is smaller than $2/n$ for any graphon. For technical reasons detailed in Section 1.5, we use here the packing number instead of the covering number, which are essentially the same measures (see Appendix 1.A for a reminder about these usual measures for metric spaces).

1.1.3 Connection with the literature

1.1.3.1 W-random graph model

The most simple random graph is the Erdös-Rényi model where each edge has the same probability p of being present, independently of the other edges. The study of this generative

model has been impressively fruitful in mathematics [Bollobás, 1998] but does not replicate even the simplest properties of real-world networks. Hence, the assumption of a constant connection probability p has been relaxed in the celebrated stochastic block model [Holland et al., 1983] where the connection probabilities may vary with the community membership of each node. Although this model has attracted a lot of attention [Abbe, 2017], it fails to catch some subtle aspects of very large graphs. Such modeling issues have led to a non-parametric view of network analysis [Bickel and Chen, 2009], in particular the introduction of the W-random graph model [Diaconis and Janson, 2007].

The universality of the W-random graphs has two parts. On the one hand, the graphon (Ω, μ, W) plays a key role in network analysis as a powerful representation of many graph properties. Indeed, it has been shown that many sequences of growing graphs can be represented by graphons. For details, see the theory of graph limits introduced by Lovász and Szegedy [2006] or the comprehensive monograph by Lovász [2012]. On the other hand, the W-random graph model is connected with the theory of exchangeable random graphs. In fact, every distribution on random graphs that is invariant by permutation of nodes can be expressed with W-random graphs [Diaconis and Janson, 2007, Aldous, 1981, Kallenberg, 1989]. Thus, the W-random graphs encompass many random graph models, including stochastic block models, random geometric graphs [Penrose et al., 2003] and random dot product graphs [Tang et al., 2013, Athreya et al., 2017].

1.1.3.2 Graphon estimation

There has been much interest in the recovery of the function W (or the matrix of probabilities $[W(\omega_i, \omega_j)]_{i,j\leq n}$) on the specific space $\Omega = [0, 1]$. Usually, authors assume the graphon has some regularity (e.g. W is Hölder continuous on $[0, 1]$) and then use an approximation by SBM, which can be seen as an approximation by constant piecewise functions of W [Borgs et al., 2015, Wolfe and Olhede, 2013, Gao et al., 2015, Klopp et al., 2017, Latouche and Robin, 2016]. We also mention an alternative approach based on neighborhood-smoothing [Zhang et al., 2015, Xu et al., 2014]. In comparison with this literature, our objective is less ambitious since we only estimate a feature of the graph (its complexity). In return, we carry out a general analysis and do not assume any smoothness condition on $\Omega = [0, 1]$. Indeed, our results on the neighborhood distance and covering number estimations are valid for all graphons. For the dimension, we make mild assumptions which are similar to those in the "intrinsic dimension estimation" literature (see subsection 1.1.3.3 for a brief description of this related problem).

In the problem of estimation of W, the latent space $[0, 1]$ is sometimes considered instead of Ω. This choice is not restrictive (if no assumption is made on the function W on $[0, 1]$) because both settings generate the same W-random graph distributions [Lovász, 2012]. However, the restricted setting $[0, 1]$ is not always convenient to work with, whereas the general setting Ω leads to simpler and cleaner situations [Lovász, 2012]. Indeed, many random graph distributions are naturally represented on Ω so that their properties are easy to interpret. See Section 1.3.2 for illustrative examples.

The l_2-neighborhood distance (1.1) is a variant of the l_1-neighborhood distance introduced by Lovász [2012]. This variant has been leveraged several times for the estimation of $[W(\omega_i, \omega_j)]_{i,j\leq n}$ [Zhang et al., 2015, Xu et al., 2014] where the authors use it as a criterion to select neighborhoods

of nodes. Here, our estimator of the l_2-neighborhood distance is inspired by the work of Zhang et al. [2015], as will be discussed later.

1.1.3.3 Intrinsic dimension estimation

There is a considerable body of literature on the estimation of intrinsic dimension of a manifold [Kim et al., 2016, Kégl, 2003, Koltchinskii, 2000, Levina and Bickel, 2005]. In the simplest setting, points are sampled on a manifold of $\mathbb{R}^m$ whose dimension is an integer, and the objective is to recover this dimension from the sample. In contrast, here we do not assume the dimension is an integer, we do not observe the n sampled points $\omega_1, \ldots, \omega_n$, and we are not in the Euclidean metric space $\mathbb{R}^m$. Indeed, the neighborhood distance r_W is unknown, and our only observation is the connections of the graph.

Outline of the paper. Section 1.2 gives a formal presentation of the problem. Section 1.3 presents the complexity index and some illustrations. In Section 1.4, we focus on statistical estimation (distance, covering number, dimension). In Section 1.5, we test the graph complexity. In Section 1.6, we provide two extensions (estimation on sparse graphs, and a polynomial-time algorithm). Proofs are deferred to the appendix.

Notation. we write $a \lesssim b$, if there exists a constant C such that $a \leq Cb$; and note $a \asymp b$, if there exist two constants c, c' such that $ca \leq b \leq c'a$. We denote by $a \vee b$ (respectively $a \wedge b$) the maximum (resp. minimum) between a and b; by $[a]_+$ the maximum between 0 and a; by $[n]$ the set $\{1, \ldots, n\}$; by $B(x, \epsilon)$ a ball of radius ϵ and center x. We note $1_{\mathcal{E}}$ the indicator function corresponding to any event $\mathcal{E}$. We write "a.e." for "almost everywhere"; and "w.r.t." for "with respect to"; and "w.h.p." for "with high probability", which means that the probability converges to 1 as the number of graph nodes tends to infinity.

1.2 Model

1.2.1 Setting

For a set of vertices $V = \{1, \ldots, n\}$, a *W-random graph* $G = (V, E)$ is generated as follows. Let (Ω, μ, W) be an unknown triplet of parameters, which is composed of a measurable set Ω, a probability measure μ on Ω, and a symmetric (measurable) function $W : \Omega \times \Omega \to [0, 1]$. For each node $i \in V$, an unknown attribute $\omega_i \in \Omega$ is drawn in an i.i.d. manner from the distribution μ. Conditionally to the attributes $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$, an edge connects two vertices i and j, independently of the other edges, with probability $W(\omega_i, \omega_j)$.

$$\mathbb{P}\big((i, j) \in E \,\big|\, \boldsymbol{\omega}\big) = W(\omega_i, \omega_j) \tag{1.3}$$

Our data are a single observation of the W-random graph. Formally, it is an adjacency matrix $A = [A_{ij}]_{i,j \leq n}$ defined by $A_{ij} = 1$ if $(i, j) \in E$, and 0 otherwise. This symmetric binary matrix with zero-entries on the diagonal represents an undirected, unweighted graph with no self edges. The distribution of A is called the data distribution and is denoted by $\mathbb{P}_{(\Omega,\mu,W)}$. The set of graphons is written $\mathcal{W}$.

1.2.2 Non-identifiability and equivalence class of graphons

From the observation A, the function W is not identifiable. Indeed, for any measure-preserving bijection $\phi : \Omega \to \Omega$, we can observe that the map $W^\phi(x, y) = W(\phi(x), \phi(y))$ leaves the data distribution unchanged, i.e.:

$$\mathbb{P}_{(\Omega,\mu,W)} = \mathbb{P}_{(\Omega,\mu,W^\phi)}.$$

In fact, even the latent space Ω is not identifiable. The full picture is described by Lovász [2012, chap.10]:

> *Two graphons (Ω, μ, W) and (Ω', μ', W') parametrize the same data distributions for all n, if and only if, there exist some measure-preserving maps $\phi : [0, 1] \to \Omega$ and $\psi : [0, 1] \to \Omega'$ such that $W^\phi(x, y) = W'^\psi(x, y)$ a.e.*

where $[0, 1]$ is the probability space endowed with the uniform measure. This characterization will be useful to prove the identifiability of our complexity index. For clarity of this future discussion, we consider the corresponding quotient space $\mathcal{W}/\sim$, which is the set of equivalence classes of graphons leading to the same data distributions.

1.3 Complexity index

Given a graphon (Ω, μ, W), we endow the latent space Ω with the *neighborhood distance*

$$r_W(\omega, \omega') = \left(\int_\Omega \left| W(\omega, \omega'') - W(\omega', \omega'') \right|^2 \mu(d\omega'') \right)^{1/2} \tag{1.4}$$

which is the l_2-norm $||W(\omega, .) - W(\omega', .)||_{2,\mu}$ between the slices of the function W in ω and ω'. Then, we measure the complexity of the pseudo-metric space (Ω, r_W) in a classic way, using its covering number $N_\Omega^{(c)}(\epsilon)$ and its Minkowski dimension:

$$dim\,\Omega := \lim_{\epsilon \to 0} \frac{\log\, N_\Omega^{(c)}(\epsilon)}{-\log \epsilon} \tag{1.5}$$

when the limit exists. See appendix 1.A for additional information about these two standard measures of metric spaces.

Unfortunately, the covering number and the Minkowski dimension of a graphon are not identifiable from the data distribution $\mathbb{P}_{(\Omega,\mu,W)}$. Indeed, they are not robust to changes of the graphon on null-sets, whereas such changes leave the data distribution unaltered (a null-set is a set of zero measure in the probability space (Ω, μ)). This fact is illustrated in the following example where two equivalent graphons (i.e. leading to the same data distributions) have two different Minkowski dimensions. As we can see, this problem is due to the presence of a "big" null-set in Ω.

Example. Let $\Omega := \{2\}$ and $\Omega' := \{2\} \sqcup [0, 1]$ be two latent spaces endowed with a common probability distribution μ such that $\mu[\{2\}] = 1$. Let W' be a function defined on $\Omega' \times \Omega'$ such that $W'(x', y') = (x' + y')/3$ for $x', y' \in [0, 1]$. Let W be any measurable function on $\Omega \times \Omega$ such that $W(2, 2) = W'(2, 2)$. Then, the two graphons (Ω, μ, W), (Ω', μ, W') are equivalent, and yet they have two different Minkowski dimensions: $dim\,\Omega = 0$ since $r_W = 0$ on Ω, while $dim\,\Omega' = 1$ since $r_{W'}(x', z') = |x' - z'|/3$ for $x', z' \in [0, 1]$. □

1.3.1 Purification process for identifiability

To define an identifiable index of complexity, we need to take care of "big" null-sets (seen in the above example). Usually, these pathological sets are not present in standard representations (Ω, μ, W) and even useless in terms of modeling. Thus, we get rid of them; we do so by using a general remedy, called pure graphon.

Definition [Lovász, 2012, chap.13] *A graphon (Ω, μ, W) is called pure if (Ω, r_W) is a complete separable metric space and the probability measure has full support (that is, every ball of non-zero radius has positive measure). Besides, there is a pure graphon in each equivalence class of graphons.*

For illustrative examples of pure graphons, see Section 1.3.2. There is no "big" null-set in pure graphons (since their measure μ has full support by definition) and the complexity index takes the same value on the pure graphons of a same equivalence class of $\mathcal{W}/\sim$ (Lemma 1.3.1).

Lemma 1.3.1 *If two pure graphons are equivalent, then their covering numbers are equal.*

The proof of Lemma 1.3.1 is written in Appendix 1.C.2. Lemma 1.3.1 directly implies that the Minkowski dimension takes the same value for two equivalent pure graphons. We now define the complexity of a W-random graph distribution $\mathbb{P}_{(\Omega,\mu,W)}$ as the covering number and the Minkowski dimension of any pure graphon from the corresponding equivalence class. According to the above lemma, these indices are therefore identifiable from $\mathbb{P}_{(\Omega,\mu,W)}$. From now on, we can work exclusively with pure graphons without the loss of generality, since there are pure graphons in each equivalence class of $\mathcal{W}/\sim$. In the remaining of the subsection, we describe two consequences of working with pure graphons.

The metric properties are preserved between equivalent pure graphons (Lemma 1.3.2).

Lemma 1.3.2 *Let (Ω, μ, W) and (Ω', μ', W') be two pure graphons, endowed with their respective neighborhood distances r_W and $r_{W'}$. If the two graphons are in a same equivalence class of $\mathcal{W}/\sim$, then for some bijective measure-preserving map $\phi : \Omega' \to \Omega$, we have*

$$r_{W'}(x, y) = r_W(\phi(x), \phi(y)) \quad \textit{almost surely on } \Omega' \times \Omega'.$$

Lemma 1.3.2 states that the metric spaces (Ω, r_W) and $(\Omega', r_{W'})$ are isometric up to a null-set, it is therefore not surprising that they share the same covering number (Lemma 1.3.1). The proof of lemma 1.3.2 is written in Appendix 1.C.1. Note that Lemma 1.3.2 ensures that the future distance estimation is a well-posed problem.

Another consequence of working with pure graphon is that the sample $\omega_1, \ldots, \omega_n$ is asymptotically dense in Ω. Lemma 1.3.3 is proved in Appendix 1.C.3.

Lemma 1.3.3 *For a pure graphon (Ω, μ, W) such that $N_\Omega^{(c)}(\epsilon) < \infty$ for all $\epsilon > 0$, the sample $\omega_1, \ldots, \omega_n$ is asymptotically dense in the metric space (Ω, r_W). That is, for all radii $\epsilon > 0$, the event*

$$\mathcal{E}(\epsilon) = \{\text{each ball of radius } \epsilon \text{ in } (\Omega, r_W) \text{ contains at least a sampled point } \omega_i\}$$

holds with a probability tending to one as $n \to \infty$.

1.3.2 Illustrative examples

We exemplify the complexity index with instances of W-random graphs that are often considered in the literature: a stochastic block model [Holland et al., 1983, Abbe, 2017], a random Hölder graph [Gao et al., 2015, Zhang et al., 2015] and a random geometric graph [Penrose et al., 2003, Arias-Castro et al., 2018, De Castro et al., 2017, Bubeck et al., 2016].

Stochastic Block Model. It produces a structure of community dividing the node set into K subsets of nodes which share a same pattern of connection. More precisely, the edges are independently sampled from each others, and the probability of an edge between two nodes only depends on their community membership. The SBM with K communities can be written in the framework of the W-random graph model, by setting $\Omega = \{c_1, \dots, c_K\}$, so that each node belongs to one of the K communities c_i, and connects to each other with probability $W(c_i, c_j)$. A natural notion of complexity for SBM is the number K of communities, which coincides with the ϵ-covering number of $\{c_1, \dots, c_K\}$ for small radii ϵ.

Approximation by SBM. In the estimation of W based on the classic approximation by SBM [Gao et al., 2015, Klopp et al., 2017], the right number of communities can be selected using the covering number. Indeed, Proposition 1.3.4 states that, for any graphon (Ω, μ, W), the function W can be "$O(\epsilon)$-approximated" in l_2-norm by an SBM with at most $N_\Omega^{(c)}(\epsilon)$ communities. The proof is written in Appendix 1.B.1.

Proposition 1.3.4 *Consider any graphon (Ω, μ, W) and its ϵ-covering number $N_\Omega^{(c)}(\epsilon)$, defined in Section 1.3. There exists a graphon $(\Omega, \mu, \overline{W})$ equivalent to an SBM with $N_\Omega^{(c)}(\epsilon)$ communities, such that,*

$$\int_{\Omega^2} (W(\omega, \omega') - \overline{W}(\omega, \omega'))^2 \mu(d\omega)\mu(d\omega') \leq (4\epsilon)^2.$$

Random Hölder graph. Let $\Omega = [0,1]^d$ be endowed with the uniform measure, and W fulfill a double Hölder condition:

$$m||\omega' - \omega||_2^\alpha \leq |W(\omega', \omega'') - W(\omega, \omega'')| \leq M||\omega' - \omega||_2^\alpha \tag{1.6}$$

for some Hölder exponent $\alpha > 0$ (and some constants $m, M > 0$). This means that each node has its specific attribute of d variables, and connects to another node with a probability that smoothly depends on the node attributes. A natural notion of complexity for this graph distribution should increase with the number d of variables, and decrease with the level α of smoothness. This intuitive notion is matched by the Minkowski dimension, which is equal to d/α. See Appendix 1.A for details.

Random geometric graph. It generates simple spatial networks placing nodes in a Euclidean metric space and connecting two nodes if their Euclidean distance is small. Let $\Omega = [0,1]^d$ be endowed with the uniform measure and the indicator function $W(\omega, \omega') = \mathbb{I}_{||\omega - \omega'||_2 \leq \delta}$ for some constant $\delta > 0$. Appendix 1.A shows that $dim\,\Omega = 2d$. Thus, the Minkowski dimension matches the Euclidean dimension of the latent space, up to a factor 2.

1.4 Estimation of the complexity index

Given a pure graphon (Ω, μ, W), assume a W-random graph is generated from the probability distribution $\mathbb{P}_{(\Omega,\mu,W)}$ defined in Section 1.2.1. From a single observation of the adjacency matrix A of this graph, we want to estimate the complexity index (introduced in Section 1.3.1). In particular, the underlying graphon (Ω, μ, W) is unknown, and the sampled points $\omega_1, \dots, \omega_n$ are not observed.

This section is organized in the following manner. We first estimate the neighborhood distance (1.4) on the sampled points $\omega_1, \dots, \omega_n$. Based on these estimated distances, we then estimate the ϵ-covering number of $(\{\omega_1, \dots, \omega_n\}, r_W)$ by plug-in. Denote by $\widehat{N}_\Omega^{(c)}(\epsilon)$ this estimator. We finally estimate the Minkowski dimension using $-\log \widehat{N}_\Omega^{(c)}(\epsilon)/\log \epsilon$ at a well chosen radius ϵ.

1.4.1 Distance-estimator

Let us explain the construction of the distance estimator. The l_2-neighborhood distance is naturally associated with a structure of inner product. Given some square-integrable functions f and g on Ω, we write their inner product $\langle f, g\rangle := \int_\Omega f(z)g(z)\mu(dz)$. Let $W(\omega_i, .)$ denote the function $x \mapsto W(\omega_i, x)$, then the neighborhood distance admits the following decomposition

$$r_W^2(\omega_i, \omega_j) = \langle W(\omega_i, .), W(\omega_i, .)\rangle + \langle W(\omega_j, .), W(\omega_j, .)\rangle - 2\langle W(\omega_i, .), W(\omega_j, .)\rangle. \tag{1.7}$$

We estimate separately the crossed term and the two quadratic terms of (1.7).

Note A_i the i^{th} row vector of the adjacency matrix A, and $\langle A_i, A_j\rangle_n = \sum_{i=1}^n A_{ik}A_{jk}/n$ the inner product between two such rows. Given ω_i, ω_j, we observe that $\langle A_i, A_j\rangle_n$ is (almost) a sum of i.i.d. random variables (up to a duplicated entry because of the symmetry of the adjacency matrix A). Indeed, the $n-2$ random variables $\{A_{ik}A_{jk} : k \in [n] \text{ and } k \neq i, j\}$ are independent with the same mean conditionally to ω_i, ω_j :

$$\mathbb{E}\left[A_{ik}A_{jk}|\omega_i, \omega_j\right] = \langle W(\omega_i, .), W(\omega_j, .)\rangle$$

where the mean $\mathbb{E}$ is taken over the data distribution $\mathbb{P}_{(\Omega,\mu,W)}$. It is therefore possible to use Hoeffding's inequality to prove that $|\langle A_i, A_j\rangle_n - \langle W(\omega_i, .), W(\omega_j, .)\rangle| \lesssim \sqrt{\log n/n}$ w.h.p. (see Proposition 1.D.1 in Appendix 1.D.1). Thus, the inner product between two *different* rows is a consistent estimator of the crossed term $\langle W(\omega_i, .), W(\omega_j, .)\rangle$ in (1.7).

To estimate the remaining quadratic term $\langle W(\omega_i, .), W(\omega_i, .)\rangle$ in (1.7), we cannot proceed in the same way since $\frac{1}{n}\langle A_i, A_i\rangle$ is an inconsistent estimator of $\langle W(\omega_i, .), W(\omega_i, .)\rangle$; indeed, we have

$$\mathbb{E}\big[A_{ik}A_{ik}|\omega_i\big] = \mathbb{E}\big[A_{ik}|\omega_i\big] = \langle W(\omega_i, .), 1\rangle \neq \langle W(\omega_i, .), W(\omega_i, .)\rangle.$$

To work around this issue, we simply approximate the quadratic term by a crossed term to be back to the previous case. Specifically, the approximation consists in replacing a sampled point by its nearest neighbor as follows: let $\omega_{m(i)} \in \{\omega_1, \dots, \omega_n\}$ denote a nearest neighbor of ω_i according to the distance r_W, that is $m(i) \in \operatorname{argmin}_{t:\, t\neq i}\ r_W(\omega_i, \omega_t)$, then we have the

following approximation:

$$\begin{aligned}|\langle W(\omega_i,.),W(\omega_i,.)\rangle - \langle W(\omega_i,.),W(\omega_{m(i)},.)\rangle| &= |\langle W(\omega_i,.),W(\omega_i,.)-W(\omega_{m(i)},.)\rangle| \\ &\leq r_W(\omega_i,\omega_{m(i)}) \end{aligned} \tag{1.8}$$

using Cauchy-Schwarz inequality. Thus, the nearest neighbor approximation (1.8) entails a bias in our estimation procedure, which is equal to the distance between ω_i and its nearest neighbor $\omega_{m(i)}$.

Since the index $m(i)$ is unknown, we define an index estimator $\widehat{m}(i)$ such that $\omega_{\widehat{m}(i)}$ is hopefully close to ω_i according to r_W, and then we use $\langle A_i, A_{\widehat{m}(i)}\rangle_n$ to estimate the quadratic term. Formally, $\widehat{m}(i)$ is a minimizer of the distance function $j \mapsto \widehat{f}(i,j)$ defined by

$$\widehat{f}(i,j) = \max_{k:\, k\neq i,j} |\langle A_k, A_i - A_j\rangle_n| \tag{1.9}$$

where $\widehat{f}(i,j)$ represents a proxy for the distance between the i^{th} and j^{th} rows of the adjacency matrix, which is enough to define the index estimator

$$\widehat{m}(i) = \operatorname*{argmin}_{j:\, j\neq i} \widehat{f}(i,j). \tag{1.10}$$

Note that $\widehat{f}(i,j)$ is small in expectation if ω_i and ω_j are close according to the neighborhood distance; indeed, $\mathbb{E}\big[\,\widehat{f}(i,j)|\omega_i,\omega_j,\omega_k\big] = \max_{k\neq i,j} |\langle W(\omega_i,.)-W(\omega_j,.),W(\omega_k,.)\rangle| \leq r_W(\omega_i,\omega_j)$ using Cauchy-Schwarz inequality.

Putting together the estimators of the crossed term and the two quadratic terms, we get the following estimator of the square distance $r_W^2(\omega_i,\omega_j)$:

$$\widehat{r}^2(i,j) = \langle A_i, A_{\widehat{m}(i)}\rangle_n + \langle A_j, A_{\widehat{m}(j)}\rangle_n - 2\,\langle A_i, A_j\rangle_n \tag{1.11}$$

for all $i,j \in [n]$, where $\widehat{m}(i)$ is given by (1.10).

Remark: The distance-estimator (1.11) is inspired by the work of Zhang et al. [2015], in which the authors want to recover the expectation of the adjacency matrix A, based on neighborhood smoothing. They rely on the proxy (1.9) to select neighborhood of points with respect to the neighborhood distance. Restricting themselves on graphons of the form $([0,1],\lambda,W)$ with λ the uniform measure and W a piecewise Lipschitz function, they derive risk bounds for the estimation of W. In contrast, here we do not make any assumption on the graphon, and our objective is to provide an estimator of the neighborhood distance per se.

1.4.2 Consistency of the distance-estimator

The statistical recovery of the set of distances $\{r_W(\omega_i,\omega_j) : i,j \in [n]\}$ is a well-posed problem, since the neighborhood distance is invariant on each equivalence class of graphons (Lemma 1.3.2). Theorem 1.4.1 gives non-asymptotic error bounds for the distance-estimator (1.11). The proof is written in Appendix 1.D.1.

Theorem 1.4.1 *Given any (pure) graphon (Ω, μ, W), consider the data distribution $\mathbb{P}_{(\Omega,\mu,W)}$ defined in model (1.3). For all $1 \leq i \leq n$, let $\omega_{m(i)} \in \{\omega_1, \ldots, \omega_n\} \setminus \{\omega_i\}$ denote a nearest neighbor of ω_i according to the distance r_W. Then, for the distance-estimator (1.11), the event*

$$\mathcal{E}_{dist} = \Big\{ \forall i, j \in [n] : \big| r_W^2(\omega_i, \omega_j) - \widehat{r}^2(i,j) \big| \\ \leq 3 r_W(\omega_j, \omega_{m(j)}) + 3 r_W(\omega_i, \omega_{m(i)}) + 36\sqrt{\log(n)/n} \Big\}$$

holds with probability $\mathbb{P}_{(\Omega,\mu,W)}[\mathcal{E}_{dist}] \geq 1 - \frac{2}{n}$.

Theorem 1.4.1 implies that the distance-estimator (1.11) is a consistent estimator of the neighborhood distance (1.4), provided that the ϵ-covering number is finite for all radii $\epsilon > 0$. Indeed, for a finite covering number, Lemma 1.3.3 ensures that the sample $\omega_1, \ldots, \omega_n$ is asymptotically dense in (Ω, r_W), which implies that the bias $r_W(\omega_i, \omega_{m(i)})$ is convergent in probability to zero as n grows to infinity.

Let us describe the upper bound of Theorem 1.4.1. On the one hand, there is a fluctuation term $\sqrt{\log(n)/n}$ that corresponds to the convergence property of the inner products between rows of A, i.e.: $|\langle A_i, A_j \rangle_n - \langle W(\omega_i, .), W(\omega_j, .) \rangle| \lesssim \sqrt{\log n/n}$ w.h.p. for $i \neq j$. On the other hand, there is a bias term $r_W(\omega_i, \omega_{m(i)})$ that results from the nearest neighbor approximation (1.8). Its value depends on the graphon regularity. For instance, in the SBM example of Section 1.3.2, the bias term $r_W(\omega_i, \omega_{m(i)})$ is equal to zero w.h.p. (indeed, ω_i and its nearest neighbor $\omega_{m(i)}$ are in the same community w.h.p., and thus separated by a distance zero w.r.t. r_W). In the random Hölder graph example, the bias term is of the order of $(\log(n)/n)^{\alpha/d}$ w.h.p..

We now discuss the optimality of the upper bound of Theorem 1.4.1. As the event $\mathcal{E}_{dist}$ is a uniform error bound on *square* distances, we may expect the bias term to be a *square* distance too (instead of a distance as in Theorem 1.4.1). Indeed, this expected bias $r_W^2(\omega_i, \omega_{m(i)})$ would improve on $r_W(\omega_i, \omega_{m(i)})$ since the neighborhood distance r_W is always smaller than 1 by definition. Then, one may wonder whether such an improvement is possible. It turns out that even replacing the bias $r_W(\omega_i, \omega_{m(i)})$ by $r_W^{1+\gamma}(\omega_i, \omega_{m(i)})$ for some $\gamma > 0$ is impossible. Indeed, no estimator $\widehat{d}$ simultaneously satisfies the following inequalities
$\forall i, j \in [n]$:

$$\frac{\left| r_W^2(\omega_i, \omega_j) - \widehat{d}^2(i,j) \right|}{r_W^{1+\gamma}(\omega_j, \omega_{m(j)}) + r_W^{1+\gamma}(\omega_i, \omega_{m(i)}) + \sqrt{\log(n)/n}} \leq C \tag{1.12}$$

w.h.p. for all graphons (Ω, μ, W) and some numerical constant C. Specifically, Theorem 1.4.2 states that, for a sequence of graphons $(\Omega, \mu, W_n)_{1 \leq n}$, the (uniform) bound (1.12) cannot be achieved by any estimator $\widehat{d}$ [1].

Theorem 1.4.2 *There exist a sequence of graphons $(\Omega, \mu, W_n)_{n \in \mathbb{N}}$ and some numerical constants $p > 0$ and $c > 0$, such that the following holds for any estimator $\widehat{d}$ and any permutation σ of*

[1] defined as a function of the adjacency matrix $A \in \{0,1\}^{n \times n}$.

the n indices. With a probability larger than p, the lower bound

$$\frac{\left|r^2_{W_n}(\omega_i,\omega_j) - \widehat{d}^2(\sigma(i),\sigma(j))\right|}{r^{1+\gamma}_{W_n}(\omega_j,\omega_{m(j)}) + r^{1+\gamma}_{W_n}(\omega_i,\omega_{m(i)}) + \sqrt{\log(n)/n}} \gtrsim \left(\sqrt{\frac{n}{\log n}}\right)^{\gamma/(1+\gamma)}$$

is satisfied for (at least) $c\,n$ different pairs (i,j).

Hence, the uniform bound (1.12) cannot be achieved, implying that the upper bound of Theorem 1.4.1 is optimal. Note that the data distribution is invariant by relabeling of the nodes, and consequently we study the problem of estimating a *set* of distances (i.e., regardless of their labels $i \in \{1,\dots,n\}$). Accordingly, the above lower bound holds for any permutation σ of the n indices $\{1,\dots,n\}$. For a proof of Theorem 1.4.2, see Appendix 1.D.2.

1.4.3 Consistency of the covering number estimator

We have defined the ϵ-covering number estimator $\widehat{N}^{(c)}_{\Omega}(\epsilon)$ as the covering number of the set $\{1,\dots,n\}$ w.r.t. the distance-estimator $\widehat{r}$. Consider e_{sup} the supremum of the errors of $\widehat{r}$:

$$e_{sup} := \sup_{i,j\in[n]} |r_W(\omega_i,\omega_j) - \widehat{r}(i,j)|\,.$$

Then, the covering number estimator is linked with the true covering number of $\{\omega_1,\dots,\omega_n\}$ by the following inequalities

$$\forall \epsilon > e_{sup}, \qquad N^{(c)}_{\omega_1,\dots,\omega_n}(\epsilon + e_{sup}) \le \widehat{N}^{(c)}_{\Omega}(\epsilon) \le N^{(c)}_{\omega_1,\dots,\omega_n}(\epsilon - e_{sup})\,.$$

To compare the covering numbers of $\{\omega_1,\dots,\omega_n\}$ and Ω, we need to measure the difference between the sample $\omega_1,\dots,\omega_n$ and the space Ω. We do so by introducing the sampling error s_ω defined as

$$s_\omega = \sup_{\omega\in\Omega} \inf_{i\in\{1,\dots,n\}} r_W(\omega,\omega_i) \tag{1.13}$$

which is the greatest distance that separates a point of Ω from the set $\{\omega_1,\dots,\omega_n\}$. Thus, the covering numbers (w.r.t. the true distance r_W) of $\omega_1,\dots,\omega_n$ and Ω are linked by the following inequalities

$$\forall \epsilon > s_\omega, \qquad N^{(c)}_{\Omega}(\epsilon + s_\omega) \le N^{(c)}_{\omega_1,\dots,\omega_n}(\epsilon) \le N^{(c)}_{\Omega}(\epsilon - s_\omega)\,.$$

Finally, for

$$b^2_{sup} := 6 \sup_{i\in[n]} r_W(\omega_i,\omega_{m(i)}) + 36\sqrt{\log(n)/n}, \tag{1.14}$$

Theorem 1.4.1 ensures that $e_{sup} \le b_{sup}$ with probability at least $1 - 2/n$. From the above displays, we obtain the following proposition.

Proposition 1.4.3 *Given any (pure) graphon (Ω,μ,W), consider the data distribution $\mathbb{P}_{(\Omega,\mu,W)}$ defined in model (1.3). Let b_{sup} and s_ω be the distance error bound (1.14) and the sampling error (1.13). Then, the estimator $\widehat{N}^{(c)}_{\Omega}$ satisfies the following non-asymptotic bounds*

$\forall \epsilon > b_{sup} + s_\omega$,

$$N_\Omega^{(c)}(\epsilon + b_{sup} + s_\omega) \leq \widehat{N}_\Omega^{(c)}(\epsilon) \leq N_\Omega^{(c)}(\epsilon - b_{sup} - s_\omega) \tag{1.15}$$

with probability at least $1 - \frac{2}{n}$ *according to the distribution* $\mathbb{P}_{(\Omega,\mu,W)}$.

As a result, we have a consistent estimation of the ϵ-covering number for almost every ϵ, provided that the covering number is finite for all radii. Indeed, if $N_\Omega^{(c)}(\epsilon) < \infty$ for all $\epsilon > 0$, then the sample $\omega_1, \dots \omega_n$ is asymptotically dense in (Ω, r_W) by Lemma 1.3.3, which implies that b_{sup} and s_ω converge in probability to zero; Then, taking the limit $n \to \infty$ in (1.15), one has the convergence in probability of $\widehat{N}_\Omega^{(c)}$ towards $N_\Omega^{(c)}(\epsilon)$, for all ϵ where the step function $\epsilon \mapsto N_\Omega^{(c)}(\epsilon)$ is continuous (i.e., for almost every ϵ).

1.4.4 Consistency of the dimension estimator

We estimate the Minkowski dimension of (Ω, r_W) using the data-function $-\log \widehat{N}_\Omega^{(c)}(\epsilon) / \log \epsilon$ at a well chosen radius ϵ. The following observation makes it clear that each graphon requires a specific choice of radius, and thus no (universal) radius is suited for all graphons.

> Observation. 1/at very small scale (i.e. very small ϵ), the covering number may just count the points of the sample $\omega_1, \dots, \omega_n$ and the data look zero-dimensional; 2/if the scale is comparable to the noise due to the distance estimation, the covering number estimator $\widehat{N}_\Omega^{(c)}(\epsilon)$ is not reliable; 3/for an intermediate scale, it is possible to have a good estimation of the dimension, as we shall see in Theorem 1.4.4; 4/at very big scale, the apparent geometry may not reflect the Minkowski dimension (which is, by definition, a measure of the complexity at infinitesimal scale).

Hence, we consider a subset of graphons for which there exists a radius that is well-suited for dimension estimation. We sometimes denote $dim\,\Omega$ by d for brevity, and write $B(\omega, \epsilon)$ the ball of center $\omega \in \Omega$ with radius ϵ (w.r.t. the neighborhood distance). Given constants $D, v, \alpha > 0$ and $M \geq 1 \geq m > 0$, we define the set $\mathcal{W}(D, \alpha, m, M, v)$ of all (pure) graphons (Ω, μ, W) satisfying

1. $dim\,\Omega \leq D$.
2. For $dim\,\Omega := d$ and all $\epsilon \in]0, v]$,

$$\alpha\,\epsilon^d \leq \mu\left[B(\omega, \epsilon)\right] \tag{$H_1^{\alpha,v}$}$$

$$m\epsilon^{-d} \leq N_\Omega^{(c)}(\epsilon) \leq M\epsilon^{-d}. \tag{$H_2^{m,M,v}$}$$

The assumption $H_2^{m,M,v}$ links the covering number with the Minkowski dimension of the graphon. The condition $H_1^{\alpha,v}$ enforces a minimal measure for each ball of (Ω, r_W); in particular, it strengthens the non-zero measure of balls of pure graphons, seen in Section 1.3.1. Mention

can be made of the problem of recovery of the dimension of a manifold, where similar hypotheses are often considered [see Koltchinskii, 2000, for example]. Besides, $H_1^{\alpha,v}$ may be seen as a small-ball condition used in learning problems [Mendelson, 2014, Lecué et al., 2018].

With the radius

$$\epsilon_D \asymp \left(\frac{\log n}{n}\right)^{1/(4\vee 2D)} \tag{1.16}$$

we consistently estimate the Minkowski dimension (Theorem 1.4.4) using the following estimator

$$\widehat{dim}_D := \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon_D)}{-\log \epsilon_D}. \tag{1.17}$$

Theorem 1.4.4 *For all graphons (Ω, μ, W) in $\mathcal{W}(D, \alpha, m, M, v)$ and all large enough n, we have*

$$\left|\widehat{dim}_D - dim\,\Omega\right| \leq \frac{C(D, \alpha, m, M)}{\log n}$$

with probability at least $1 - C'(\alpha, M)/n$ w.r.t. the distribution $\mathbb{P}_{(\Omega,\mu,W)}$, and for some constants $C'(\alpha, M)$ and $C(D, \alpha, m, M)$ that are independent of n.

Theorem 1.4.4 is a corollary of Theorem 1.E.1 in Appendix 1.E.1, which gives a non-asymptotic high probability bound for $-\log \widehat{N}_\Omega^{(c)}(\epsilon)/\log \epsilon$ at any radius ϵ.

One can observe that the convergence rate $\log^{-1} n$ of Theorem 1.4.4 is optimal, in the sense that faster convergence rates cannot be achieved by any estimator of the form $\log \widehat{N}_\Omega^{(c)}(\hat{\epsilon})/ - \log \hat{\epsilon}$ [2]. To see it, take a graphon of dimension $d > 1$ with covering number $N_\Omega^{(c)}(\epsilon) = m\epsilon^{-d}$ for some constant $m > 1$. Even if there exists a covering number estimator that gives a perfect estimation, i.e. $\widehat{N}_\Omega^{(c)} = N_\Omega^{(c)}$, this still entails an error for the dimension estimation. Indeed, in such a case we have:

$$\left|\frac{\log \widehat{N}_\Omega^{(c)}(\epsilon)}{-\log \epsilon} - d\right| = \frac{\log m}{-\log \epsilon}$$

which is (at least) of the order $\log^{-1} n$ since the radius ϵ cannot be taken smaller than n^{-1} in general (otherwise, the estimator of the covering number may just count the n sampled points). Thus, the convergence rate $\log^{-1} n$ is optimal for the classical method of estimation of the Minkowski dimension, which is based on the the plug-in of a covering number estimate into formula (1.5).

Next we show that no estimator [3] can improve on the error bound $\log^{-1} n$, over the following sequence of sets. Given $n > 0$, let $\mathcal{W}_n(D, \alpha, m, M, v)$ be the class of all (pure)

[2] where $\widehat{N}_\Omega^{(c)}$ is any consistent estimator of the covering number, and $\hat{\epsilon}$ is any estimator of a "well chosen radius"

[3] defined as a function of the adjacency matrix $A \in \{0,1\}^{n\times n}$.

graphons fulfilling, for all $\epsilon > 1/n$, the conditions of the above set $\mathcal{W}(D, \alpha, m, M, v)$. On this sequence of sets, one can readily extend Theorem 1.4.4 and retrieve the same error bound, using the same estimator (1.17). This means that there exist some constants $C(D, \alpha, m, M)$ and $C'(\alpha, M)$ that are independent of n, such that for all graphons in $\mathcal{W}_n(D, \alpha, m, M, v)$ and all large enough n, the following error bound holds

$$\left| \widehat{dim}_D - dim\, \Omega \right| \leq \frac{C(D, \alpha, m, M)}{\log n} \tag{1.18}$$

with probability at least $1 - C'(\alpha, M)/n$. Then, Theorem 1.4.5 shows that no estimator can improve on the (order of the) bound (1.18). The proof is written in Appendix 1.E.2.

Theorem 1.4.5 *For any $D > 2$, some numerical constants $\alpha, m, M, v > 0$ and all large enough n, we have*

$$\inf_{\widehat{d}} \sup_{\mathcal{W}_n(D,\alpha,m,M,v)} \mathbb{P}_{(\Omega,\mu,W)} \left[|\hat{d} - dim\, \Omega| \geq \frac{1}{2\log(n)} \right] \geq \frac{1}{4}$$

where $\inf_{\widehat{d}}$ is the infimum over all estimators.

Let us discuss the minimal aspect of the conditions defining $\mathcal{W}_n(D, \alpha, m, M, v)$. First, the assumption that the dimension is upper bounded seems natural, as our available data $A \in \{0,1\}^{n\times n}$ is a finite set. Indeed, for metric spaces (Ω_n, r_{W_n}) with arbitrary large dimensions (like $dim\, \Omega_n/n \to \infty$ for instance), a finite sample $\omega_1, \ldots, \omega_n$ may look like a set of distant and isolated points, which does not reflect the true geometry of (Ω_n, r_{W_n}). Since this situation is not conducive to accurate estimates of the complexity of Ω_n, we avoid it by assuming the dimension is upper bounded. Second, we show that the assumptions $H_1^{\alpha,v}$ and $H_2^{m,M,v}$ are minimal, in the sense that, removing any one of them entails a large loss for any estimator. Specifically, let $\mathcal{W}_n^{min(j)}(D, \alpha, m, M, v)$ be the collection of all (pure) graphons satisfying all conditions of the set $\mathcal{W}_n(D, \alpha, m, M, v)$ except the condition H_j (where H_j denotes $H_1^{\alpha,v}$ or $H_2^{m,M,v}$ according to the value of $j \in \{1, 2\}$). Then, Theorem 1.4.6 shows that any estimator suffers from an error of the order D, over the class $\mathcal{W}_n^{min(j)}(D, \alpha, m, M, v)$. The proof is written in Appendix 1.E.2.

Theorem 1.4.6 *For any $D > 2$, some numerical constants $\alpha, m, M, v > 0$, all $j \in \{1, 2\}$ and all large enough n, we have*

$$\inf_{\widehat{d}} \sup_{\mathcal{W}_n^{min(j)}(D,\alpha,m,M,v)} \mathbb{P}_{(\Omega,\mu,W)} \left[|\hat{d} - dim\, \Omega| \geq \frac{D}{2} \right] \geq \frac{1}{4}$$

where $\inf_{\widehat{d}}$ is the infimum over all estimators.

Remark: our optimal rate of estimation may seem at odds with the faster rates of convergence in the literature about intrinsic dimension estimation, see [Kim et al., 2016] for

instance. This is due to the important differences in the modeling assumptions. In the work of Kim et al. [2016], for example, the observed data are n i.i.d. sampled points from a well-behaved manifold in $\mathbb{R}^m$ whose dimension is an integer. In contrast, here we do not assume the dimension is an integer, not observe the n sampled points $\omega_1, \dots, \omega_n$, and not know the metric r_W.

COMMENTS ON $H_1^{\alpha,v}$, $H_2^{m,M,v}$: we only make the assumptions $H_1^{\alpha,v}$, $H_2^{m,M,v}$ at a small scale, that is for $\epsilon \in]0, v]$. Besides, the right hand side of $H_2^{m,M,v}$ is almost free since it is already implied by $H_1^{\alpha,v}$ for $M = 2^d/\alpha$. Let us briefly explain how these assumptions imply the error bound of Theorem 1.4.4. The assumption $H_1^{\alpha,v}$ ensures that the difference between the sampled points $\omega_1, \dots, \omega_n$ and the latent space Ω is not too large. By definition, this implies that the sampling error (1.13) and the distance error (1.14) are small. Accordingly, we can choose a radius ϵ_D that is larger than these two errors, and reliably estimate the ϵ_D-covering number $N_\Omega^{(c)}(\epsilon_D)$ by Proposition 1.4.3. Then, we use a plug-in to estimate the quantity $-\log N_\Omega^{(c)}(\epsilon_D)/\log \epsilon_D$, which is a good approximation of the dimension by assumption $H_2^{m,M,v}$. To sum up, the radius ϵ_D must be larger than the sampling and distance errors, but still small enough to well approximate the Minkowski dimension with $-\log N_\Omega^{(c)}(\epsilon_D)/\log \epsilon_D$.

1.5 Testing the complexity

Given the adjacency matrix of a W-random graph, we want to known if the graph is simple or complex. In other words, we would like to test the null-hypothesis $N_\Omega^{(c)}(\epsilon) \leq K$ for a given $K > 0$, with a specific care for minimizing the assumptions on the graphon. However, instead of using the covering number we use the packing number $N_\Omega^{(p)}(\epsilon)$ for some reasons to be specified in Section 1.5.1. For now, note that it is essentially the same measure as the covering number, and all previous results of the paper can be adapted to the packing number (without any significant difference). See Appendix 1.A for a reminder of this usual measure for metric spaces.

In hypothesis testing, it is common to be conservative and focus on the minimization of the type I error, which is the probability of rejecting the null-hypothesis incorrectly. Accordingly, our objective is to control the type I error without any assumption on the graphon, while keeping a control of the type II error under reasonable assumptions. (the type II error is the probability of accepting the null-hypothesis incorrectly)

1.5.1 Testing the null-hypothesis without assumption on the graphon, via under-estimation of the packing number

To test the null-hypothesis without assumption on the graphon, we want to define a complexity estimator that does not overestimate the true complexity w.h.p.. Unfortunately, the inequality on the covering number estimator from Proposition 1.4.3

$$\widehat{N}_\Omega^{(c)}(\epsilon + b_{sup} + s_\omega) \leq N_\Omega^{(c)}(\epsilon)$$

is difficult to leverage for an under-estimation since the errors b_{sup} and s_ω are unknown and take specific values for each graphon. However, we show below that the sampling error s_ω can be removed, by working with the packing number instead of the covering number. Then, we show that the distance error bound b_{sup} can be handled with a slight modification of the distance-estimator $\widehat{r}$, defined earlier by (1.11).

Based on the distance estimator $\widehat{r}$, we can define a plug-in estimator $\widehat{N}_\Omega^{(p)}(\epsilon)$ of the packing number, as we did for the covering number estimator. This estimator satisfies almost the same non-asymptotic bounds as the covering number estimator, see the following proposition, which is a slight variant of Proposition 1.4.3. The proof is omitted.

Proposition 1.5.1 *Given any graphon (Ω, μ, W), consider the data distribution $\mathbb{P}_{(\Omega,\mu,W)}$ defined in model (1.3). Let b_{sup} and s_ω be the distance error bound (1.14) and the sampling error (1.13). Then, the packing number estimator $\widehat{N}_\Omega^{(p)}$ satisfies the following inequalities*

$$\forall \epsilon > b_{sup}, \qquad N_\Omega^{(p)}\left(\epsilon + b_{sup} + 2s_\omega\right) \leq \widehat{N}_\Omega^{(p)}\left(\epsilon\right) \leq N_\Omega^{(p)}\left(\epsilon - b_{sup}\right)$$

with probability at least $1 - \frac{2}{n}$ with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W)}$.

Hence, we have

$$\widehat{N}_\Omega^{(p)}\left(\epsilon + b_{sup}\right) \leq N_\Omega^{(p)}\left(\epsilon\right)$$

without the sampling error s_ω anymore.

The next step is to control the remaining error term b_{sup}. We do so by modifying the previous estimator $\widehat{r}$ as follows:

$$\widehat{r}^2_{new}(i,j) := \left[\langle A_i, A_{\widehat{m}(i)}\rangle_n + \langle A_j, A_{\widehat{m}(j)}\rangle_n - 2 \max_{k\in\{i,\widehat{m}(i)\}, l\in\{j,\widehat{m}(j)\}} \langle A_k, A_l\rangle_n\right]_+ \tag{1.19}$$

which satisfies the same upper bound as $\widehat{r}$ in Theorem 1.4.1, up to a numerical constant $5/3$ (see Lemma 1.G.1 in Appendix 1.G.1). The new packing number estimator based on $\widehat{r}_{new}$ is denoted by $\widehat{N}_\Omega^{(p.new)}$, and provides the under-estimation of the packing number (Theorem 1.5.2). The proof is written in Appendix 1.G.1.

Theorem 1.5.2 *Given any graphon (Ω, μ, W), consider the data distribution $\mathbb{P}_{(\Omega,\mu,W)}$ defined in model (1.3). Then, for the radius $\widehat{\epsilon} = \sqrt{\epsilon^2 + t_n}$ with $t_n = 12\sqrt{\frac{\log n}{n}}$, the estimator $\widehat{N}_\Omega^{(p.new)}$ satisfies the following inequalities*

$$\forall \epsilon > 0, \qquad N_\Omega^{(p)}\left(\widehat{\epsilon} + \frac{5}{3}b_{sup} + 2s_\omega\right) \leq \widehat{N}_\Omega^{(p.new)}(\widehat{\epsilon}) \leq N_\Omega^{(p)}(\epsilon) \tag{1.20}$$

with probability at least $1 - \frac{2}{n}$ with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W)}$.

Thus, without any assumption on the graphon, the estimator $\widehat{N}_\Omega^{(p.new)}(\widehat{\epsilon})$ does not overestimate the ϵ-packing number with high probability. Besides, the left hand side of (1.20) shows that it does not under-estimate (significantly) more than the previous estimator $\widehat{N}_\Omega^{(p)}$ of the packing number (seen in Proposition 1.5.1).

1.5.2 Results on the packing number test

We accept the null hypothesis if and only if $\widehat{N}_{\Omega}^{(p.new)}(\widehat{\epsilon}) \leq K$. The upper bound (1.20) ensures that the type I error is controlled for all graphons, which gives the following result.

Corollary 1.5.3 *For any graphon, the type I error is lower than $\frac{2}{n}$ with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W)}$.*

By definition of the packing number, the type II error is small as soon as $K+1$ sampled points are separated by at least a distance $\widehat{\epsilon}+err$, where err upper bounds all errors of distance estimation between the $K+1$ points. This condition on the sampled points is satisfied w.h.p. by each of the following graphons.

Given two parameters $\eta > 0$ and $\beta > 1/n$, let $\mathcal{W}(\eta, \beta)$ denote a collection of graphons for which there exist $K+1$ balls $B(x_1, \eta_1), \ldots, B(x_{K+1}, \eta_{K+1})$ in (Ω, r_W) such that

1. the $K+1$ balls are weighted enough: $\mu\left[B(x_i, \eta_i)\right] \geq \beta$ for all $i \in [K+1]$,
2. the radii are small enough: $\eta_i \leq \eta/2$ for all $i \in [K+1]$,
3. the centers are spaced enough: $r_W(x_i, x_j) \geq \sqrt{\epsilon^2 + 10\eta + 6t_n} + \eta$.

The small-ball condition 1. is similar to the assumption $H_1^{\alpha,v}$ for the dimension estimation; it ensures that some of the sampled points $\omega_1, \ldots, \omega_n$ belong to the $K+1$ balls w.h.p.. The third condition 3. ensures that these balls are enough distant from each other, so that the sampled points in these balls are separated enough, in order to have $\widehat{N}_{\Omega}^{(p.new)}(\widehat{\epsilon}) \geq K+1$ and confirm the alternative hypothesis correctly.

Theorem 1.5.4 *Assume the graphon (Ω, μ, W) belongs to $\mathcal{W}(\eta, \beta)$ for some $\beta > 1/n$. Then, the type II error is smaller than*

$$\frac{2}{n} + 2\,\beta n(K+1)\exp[-\beta(n-1)]$$

with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W)}$.

The proof of Theorem 1.5.4 is written in Appendix 1.G.2. This result implies that, for any graphon in $\mathcal{W}(\eta, \beta)$, the type II error is convergent to zero as soon as the measure of each ball $B(x_i, \eta_i)$ is large enough to satisfy $\beta \gtrsim n^{-1}$. For example, if each of the $K+1$ balls has a measure that is larger than $\log[Kn]/n$, then the type II error is smaller than $\log(n)/n$ up to some numerical constant. In Appendix 1.A.3, Theorem 1.5.4 is improved by using the graphon regularity at a finer level (see Theorem 1.A.1).

1.6 Further considerations

1.6.1 Estimation of the complexity with sparse observations

In the W-random graph model (1.3), each node has an average degree that is linear with n the total number of nodes. However, real-world networks are often sparse with node degrees

varying from zero to n. This motivates to consider a model of sparse graph where the node degree can be an order of magnitude smaller than n.

Given a sequence ρ_n such that $\rho_n \to 0$, the definition of model (1.3) can be modified to have average node degrees of the order of $\rho_n n$. Consider the adjacency matrix A, defined by model (1.3), whose edges are independently retained with probability ρ_n and erased with probability $1-\rho_n$. We refer to this set-up as "the sparse setting" and denote by $\mathbb{P}_{(\Omega,\mu,W),\rho_n}$ the corresponding data distribution. This model has been considered several times in the literature [see Bickel et al., 2011, Wolfe and Olhede, 2013, Klopp et al., 2017, Xu et al., 2014].

We now extend the results of Section 1.4 to this sparse setting. Corollary 1.6.1 gives non-asymtotic error bounds for the distance estimation. It is a slight variant of Theorem 1.4.1. For completeness, the proof is written in Appendix 1.F.1.

Corollary 1.6.1 *Assume the scaling parameter ρ_n is lower bounded by*

$$\rho_n \geq 2\sqrt{\log(n)/(n-2)}. \tag{1.21}$$

Then, the following event

$$\mathcal{E}^{sp}_{dist} = \Big\{ \forall i,j \in [n] : \big|\rho_n^2 r_W^2(\omega_i,\omega_j) - \widehat{r}^2(i,j)\big| \\ \leq 3\rho_n\Big(\rho_n r_W(\omega_j,\omega_{m(j)}) + \rho_n r_W(\omega_i,\omega_{m(i)}) + 20\sqrt{\log(n)/n}\Big)\Big\}$$

holds with probability $\mathbb{P}_{(\Omega,\mu,W),\rho_n}\left(\mathcal{E}^{sp}_{dist}\right) \geq 1-\frac{2}{n}$.

As in Section 1.4, it is possible to show a matching lower bound here, implying that Corollary 1.6.1 is optimal.

We estimate the Minkowski dimension using the following radius

$$\epsilon_{D,\rho_n} \asymp \left(\frac{\log n}{n}\right)^{1/(2D)} \vee \rho_n^{-1/2}\left(\frac{\log n}{n}\right)^{1/4} \tag{1.22}$$

Corollary 1.6.2 is an adaptation of Theorem 1.4.4 for the sparse setting. The proof is written in Appendix 1.F.2.

Corollary 1.6.2 *For all graphons (Ω,μ,W) in $\mathcal{W}(D,\alpha,m,M,v)$, all scaling parameters ρ_n fulfilling (1.21), and all radii satisfying (1.22), the following rate of estimation of the dimension holds with probability tending to 1 as $n\to\infty$ (w.r.t. the distribution $\mathbb{P}_{(\Omega,\mu,W),\rho_n}$).*

$$\left|\frac{\log \widehat{N}^{(c)}_{\Omega}(\epsilon_{D,\rho_n})}{-\log \epsilon_{D,\rho_n}} - d\right| \leq C(D,\alpha,m,M,t)\begin{cases} 1 & \text{if } \rho_n \asymp \sqrt{\log(n)/n}, \\ (\log n)^{-1} & \text{if } \rho_n \asymp (\log(n)/n)^{(1/2)-t}, \end{cases}$$

where $t\in]0,1/2[$ and $C(D,\alpha,m,M,t)$ is some constant independent of n.

1.6.2 Polynomial-time algorithm (with some theoretical guarantees)

In contrast with the previous sections, here we take into account the computational aspect of the problem. Computing the covering number of a finite set is NP-hard, hence we approximate it with a greedy algorithm [Chvatal, 1979].

For completeness, the polynomial-time procedure for estimating $N_\Omega^{(c)}(\epsilon)$ is described below. The algorithm proceeds in two steps: Step 1 computes all distances $\widehat{r}(i,j)$ using the distance-estimator (1.11); in particular, this step requires the computation of all index estimators $\widehat{m}(j)$ defined by (1.10). Step 2 approximates the ϵ-covering number of $\{1,\dots,n\}$ w.r.t. the distance estimator $\widehat{r}$, by sequentially selecting balls (of radius ϵ) according to one rule: at each stage, select the ball that contains the largest number of uncovered elements. At the end of the process, the number of selected balls is returned. This output is denoted by $\widehat{N}_\Omega^{(ap.c)}(\epsilon)$.

COVERING NUMBER ALGORITHM

Input: $A = [A_{ij}]$ adjacency matrix of size $n \times n$, a radius ϵ.

Step 1 : constructing the distance-estimator $\widehat{r}$

1. Compute the nearest neighbor's index of each sampled point ω_i:
   $\forall i \in \{1,\dots,n\}, \quad \widehat{m}(i) = \underset{j:\ j\neq i}{\operatorname{argmin}} \ \max_{k:\ k\neq i,j} \left|\langle A_k, A_i - A_j\rangle_n\right|.$

2. Compute all the distances:
   $\forall i,j \in \{1,\dots,n\}, \quad \widehat{r}(i,j) = \langle A_i, A_{\widehat{m}(i)}\rangle_n + \langle A_j, A_{\widehat{m}(j)}\rangle_n - 2\langle A_i, A_j\rangle_n.$

Step 2 : computing an approximation of the ϵ-covering number

3. In the space $\mathcal{S}_0 = \{1,\dots,n\}$ endowed with the distance function $\widehat{r}$, consider $\mathcal{B}_0 = \{B_j\}_{j\leq n}$ the set of all the balls of radius ϵ.

4. Obtain a cover of $\{1,\dots,n\}$ as follows:
   Set $i = 0$. While $\mathcal{S}_i \neq \emptyset$, do:

   (a) Select a ball B in $\mathcal{B}_i$ that contains the largest number of elements of $\mathcal{S}_i$.

   (b) Set $\mathcal{S}_{i+1} = \mathcal{S}_i \setminus B$ to remove the elements covered by B,

   (c) Set $\mathcal{B}_{i+1} = \mathcal{B}_i \setminus \{B\}$ to update the set of available balls,

   (d) Set $i = i + 1$ to continue the algorithm.

Output: the number i of selected balls, denoted by $\widehat{N}_\Omega^{(ap.c)}(\epsilon)$.

We also suggest an heuristic for tuning ϵ in the estimation of the Minkowski dimension. First, run several times COVERING NUMBER ALGORITHM for a range of different radii $\epsilon_1,\dots,\epsilon_t$, and then plot $\log \widehat{N}_\Omega^{(ap.c)}(\epsilon_j)/\log \epsilon_j$ for $j = 1,\dots,t$. As in Figure 1.1, we look for a graph function that (roughly) admits the three following parts: 1/for big radii, the shape of the curve is irregular and seems sawtooth; 2/for medium radii, there is almost a plateau whose value is the dimension estimate; 3/for small radii, there is an abrupt drop towards zero.

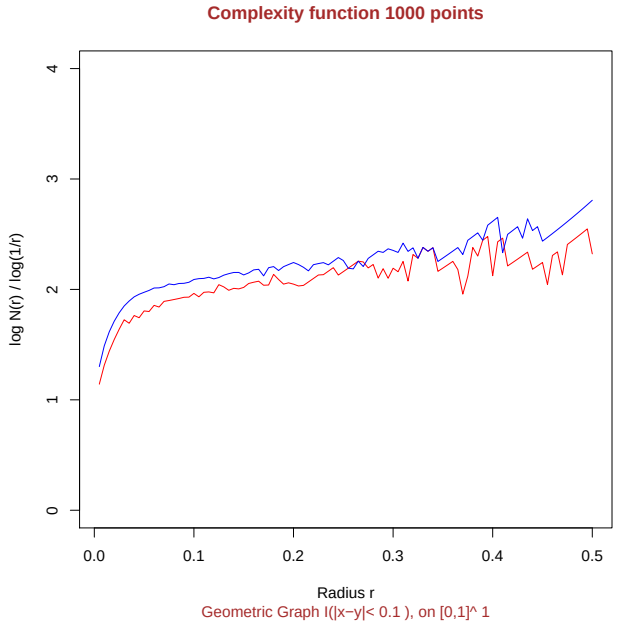


Figure 1.1: W-random graph with Minkowski dimension 2

According to the theoretical guarantee of the greedy algorithm [Chvatal, 1979], one has

$$\widehat{N}_{\Omega}^{(c)}(\epsilon) \leq \widehat{N}_{\Omega}^{(ap.c)}(\epsilon) \leq 2\log(n)\widehat{N}_{\Omega}^{(c)}(\epsilon)$$

where $\widehat{N}_{\Omega}^{(c)}(\epsilon)$ is the consistent estimator introduced in Section 1.4. Then, for graphons fulfilling the assumptions of Theorem 1.4.4, there exist some radii ϵ such that $-\log \widehat{N}_{\Omega}^{(ap.c)}(\epsilon)/\log \epsilon$ is close to the Minkowski dimension up to a small error term $-\log\big(2\log(n)\big)/\log \epsilon$.

We shortly illustrate the empirical performance of our algorithm on the random geometric graph, introduced in Section 1.3.2. Consider the latent space $[0,1]$, endowed with the uniform measure and the function $W(x,y) = \mathbb{I}_{||x-y||_2 \leq 0.1}$, which has a Minkowski dimension 2 and satisfies the assumptions of Theorem 1.4.4. We sample $n = 1000$ points uniformly on $[0,1]$ and plot the outputs $-\log \widehat{N}_{\Omega}^{(ap.c)}(\epsilon)/\log \epsilon$ over the range of radii $\epsilon \in \Big\{0.005 + k * 0.005\,;\, k \in \{0,\dots,100\}\Big\}$. This is represented by the red curve in Figure 1.1. As we can see, it is close to the true dimension at some intermediate radii, which coincides with our theoretical results. Specifically, we observe the three typical parts in the graph function: 1/ on the right of the figure, the sawtooth-shaped curve means that the radius is too big for approaching the Minkowski dimension (which is by definition a limit in $\epsilon \to 0$); 2/ on the middle, there is a plateau whose value is close to the dimension; 3/ on the left, there is an abrupt drop because the covering number estimator eventually just counts the sampled points $\omega_1, \dots, \omega_n$. As a reference, we also plot $-\log N_{\omega_1,\dots,\omega_n}^{(ap.c)}(\epsilon)/\log \epsilon$ in blue, where $N_{\omega_1,\dots,\omega_n}^{(ap.c)}(\epsilon)$ is the *approximated* covering number of the sample $\{\omega_1, \dots, \omega_n\}$ w.r.t. to the true distance r_W.

Appendices

1.A Additional information

1.A.1 Basic information on the covering and packing numbers and the Minkowski dimension

Given any set S, its covering number $N^{(c)}(\epsilon)$ is the minimal number of balls of radius ϵ required to entirely cover S, with the constraint that the ball centers are in S. This measure is widely used for general metric spaces. Likewise, the packing number $N^{(p)}(\epsilon)$ is the maximum number of points in a given space (strictly) separated by at least a given distance ϵ. Both measures are similar and linked by the following inequalities $N^{(c)}(\epsilon) \leq N^{(p)}(\epsilon) \leq N^{(c)}(\epsilon/2)$. In all the paper (except the last subsection 1.5), our results are mostly stated with the covering number, but each of them can be adapted to the packing number.

The covering number requires to choose the scale ϵ at which we look at the data. To get rid of this parameter, it is common to consider the Minkowski dimension which is defined by $\lim_{\epsilon \to 0} \; -\log N^{(c)}(\epsilon)/\log \epsilon$. Note that the same formula holds with the packing number instead. The Minkowski dimension is useful for infinite (separable) spaces, when the covering number diverges to infinity as ϵ goes to zero. This dimension is therefore complementary to the covering number. It is known to match with some other classical notions of dimension in simple cases, for example the Minkowski dimension of the hypercube $[0,1]^d$ is equal to its Euclidean dimension d. The Minkowski dimension has the advantage to be applicable on a wide range of spaces (whose dimension is not necessarily an integer) and to be easy to compute (in comparison with the Hausdorff dimension for example).

1.A.2 Details on the illustrative examples

Random Hölder graph. Recall that the graphon (Ω, μ, W) is $([0,1]^d, \lambda, W)$ where λ is the uniform measure on $[0,1]^d$ and W satisfies the following condition: there exist three constants $m, M, \alpha > 0$ such that for all $\omega, \omega', \omega'' \in [0,1]^d$,

$$m\big|\big|\omega' - \omega\big|\big|_2^\alpha \leq \big|W(\omega', \omega'') - W(\omega, \omega'')\big| \leq M\big|\big|\omega' - \omega\big|\big|_2^\alpha$$

where α is the level of regularity of the function W and $||\omega' - \omega||_2$ is the Euclidean distance between ω' and ω in $[0,1]^d$. From the above display, we directly deduce some bounds on the neighborhood distance (1.4) :
$\forall \omega, \omega' \in [0,1]^d$,

$$m||\omega' - \omega||_2^\alpha \leq r_W(\omega', \omega) \leq M||\omega' - \omega||_2^\alpha.$$

Thus, the distance r_W behaves (up to some constants) like the Euclidean distance on $[0,1]^d$ raised to the power of α. As the covering number of the Euclidean hypercube $([0,1]^d, ||.||_2)$ is approximately equal to ϵ^{-d} for small radii, we have

$$(\epsilon/m)^{-d/\alpha} \lesssim N_\Omega^{(c)}(\epsilon) \lesssim (\epsilon/M)^{-d/\alpha}\,.$$

Hence $dim\,\Omega = d/\alpha$, which means that the Minkowski dimension of (Ω, r_W) is equal to the ratio between the Euclidean dimension of the latent space $[0,1]^d$ and the regularity of the function W.

Random geometric graph example. Recall that the graphon is $([0,1]^d, \lambda, W)$ where λ is the uniform measure, and W is defined as $W(\omega,\omega') = \mathbb{I}_{||\omega-\omega'||_2 \leq \delta}$ for some parameter $\delta \in]0,1[$, and $||\omega-\omega'||_2$ is the Euclidean distance between $\omega, \omega' \in [0,1]^d$. Here, the bounds on the neighborhood distance are rather involved and deferred to the Appendix 1.B.2. The main message is that

$$r_W(\omega,\omega') \asymp \sqrt{||\omega-\omega'||_2}$$

if $||\omega-\omega'||_2$ is small enough, which means that the distance r_W behaves like the squared root of the Euclidean norm in $[0,1]^d$. Following the line of the Random Hölder graph example, we can see that $N_\Omega^{(c)}(\epsilon)$ behaves like ϵ^{-2d} for ϵ small enough. By definition of the Minkowski dimension, it follows that $dim\,\Omega = 2d$.

1.A.3 Test: improvement of the type II error

The control of the type II error can be refined using the graphon regularity at a finer level. Instead of considering the set $\mathcal{W}(\eta,\beta)$ of graphons with $K+1$ well separated balls (Theorem 1.5.4), here we consider the new set $\mathcal{W}(\eta,\beta,M,K')$ of graphons with M disjoint collections of $K+1+K'$ separated balls. That is, for a collection of $K+1+K'$ balls, we assume the same conditions of separation, size and measure as in a collection of $K+1$ balls defined by $\mathcal{W}(\eta,\beta)$ (in Theorem 1.5.4). In addition, we assume that the M formations of $K+1+K'$ balls do not intersect each other (i.e. no ball from a collection overlaps a ball from another collection). Thus, the new set $\mathcal{W}(\eta,\beta,M,K')$ of graphons is linked with the previous one by the following equality $\mathcal{W}(\eta,\beta,1,0) = \mathcal{W}(\eta,\beta)$.

Theorem 1.A.1 *If the underlying graphon belongs to $\mathcal{W}(\eta,\beta,M,K')$ with $\beta \geq 1/n$, then the type II error is smaller than $\frac{2}{n} + \widetilde{p}_n^M$, where $\widetilde{p}_n$ admits the following upper bound*

$$\binom{K+K'+1}{K'+1}\Big(2\beta n\,\exp[-\beta(n-1)]\Big)^{(K'+1)}.$$

The proof of Theorem 1.5.4 is written in Appendix 1.G.2.

1.B Proofs for illustrative examples

1.B.1 Proof of Proposition 1.3.4: approximation by SBM

Given a graphon (Ω,μ,W) and a radius $\epsilon > 0$, we consider a cover of (Ω, r_W) whose the cardinality is $N_\Omega^{(c)}(\epsilon)$ (written N for brevity), and the ball centers are $x_1,\dots,x_N$. The Voronoi cell V_j of x_j is the set of all elements in Ω that are closer to x_j than to any other x_k, $k \neq j$, according to the metric r_W. In the case of equality, where a point ω is at equal distance of several ball centers x_i, it belongs to the Veronoi cell of smallest index i.

$$V_j := \Big\{\omega \in \Omega : \quad r_W(\omega,x_j) < r_W(\omega,x_k) \text{ if } k<j, \\ \text{and } r_W(\omega,x_j) \leq r_W(\omega,x_k) \text{ otherwise}\Big\}$$

Define the SBM approximation of W as follows:

$$\overline{W}(x,y) = \sum_{i,j=1}^{N} 1_{x\in V_i} 1_{y\in V_j} \frac{1}{\mu(V_i)\mu(V_j)} \int_{V_i}\int_{V_j} W(z_1,z_2)\mu(dz_1)\mu(dz_2)$$

By triangular inequality and Jensen inequality, the expression

$$\int_{\Omega^2} (W(x,y) - \overline{W}(x,y))^2 \mu(dx)\mu(dy)$$

is upper bounded by

$$\leq \quad 2\int_{\Omega^2} \sum_{j=1}^{N} 1_{y\in V_j} \Big[\frac{1}{\mu(V_j)} \int_{V_j} [W(x,y) - W(x,z_2)]^2 \mu(dz_2)\Big]\mu(dx)\mu(dy) \quad +$$

$$2\int_{\Omega^2} \Big[\sum_{i,j=1}^{N} 1_{x\in V_i} 1_{y\in V_j} \frac{1}{\mu(V_i)\mu(V_j)} \int_{V_i}\int_{V_j} [W(x,z_2) - W(z_1,z_2)]^2 \mu(dz_1)\mu(dz_2)\Big]\mu(dx)\mu(dy)$$

Note that the first term is smaller than $8\epsilon^2$ by integrating with respect to x and using the fact that y and z_2 belong to the same Voronoi cell. The second term simplifies

$$2\int_{\Omega^2} \Big[\sum_{i,=1}^{N} 1_{x\in V_i} \frac{1}{\mu(V_i)} \int_{V_i} [W(x,y) - W(z_1,y)]^2 \mu(dz_1)\Big]\mu(dx)\mu(dy)$$

which is again smaller than $8\epsilon^2$. The approximation error of W by $\overline{W}$ is therefore lower than 4ϵ in l_2-norm. The proposition is proved. □

1.B.2 The neighborhood distance for the random geometric graph example

Lemma 1.B.1 gives bounds on the neighborhood distance for the random geometric graph of Section 1.3.2. For simplicity, we neglect the side effects associated with a point too close of the side of $\Omega = [0,1]^d$. That is, we assume the parameter δ is small compared to 1 (where 1 is the length of a side of $[0,1]^d$). Write V_d the volume of the unit ball in $[0,1]^d$ endowed with the Euclidean norm $||.||_2$, and write $I_x(.,.)$ the (regularized) incomplete beta function [see DLMF, Eq.8.17.2 for a definition].

Lemma 1.B.1 *If $||x-y||_2 > 2\delta$, then $r_W^2(x,y) = 2V_d\delta^d$; otherwise $r_W^2(x,y) = 2V_d\delta^d I_x(\frac{1}{2}, \frac{d+1}{2})$ for $x = \left(\frac{||x-y||_2}{2\delta}\right)^2$. As a consequence, $\sqrt{||x-y||_2} \lesssim r_W(x,y) \lesssim \sqrt{||x-y||_2}$ as soon as $||x-y||_2$ is small enough (compared to δ).*

According to the above lemma, the neighborhood distance r_W behaves like the squared root of the Euclidean norm of $[0,1]^d$ if $||x-y||_2$ is small enough. For lower dimensions, for instance $d=3$, we can also use the paper of Li [2011] to get the simpler formula:
if $||x-y||_2 < 2\delta$, then

$$r_W^2(x,y) = 2\pi\left(\delta^2 - \frac{||x-y||_2^2}{12}\right)||x-y||_2.$$

Proof of Lemma 1.B.1. For the random geometric graph, observe that the computation of the neighborhood distance is equivalent to the computation of the volumes of hypersherical caps. Using the formula (3) in the paper of Li [2011] (and neglecting the side effects due to the boundary of the latent space), we have:
if $||x-y||_2 < 2\delta$, then

$$r_W^2(x,y) = 2V_d\delta^d\left[1 - I_x(\frac{d+1}{2}, \frac{1}{2})\right]$$

where $x = 1 - \left(\frac{||x-y||_2}{2\delta}\right)^2$. Basic properties of the (regularized) incomplete beta function [see DLMF, Eq.8.17.4] allows to rewrite the last formula:
if $||x-y||_2 < 2\delta$, then

$$r_W^2(x,y) = 2V_d\delta^d I_x(\frac{1}{2}, \frac{d+1}{2}) \tag{1.23}$$

where $x = \left(\frac{||x-y||_2}{2\delta}\right)^2$. Let $B(a,b)$ denote the beta function [DLMF, Eq.5.12.1], then the above formula (1.23) can be developed using the recurrence formula $I_x(a,b+1) = I_x(a,b) + \frac{x^a(1-x)^b}{bB(a,b)}$ [DLMF, Eq.8.17.21]. It follows that r_W satisfies the following bounds: $\sqrt{||x-y||_2} \lesssim r_W(x,y) \lesssim \sqrt{||x-y||_2}$ as soon as $||x-y||_2$ is small enough. □

1.C Proof of identifiability

1.C.1 Proof of Lemma 1.3.2 : invariance of the neighborhood distance

Given two equivalent pure graphons (Ω, μ, W) and (Ω', μ', W'), let us show that their respective neighborhood distances r_W and $r_{W'}$ are linked by the following $\mu' \otimes \mu'$-almost surely equality

$$r_W\left(\phi(x), \phi(y)\right) = r_{W'}\left(x, y\right)$$

for some measure-preserving bijection $\phi : \Omega' \to \Omega$.

It follows from Lemma 1.C.1, which links any two equivalent pure graphons. Denote by W^ϕ the function $(x,y) \mapsto W(\phi(x), \phi(y))$.

Lemma 1.C.1 [**Lovász, 2012, Section 13.3**] *If two pure graphons (Ω, μ, W) and (Ω', μ', W') are equivalent, then there exists a bijective measure-preserving map $\phi : \Omega' \to \Omega$ such that $W^\phi(x,y) = W'(x,y)$ $\mu' \otimes \mu'$-almost surely.*

Indeed, by definition of the neighborhood distance,

$$r_{W'}\left(x, y\right) = \left(\int_{\Omega'} \left|W'(x,z') - W(y,z')\right|^2 \mu'(dz')\right)^{1/2}$$

which gives the following $\mu' \otimes \mu'$-almost surely equality by Lemma 1.C.1,

$$r_{W'}\left(x, y\right) = \left(\int_{\Omega'} \left|W(\phi(x),\phi(z')) - W(\phi(y),\phi(z'))\right|^2 \mu'(dz')\right)^{1/2}$$

for some measure-preserving bijection $\phi : \Omega' \to \Omega$. Then, using a pushforward measure (or image measure),

$$r_{W'}(x,y) = \left(\int_{\Omega} |W(\phi(x),z) - W(\phi(y),z)|^2 \, \mu(dz) \right)^{1/2}$$

$\mu' \otimes \mu'$-almost surely, so that, by definition of the neighborhood distance,

$$r_{W'}(x,y) = r_W(\phi(x),\phi(y))$$

$\mu' \otimes \mu'$-almost surely. Lemma 1.3.2 is proved. □

1.C.2 Proof of Lemma 1.3.1 : identifiability of the covering number

Given two equivalent pure graphons (Ω,μ,W) and (Ω',μ',W'), let us prove that their respective covering numbers are equal: $N_{\Omega}^{(c)}(\epsilon) = N_{\Omega'}^{(c)}(\epsilon)$ for all $\epsilon > 0$.

According to Lemma 1.3.2, there exists a measure-preserving bijection ϕ, such that the two metric spaces (Ω, r_W) and $(\Omega', r_{W'})$ are linked by the equality $r_{W'}(x,y) = r_W(\phi(x),\phi(y))$ on a subset of measure 1, say $\Sigma \subseteq \Omega'$ with $\mu'(\Sigma) = 1$. This means that both subpaces $(\phi(\Sigma), r_W)$ and $(\Sigma, r_{W'})$ are linked by a bijection that preserves the distances, which directly implies equality between their covering numbers: $N_{\phi(\Sigma)}^{(c)}(\epsilon) = N_{\Sigma}^{(c)}(\epsilon)$ for all $\epsilon > 0$.

Then, for proving Lemma 1.3.1, it is enough to show the two following inequalities

$$N_{\Omega}^{(c)}(\epsilon) \geq N_{\phi(\Sigma)}^{(c)}(\epsilon+\delta) \tag{1.24}$$

$$N_{\Sigma}^{(c)}(\epsilon+\delta) \geq N_{\Omega'}^{(c)}(\epsilon+\delta) \tag{1.25}$$

for any $\delta > 0$. Indeed, combining these two inequalities with the covering number equality from the above paragraph, one has $N_{\Omega}^{(c)}(\epsilon) \geq N_{\Omega'}^{(c)}(\epsilon+\delta)$. Taking the limit $\delta \to 0$ and using the right-continuity of the covering number (Lemma 1.C.2), this gives $N_{\Omega}^{(c)}(\epsilon) \geq N_{\Omega'}^{(c)}(\epsilon)$. As the reverse inequality holds by symmetry of the proof, one obtain the equality $N_{\Omega}^{(c)}(\epsilon) = N_{\Omega'}^{(c)}(\epsilon)$ of Lemma 1.3.1.

Lemma 1.C.2 *Given a pure graphon (Ω,μ,W), the function $\epsilon \mapsto N_{\Omega}^{(c)}(\epsilon)$ is piecewise constant and right-continuous (note that we use closed balls in the definition).*
Likewise, $\epsilon \mapsto N_{\Omega}^{(p)}(\epsilon)$ is a right continuous piecewise function.

Assume Σ is dense in $(\Omega', r_{W'})$. Each cover of Σ is closed as a finite union of closed balls. Hence it is also a cover of Ω' by density of Σ in Ω'. This proves (1.25). Likewise, assume $\phi(\Sigma)$ is dense in (Ω, r_W). An ϵ-cover of Ω can be transformed into an $(\epsilon+\delta)$-cover of $\phi(\Sigma)$ by moving the ball centers from Ω to Σ and increasing the ball radius of δ (for arbitrary small δ). This proves (1.24) for any $\delta > 0$.

Let us show the density of $\phi(\Sigma)$ in (Ω, r_W). One has $\mu(\phi(\Sigma)) = \mu'(\Sigma) = 1$ by definition of a (bijective) measure-preserving map, which implies that $\phi(\Sigma)$ intersects each ball of non-zero measure in (Ω, r_W). As the measure of a pure graphon has full-support by definition,

then each ball of non-zero radius has a non-zero measure. Thus, $\phi(\Sigma)$ intersects each ball of non-zero radius in (Ω, r_W), which means that $\phi(\Sigma)$ is dense in (Ω, r_W). Similarly, we can show the density of Σ in $(\Omega', r_{W'})$.

Lemma 1.3.1 is proved for the covering number. The proof for the packing number is similar and omitted. □

Proof of Lemma 1.C.2. The function $\epsilon \mapsto N_\Omega^{(c)}(\epsilon)$ is non-increasing from $[0, \infty[$ to the set of all non-negative integers, it is therefore a piecewise constant function. Thus, for any radius $\epsilon_0 > 0$, there exists a (strictly) larger radius ϵ_1 such that the covering number $N_\Omega^{(c)}(\epsilon)$ is equal to a constant, say N, over the interval $]\epsilon_0, \epsilon_1[$. To prove the right continuity in ϵ_0, let us show the inequality $N_\Omega^{(c)}(\epsilon_0) \leq N$ (since we already know the reverse inequality by monotonicity of the covering number function), or equivalently that there exists a cover of Ω that is composed of N balls of radius ϵ_0.

Given a radius ϵ and K points $c = (c_1, \ldots, c_K) \in \Omega^K$, denote by $C_\Omega(c, \epsilon)$ the union of K balls of centers $c_1, \ldots, c_K$. In the following, we prove: 1/ the existence of some $c_0 \in \Omega^N$ such that $C_\Omega(c_0, \epsilon)$ covers Ω for all $\epsilon \in]\epsilon_0, (\epsilon_1 + \epsilon_0)/2]$; 2/ for such a c_0, $C_\Omega(c_0, \epsilon_0)$ covers Ω. Thus, Lemma 1.C.2 will be proved.

1/ Define the set $E_\Omega(\epsilon) := \{c \in \Omega^N : \Omega \subseteq C_\Omega(c, \epsilon)\}$ for any given radius $\epsilon > 0$. Then, consider the following sequence of nested sets $\widetilde{E}_k := E_\Omega(\epsilon_0 + (\epsilon_1 - \epsilon_0)/k)$ where $k \geq 2$ is an integer. The Cantor's intersection theorem (recalled in Lemma 1.C.3 below) ensures that $\cap_{k \geq 2} \widetilde{E}_k \neq \emptyset$, provided that the assumptions of the theorem hold. For clarity, this verification is deferred to the end of the proof. As the set $\cap_{\epsilon_0 < \epsilon < \epsilon_1} E_\Omega(\epsilon)$ is equal to $\cap_{k \geq 2} \widetilde{E}_k$, one has $\cap_{\epsilon_0 < \epsilon < \epsilon_1} E_\Omega(\epsilon) \neq \emptyset$, which means that there exists some $c_0 \in \Omega^N$ such that $C_\Omega(c_0, \epsilon)$ covers Ω for all $\epsilon \in]\epsilon_0, (\epsilon_1 + \epsilon_0)/2]$.

2/By contradiction, let us prove that $C_\Omega(c_0, \epsilon_0)$ covers Ω. If $C_\Omega(c_0, \epsilon_0)$ does not cover Ω, then there exists some y in the open set $\Omega \setminus C_\Omega(c_0, \epsilon_0)$, which implies that there exists an open ball $B(y, \eta)$ in $\Omega \setminus C_\Omega(c_0, \epsilon_0)$ for some radius $\eta > 0$. Hence, $r_W(y, c_{0,j}) \geq \eta + \epsilon_0$ for all $j \in \{1, \ldots, N\}$, which means that $C_\Omega(c_0, \epsilon)$ does not cover Ω for the radius $\epsilon = \epsilon_0 + \eta/2$ for instance. This is a contradiction with point 1/ above.

Lemma 1.C.3 (Cantor's intersection theorem) *Suppose that (X, d) is a complete metric space, and C_n is a sequence of non-empty closed nested subsets of X whose diameters tend to zero. Then the intersection of the C_n contains exactly one point, that is $\cap_{k=1}^{\infty} C_k = \{x\}$ for some x in X.*

Verification of the assumptions of Lemma 1.C.3. Since (Ω, r_W) is a complete metric space by definition of a pure graphon, the product space (Ω^N, r_W^{sup}) is also complete for the sup-distance $r_W^{sup}(x, y) := \sup_{1 \leq j \leq N} r_W(x_j, y_j)$ with $x = (x_1, \ldots, x_N), y = (y_1, \ldots, y_N) \in \Omega^N$. By definition of $\widetilde{E}_k$, the sequence $(\widetilde{E}_k)_k$ is composed of nested sets, which are also non-empty since $N_\Omega^{(c)}(\epsilon) = N$ over $]\epsilon_0, \epsilon_1[$. To prove that each $\widetilde{E}_k$ is a closed subset of Ω^N, it is enough to show that $E_\Omega(\epsilon)$ is closed for any $\epsilon \in]\epsilon_0, \epsilon_1[$. Let $(x^k)_{k \geq 0}$ be a sequence in $E_\Omega(\epsilon)$ such that $x^k \to x \in \Omega^N$ as $k \to \infty$. Then, for any $\eta > 0$, there exists some k_0 such that the sup-distance between $x^{k_0} = (x_1^{k_0}, \ldots, x_N^{k_0})$ and $x = (x_1, \ldots, x_N)$ is at most η. As $x^{k_0} \in E_\Omega(\epsilon)$, one know that, for any $y \in \Omega$, there exists some j_0 such that $r_W(y, x_{j_0}^{k_0}) \leq \epsilon$. Thus, using the triangle

inequality, one has for any $\eta > 0$,

$$r_W(y, x_{j_0}) \leq r_W(y, x_{j_0}^{k_0}) + r_W(x_{j_0}^{k_0}, x_{j_0}) \leq \epsilon + \eta$$

which implies that $r_W(y, x_{j_0}) \leq \epsilon$. Hence, $y \in C_\Omega(x, \epsilon)$ for any $y \in \Omega$, which means that $x \in E_\Omega(\epsilon)$. $E_\Omega(\epsilon)$ is therefore a closed subset of Ω^N. All the conditions of Lemma 1.C.3 are checked.

The part of Lemma 1.C.2 on the covering number is proved. For the packing number, the proof is similar and omitted. □

1.C.3 Proof of Lemma 1.3.3: asymptotic density of the sample

Given $\epsilon > 0$, consider a cover of (Ω, r_W) whose cardinality is the integer $N_\Omega^{(c)}(\epsilon/4)$ (written N for brevity) and whose balls are written $B_1, \ldots, B_N$. Let us upper bound the probability that (at least) one of these balls contains zero sampled point ω_i. Using the union bound, this probability is smaller than

$$\sum_{j=1}^{N} \mathbb{P}_{(\Omega,\mu,W)} \{B_j \text{ contains zero sampled point among } \omega_1, \ldots, \omega_n\}$$

which is upper bounded by $N(1-\mu(B_j))^n \leq N(1-\beta)^n$ where $\beta := \min_{j\in[N]} \mu(B_j)$. One has $\beta > 0$ since each ball of a pure graphon has non-zero measure. And as N is not equal to infinity by assumption, this probability tends to zero with n. Thus, with high probability, all balls B_j from the cover contains at least a sampled point. Finally, the asymptotic density of the sample follows from the fact that each ball of radius ϵ of (Ω, r_W) contains a ball B_j from the cover. Lemma 1.3.3 is proved. □

1.D Proofs for the estimation of the neighborhood distance

1.D.1 Proof of Theorem 1.4.1 : the upper bound

Theorem 1.4.1 is a direct consequence of the two following propositions. Proposition 1.D.1 shows the consistency of the inner products between the rows of the adjacency matrix A. That is, $\langle A_i, A_j \rangle_n$ is convergent in probability towards $\langle W(\omega_i, .), W(\omega_j, .) \rangle$ if $i \neq j$. Actually, Proposition 1.D.1 gives a uniform convergence over all $i, j \in [n]$, $i \neq j$.

Proposition 1.D.1 *The following event on inner products*

$$\mathcal{E}_{in} := \left\{ \forall i, j \in [n] : \ |\langle A_i, A_j \rangle_n - \langle W(\omega_i, .), W(\omega_j, .) \rangle| \leq 3\sqrt{\frac{\log n}{n}} \right\}$$

holds with probability $\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{in}) \geq 1 - \frac{2}{n}$ *as soon as* $n \geq 6$.

We have seen that the neighborhood distance r_W can be decomposed into one crossed term and two quadratic terms as follows

$$r_W^2(\omega_i,\omega_j) = \langle W(\omega_i,.),W(\omega_i,.)\rangle + \langle W(\omega_j,.),W(\omega_j,.)\rangle - 2\langle W(\omega_i,.),W(\omega_j,.)\rangle. \qquad (1.26)$$

Proposition 1.D.1 ensures that the crossed term is consistently estimated. Proposition 1.D.2 deals with the quadratic terms $\langle W(\omega_i,.),W(\omega_i,.)\rangle$.

Proposition 1.D.2 *Conditionally to the event $\mathcal{E}_{in}$ (defined above), the following inequalities*

$$\forall i \in [n]: \quad \left|\langle A_i, A_{\widehat{m}(i)}\rangle_n - \langle W(\omega_i,.),W(\omega_i,.)\rangle\right| \leq 3\, r_W(\omega_i,\omega_{m(i)}) \,+\, 15\sqrt{\log(n)/n}$$

hold simultaneously as soon as $n \geq 6$.

The estimation error of (1.26) by our distance estimator

$$\widehat{r}^2(i,j) = \langle A_i, A_{\widehat{m}(i)}\rangle_n + \langle A_j, A_{\widehat{m}(j)}\rangle_n - 2\,\langle A_i, A_j\rangle_n$$

follows directly from Propositions 1.D.1 and 1.D.2. Theorem 1.4.1 is proved. □

Proof of Proposition 1.D.1. By triangle inequality, the expression

$$\Big|\sum_k \frac{A_{ik}A_{kj}}{n} - \int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big|$$

is smaller than

$$\leq \frac{1}{n}\Big|\sum_{k\neq i,j} A_{ik}A_{kj} - (n-2)\int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big|$$
$$+\frac{1}{n}\Big[(A_{ii}+A_{jj})A_{ij} + 2\int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big]$$

which is upper bounded by

$$\leq \frac{1}{n-2}\Big|\sum_{k\neq i,j} A_{ik}A_{kj} - (n-2)\int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big| + \frac{4}{n}.$$

Conditionally to ω_i,ω_j (with $i\neq j$), the $n-2$ random variables $\{A_{ik}A_{kj} : k\in[n], k\neq i,j\}$ are independent with a mean $\mathbb{E}\left[A_{ik}A_{kj}|\omega_i,\omega_j\right] = \int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)$ for all $k\neq i,j$ (where $\mathbb{E}$ is the expectation with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W)}$). It follows from Hoeffding's inequality that

$$\mathbb{P}_{(\Omega,\mu,W)}\left(\frac{1}{n-2}\Big|\sum_{k\neq i,j} A_{ik}A_{kj} - (n-2)\int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big| \geq \epsilon \;\Big|\omega_i,\omega_j\right)$$

is lower than

$$\leq 2\exp\left(-2(n-2)\epsilon^2\right) \leq 2\exp\left(-n\epsilon^2\right)$$

for $\epsilon > 0$ and $n \geq 4$. Since the above inequality is satisfied for almost every $\omega_i, \omega_j \in \Omega$, one has the same upper bound with probability 1 without conditioning. Hence, taking a union bound over all $i \neq j$ one obtain

$$\mathbb{P}_{(\Omega,\mu,W)}\left(\bigcup_{i,j:i\neq j}\left\{\frac{1}{n-2}\Big|\sum_{k\neq i,j} A_{ik}A_{kj} - (n-2)\int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big| \geq \epsilon\right\}\right)$$

lower than

$$\leq 2n^2 \exp\left(-n\epsilon^2\right).$$

Then, setting $\epsilon = \sqrt{\frac{3\log n}{n}}$ gives

$$\mathbb{P}_{(\Omega,\mu,W)}\left(\bigcup_{i,j:i\neq j}\left\{\frac{1}{n-2}\Big|\sum_{k\neq i,j} A_{ik}A_{kj} - (n-2)\int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big| \geq \sqrt{\frac{3\log n}{n}}\right\}\right)$$

smaller than $2/n$.

Combining the above expressions, we get the following inequality

$$\max_{i,j:i\neq j}\Big|\sum_k \frac{A_{ik}.A_{kj}}{n} - \int_\Omega W(\omega_i,z)W(\omega_j,z)\mu(dz)\Big| \leq \sqrt{\frac{3\log n}{n}} + \frac{4}{n} \leq 3\sqrt{\frac{\log n}{n}}$$

with probability at least $1 - \frac{2}{n}$ as soon as $n \geq 6$. □

Proof of Proposition 1.D.2.

$$\begin{aligned}\left|\langle A_i, A_{\widehat{m}(i)}\rangle_n - \langle W(\omega_i,.), W(\omega_i,.)\rangle\right| &\leq \left|\langle A_i, A_{\widehat{m}(i)} - A_{m(i)}\rangle_n\right| \\ &\quad + \left|\langle A_i, A_{m(i)}\rangle_n - \langle W(\omega_i,.), W(\omega_i,.)\rangle\right| \qquad (1.27)\end{aligned}$$

For the second term of the upper bound (1.27),

$$\begin{aligned}\left|\langle A_i, A_{m(i)}\rangle_n - \langle W(\omega_i,.), W(\omega_i,.)\rangle\right| &\leq \left|\langle A_i, A_{m(i)}\rangle_n - \langle W(\omega_i,.), W(\omega_{m(i)},.)\rangle\right| \\ &\quad + \left|\langle W(\omega_i,.), W(\omega_{m(i)},.) - W(\omega_i,.)\rangle\right| \\ &\leq 3\sqrt{\log(n)/n} + r_W(\omega_i, \omega_{m(i)})\end{aligned}$$

by Proposition 1.D.1 and Cauchy-Schwarz inequality. For the first term of the upper bound (1.27), if $\widehat{m}(i) \neq m(i)$,

$$\begin{aligned}\left|\langle A_i, A_{\widehat{m}(i)} - A_{m(i)}\rangle_n\right| &\leq \left|\langle A_i - A_{m(i)}, A_{\widehat{m}(i)}\rangle_n\right| + \left|\langle A_i - A_{\widehat{m}(i)}, A_{m(i)}\rangle_n\right| \\ &\leq \widehat{f}(i, m(i)) + \widehat{f}(i, \widehat{m}(i)) \\ &\leq 2\widehat{f}(i, m(i))\end{aligned}$$

by definition of $\widehat{m}(i)$ and $\widehat{f}$ in (1.9). We upper bound $\widehat{f}(i, m(i))$ as follows.

$$\begin{aligned}\widehat{f}(i,m(i)) := \max_{k\neq i,m(i)}\left|\langle A_k, A_i - A_{m(i)}\rangle_n\right| &\leq \max_{k\neq i,m(i)}\left|\langle W(\omega_k,.), W(\omega_i,.) - W(\omega_{m(i)},.)\rangle\right| \\ &\quad + 2\max_{l,t:\, l\neq t}\left|\langle A_l, A_t\rangle_n - \langle W(\omega_l,.), W(\omega_t,.)\rangle\right| \\ &\leq r_W(\omega_i, \omega_{m(i)}) + 6\sqrt{\log(n)/n}\end{aligned}$$

by Proposition 1.D.1 and Cauchy-Schwarz. Combining the upper bounds on (1.27), Proposition 1.D.2 is proved. □

1.D.2 Proof of Theorem 1.4.2 : the lower bound

Theorem 1.4.2 is a corollary of Theorem 1.D.3 (written below). Let $\mathbf{r}_\omega$ denote the $n \times n$ symmetric matrix with entries $r_W(\omega_i, \omega_j)$, $1 \leq i \leq j \leq n$. Given a real $\delta > 0$, a graphon (Ω, μ, W), a permutation σ of $\{1, \dots, n\}$ and an estimator $\widehat{d}$, we define

$$\mathcal{S}_{(\Omega,\mu,W)}(\widehat{d}, \sigma, \mathbf{r}_\omega) = \Big\{ (i,j) \ : \ 32 \left| \widehat{d}^2(\sigma(i), \sigma(j)) - r_W^2(\omega_i, \omega_j) \right| \ \geq \ 2\,\delta \\ \text{and} \quad 2\,\delta \geq \ r_W(\omega_i, \omega_{m(i)}) + r_W(\omega_j, \omega_{m(j)}) \Big\}$$

and

$$\Phi_{(\Omega,\mu,W)}(\widehat{d}, \mathbf{r}_\omega) = \inf_\sigma \ \text{Card} \quad \mathcal{S}_{(\Omega,\mu,W)}(\widehat{d}, \sigma, \mathbf{r}_\omega) \tag{1.28}$$

where $\Phi_{(\Omega,\mu,W)}(\widehat{d}, \mathbf{r}_\omega)$ is the number of pairs (i,j) where the estimator $\widehat{d}$ is no better than our estimator $\widehat{r}$, roughly speaking. That is, $\Phi_{(\Omega,\mu,W)}(\widehat{d}, \mathbf{r}_\omega)$ counts the pairs (i,j) for which the error of $\widehat{d}$ is larger than the bias of our distance estimator $\widehat{r}$, which is $r_W(\omega_i, \omega_{m(i)}) + r_W(\omega_j, \omega_{m(j)})$ up to some numerical constants. We put an infimum over all permutations σ of the n indices because we consider the problem of recovery of the set of distances $r_W(\omega_i, \omega_j)$, $1 \leq i \leq j \leq n$, regardless of their labeling. According to Theorem 1.D.3, there exists a sequence of graphons (Ω, μ, W_n) such that for any estimator $\widehat{d}$, the quantity $\Phi_{(\Omega,\mu,W_n)}(\widehat{d}, \mathbf{r}_\omega)$ grows linearly with n (on an event of positive probability).

Theorem 1.D.3 *There exists a sequence $(\Omega, \mu, W_n)_{n\geq 0}$ of SBM such that for all $n \geq 10$, all $\delta \in]\sqrt{\frac{8}{n-2}}, 1/40[$ and some numerical constants $c > 0$ and $p > 0$, the following lower bound holds*

$$\inf_{\widehat{d}} \ \mathbb{P}_{(\Omega,\mu,W_n)}\Big[\Phi_{(\Omega,\mu,W_n)}(\widehat{d}, \mathbf{r}_\omega) > cn \Big] \geq p \tag{1.29}$$

where $\inf_{\widehat{d}}$ is the infimum over all estimators.

Theorem 1.4.2 follows from Theorem 1.D.3, choosing $\delta = \left(\sqrt{\frac{\log n}{n}}\right)^{1/(1+\gamma)}$. □

Proof of Theorem 1.D.3. The proof follows the general scheme of reduction for testing two hypotheses [see Yu, 1997, Tsybakov, 2009]. We start with the definition of some SBM with five communities where the latent space Ω is $\{\mathcal{C}_1, \dots, \mathcal{C}_5\}$. We then show that for these SBM, any distance estimator suffers from a large loss.

Let $n \geq 10$ and $\delta \in]\sqrt{8/n-2}, 1/40[$. Consider the symmetric functions $W_n : \{\mathcal{C}_1, \dots, \mathcal{C}_5\}^2 \to \{\mathcal{C}_1, \dots, \mathcal{C}_5\}$ as described in Table 1.1 below. That is, for the two diagonal blocks $\{\mathcal{C}_1, \mathcal{C}_2\}^2$ and $\{\mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}^2$, it is a constant function:

$$W_n(x,y) = \begin{cases} 1/2 & \text{if } (x,y) \in \{\mathcal{C}_1, \mathcal{C}_2\}^2, \\ 1/2 & \text{if } (x,y) \ \in \{\mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}^2, \end{cases}$$

and for the upper right corner block $\{\mathcal{C}_1, \mathcal{C}_2\} \times \{\mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}$:

$$W_n(x,y) = \begin{cases} 1/2 + u_x\sqrt{\delta/2} & \text{if } y \in \mathcal{C}_3, \\ 1/2 + u_x\delta & \text{if } y \in \mathcal{C}_4, \\ 1/2 + u_x/2 & \text{if } y \in \mathcal{C}_5, \end{cases} \qquad u_x = \begin{cases} +1 & \text{if } x \in \mathcal{C}_1, \\ -1 & \text{if } x \in \mathcal{C}_2. \end{cases}$$

The latent space $\{\mathcal{C}_1, \dots, \mathcal{C}_5\}$ is endowed with the probability measure μ defined as follows:

$$\mu(\mathcal{C}_1) = \mu(\mathcal{C}_2) = \frac{1-2\eta}{2}$$

$$\frac{1}{2}\mu(\mathcal{C}_3) = \mu(\mathcal{C}_4) = \mu(\mathcal{C}_5) = \frac{\eta}{2}$$

where $\eta = 2/(n-2)$.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
$\mathcal{C}_1$	$1/2$		$1/2 + \sqrt{\delta/2}$	$1/2 + \delta$	1
$\mathcal{C}_2$			$1/2 - \sqrt{\delta/2}$	$1/2 - \delta$	0
$\mathcal{C}_3$			$1/2$		
$\mathcal{C}_4$					
$\mathcal{C}_5$					

Table 1.1: values of $W_n(\mathcal{C}_i, \mathcal{C}_j)$

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
$\mathcal{C}_1$	$\leq 2\eta$		$\geq \delta/4$	$\leq 5\,\delta^2$	$\geq 1/4$
$\mathcal{C}_2$					

Table 1.2: bounds on $r_W^2(\mathcal{C}_i, \mathcal{C}_j)$

We compute some bounds on the neighborhood distance associated with the above SBM, see Table 1.2 for a summary. These bounds follow easily from the definition (1.4) of the distance. For example,

$$r_W^2(\mathcal{C}_1, \mathcal{C}_3) \geq \int_{\{\mathcal{C}_1, \mathcal{C}_2\}} |W(\mathcal{C}_1, z) - W(\mathcal{C}_3, z)|^2 \ \mu(dz) \geq \big(\mu(\mathcal{C}_1) + \mu(\mathcal{C}_2)\big)\delta/2 \geq (1-2\eta)\delta/2$$

which is larger than $\delta/4$ since $\eta = 2/(n-2)$ and $n \geq 10$.

We now introduce two events $\mathcal{R}_1$ and $\mathcal{R}_2$ on the sampled points $\omega_1, \dots, \omega_n$, which lead to different sets of distances (for r_W), and yet are difficult to decipher for any estimator based on the adjacency matrix A. In addition, we want these two events to happen with a positive probability p that is independent of n. Observe that the union of the two communities $\mathcal{C}_1, \mathcal{C}_2$ have a total weight $1 - 2\eta = 1 - 4/(n-2)$ and thus concentrate most of the probability measure, whereas each of the remaining communities $\mathcal{C}_3, \mathcal{C}_4$, $\mathcal{C}_5$ has a weight of the order of n^{-1}. It follows that most of the sampled points $\omega_1, \dots, \omega_n$ belong to the communities $\mathcal{C}_1, \mathcal{C}_2$ with large probability. In particular, the two following events

$$\begin{aligned} \mathcal{R}_1 &= \Big\{\mathcal{C}_1 \cup \mathcal{C}_2,\ \mathcal{C}_4,\ \mathcal{C}_5 \text{ respectively contain n-2, 1, 1 sampled points}\Big\} \\ \mathcal{R}_2 &= \Big\{\mathcal{C}_1 \cup \mathcal{C}_2,\ \mathcal{C}_3 \text{ respectively contain n-2, 2 sampled points}\Big\} \end{aligned}$$

happen with a positive probability that is independent of n (Lemma 1.D.4).

Lemma 1.D.4 *The probability of each event $\mathcal{R}_1$ and $\mathcal{R}_2$ is lower bounded by some numerical constant $p > 0$:*

$$\mathbb{P}(\mathcal{R}_2) \geq \mathbb{P}(\mathcal{R}_1) \geq p$$

where $\mathbb{P}(\mathcal{R}_k) := \mathbb{P}_{(\Omega,\mu,W_n)}(\mathcal{R}_k) = \int_{(\omega_1,\ldots,\omega_n)\in\{\mathcal{C}_1,\ldots,\mathcal{C}_5\}^n} 1_{\mathcal{R}_k}(\omega_1,\ldots,\omega_n)\,\mathrm{d}\mu(\omega_1)\ldots\mu(\omega_n)$.

One of the interests of the two events $\mathcal{R}_1, \mathcal{R}_2$ is to lead to different sets of distances. Specifically, if $\mathcal{R}_1$ (resp. $\mathcal{R}_2$) holds, the random matrix $\mathbf{r}_\omega = [r_W(\omega_i,\omega_j)]_{i,j\in[n]}$ of distances is denoted by $\mathbf{r}_1 = [r_1(i,j)]_{i,j\in[n]}$ (resp. $\mathbf{r}_2 = [r_2(i,j)]_{i,j\in[n]}$). We measure the difference between both matrices $\mathbf{r}_1$, $\mathbf{r}_2$ of distances as follows:

$$\widetilde{\Phi}(\mathbf{r}_1,\mathbf{r}_2) = \inf_{\sigma} \mathrm{Card}\left\{(i,j) \,:\, \begin{array}{l} 16\left|r_2^2(i,j) - r_1^2(\sigma(i)\,\sigma(j))\right| \geq 2\,\delta \\ r_2(i,m(i)) + r_2(j,m(j)) \leq 2\,\delta \\ r_1(\sigma(i),m(\sigma(i)) + r_1(\sigma(j),m(\sigma(j)) \leq 2\,\delta \end{array}\right\} \tag{1.30}$$

where $\widetilde{\Phi}$ is the number of pairs (i,j) on which $\mathbf{r}_1$ and $\mathbf{r}_2$ are separated by at least the bias of our distance estimator $\widehat{r}$ (up to some numerical constants). Note that this measure is independent of the labeling $i \in \{1\ldots,n\}$ since an infimum is taken over all permutations σ of the n indices. Lemma 1.D.5 ensures that $\mathbf{r}_1$ and $\mathbf{r}_2$ are different enough for a number of pairs (i,j) that is linear with n, regardless of their labeling.

Lemma 1.D.5 *There exists a numerical constant c such that $\widetilde{\Phi}(\boldsymbol{r}_1,\boldsymbol{r}_2) \geq 2c\,n$.*

So far, we have two events of positive probability which lead to two different sets of distances. It remains to see that they are ult to decipher from the observed adjacency matrix A (Lemma 1.D.6). For simplicity, write $\mathbb{P}$ for $\mathbb{P}_{(\Omega,\mu,W)}$ in the following, and $\boldsymbol{\omega}$ the n-tuple $(\omega_1,\ldots,\omega_n)$, and $\{0,1\}^{n\times n}_{sym}$ the set of binary symmetric matrices of size $n\times n$.

Lemma 1.D.6 *For any $M \in \{0,1\}^{n\times n}_{sym}$, one has*

$$\mathbb{P}[A = M|\boldsymbol{\omega}\in\mathcal{R}_1] = \mathbb{P}[A = M|\boldsymbol{\omega}\in\mathcal{R}_2].$$

We now have all the ingredients to lower bound $\mathbb{P}\Big[\Phi_{(\Omega,\mu,W_n)}(\widehat{d},\mathbf{r}_\omega) > cn\Big]$ and prove Theorem 1.D.3. For clarity, $\Phi_{(\Omega,\mu,W_n)}(\widehat{d},\mathbf{r}_\omega)$ is denoted by $\Phi(\widehat{d},\mathbf{r}_\omega)$ in the following. Then, one has

$$\begin{aligned}\mathbb{P}\Big[\Phi(\widehat{d},\mathbf{r}_\omega) > cn\Big] \geq \mathbb{P}\Big[\Phi(\widehat{d},\mathbf{r}_\omega) > cn|\mathcal{R}_1\Big]\mathbb{P}(\mathcal{R}_1) \\ + \mathbb{P}\Big[\Phi(\widehat{d},\mathbf{r}_\omega) > cn|\mathcal{R}_2\Big]\mathbb{P}(\mathcal{R}_2)\end{aligned}$$

By definition of the SBM, the matrix $\mathbf{r}_1$ remains the same for any $\boldsymbol{\omega}\in\mathcal{R}_1$, up to a permutation of the labeling. Combining with the fact that Φ is independent of the labeling, one obtain

that $\Phi(\widehat{d}, \mathbf{r}_1)$ takes a same value for all $\boldsymbol{\omega} \in \mathcal{R}_1$. Similarly, $\Phi(\widehat{d}, \mathbf{r}_2)$ takes the same value for all $\boldsymbol{\omega} \in \mathcal{R}_2$. Hence, the above display says that $\mathbb{P}\Big[\Phi(\widehat{d}, \mathbf{r}_\omega) > cn\Big]$ is larger than

$$\Big(\mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_1) > cn | \mathcal{R}_1\big] + \mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_2) > cn | \mathcal{R}_2\big]\Big)\Big(\mathbb{P}(\mathcal{R}_1) \wedge \mathbb{P}(\mathcal{R}_2)\Big)$$

and since $\mathbb{P}(\mathcal{R}_1) \wedge \mathbb{P}(\mathcal{R}_2) \geq p$ by Lemma 1.D.4, one has

$$\mathbb{P}\Big[\Phi(\widehat{d}, \mathbf{r}_\omega) > cn\Big] \geq \Big(\mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_1) > cn | \mathcal{R}_1\big] + \mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_2) > cn | \mathcal{R}_2\big]\Big)p$$

Now assume that

$$\mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_2) > cn | \mathcal{R}_2\big] \geq \mathbb{P}\big[cn > \Phi(\widehat{d}, \mathbf{r}_1) | \mathcal{R}_1\big]. \tag{1.31}$$

Then, combining the two last inequalities gives

$$\mathbb{P}\Big[\Phi(\widehat{d}, \mathbf{r}_\omega) > cn\Big] \geq p$$

which gives the lower bound of Theorem 1.D.3.

Let us show that (1.31) holds. Lemma 1.D.6 gives

$$\mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_2) > cn | \mathcal{R}_2\big] = \mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_2) > cn | \mathcal{R}_1\big].$$

Then, we use the generalized triangle inequality of Lemma 1.D.7 with $B = \widehat{d}$.

Lemma 1.D.7 *For any $B \in \{0,1\}^{n \times n}_{sym}$, we have $\Phi(B, \boldsymbol{r}_1) + \Phi(B, \boldsymbol{r}_2) \geq \widetilde{\Phi}(\boldsymbol{r}_1, \boldsymbol{r}_2)$.*

That is,

$$\Phi(\widehat{d}, \mathbf{r}_2) \geq \widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2) - \Phi(\widehat{d}, \mathbf{r}_1)$$

which is larger than

$$2cn - \Phi(\widehat{d}, \mathbf{r}_1)$$

by Lemma 1.D.5. Combing the above displays, one has

$$\mathbb{P}\big[\Phi(\widehat{d}, \mathbf{r}_2) > cn | \mathcal{R}_2\big] \geq \mathbb{P}\big[cn > \Phi(\widehat{d}, \mathbf{r}_1) | \mathcal{R}_1\big].$$

The line (1.31) is therefore proved and Theorem 1.D.3 follows. □

We now show the technical lemmas, used in the proof of Theorem 1.D.3.

Proof of Lemma 1.D.4. Let $n \geq 10$. We show that each of the two events $\mathcal{R}_1, \mathcal{R}_2$ occurs with a positive probability that is independent of n. By definition of the events, one has

$$\mathbb{P}(\mathcal{R}_2) \geq \mathbb{P}(\mathcal{R}_1) = \frac{n(n-1)}{2}\left(\frac{\eta}{2}\right)^2 (1-2\eta)^{n-2}$$

which is equal to the following expression for $\eta = 2/(n-2)$,

$$\frac{n(n-2)}{2}\left(\frac{1}{n-2}\right)^2 \exp\left[(n-2)\log\left(1-\frac{4}{n-2}\right)\right]$$

Using $\log(1-x) \geq -x/(1-x)$ for all x in $]0,1[$,

$$\mathbb{P}(\mathcal{R}_1) \geq \exp\left[-\frac{4}{1-\frac{4}{n-2}}\right]$$

which is larger than some positive numerical constant. Hence, Lemma 1.D.4 is proved. □

Proof of Lemma 1.D.5. The proof consists in finding a lower bound of $\widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2)$ that is linear with n. As $\widetilde{\Phi}$ is independent of the labeling of the set of distances $\mathbf{r}_1$ and $\mathbf{r}_2$, one can assume the two following labelings without the loss of generality. For the matrix $\mathbf{r}_1$ (defined on the event $\mathcal{R}_1$), assume the $(n-1)^{\text{th}}$ and n^{th} columns correspond to the two sampled points in $\{\mathcal{C}_4, \mathcal{C}_5\}$. For $\mathbf{r}_2$ (defined on $\mathcal{R}_2$), assume the $(n-1)^{\text{th}}$ and n^{th} columns correspond to the two sampled points in $\mathcal{C}_3$. Accordingly, the $n-2$ first columns of $\mathbf{r}_1$ and $\mathbf{r}_2$ are associated with the sampled points in $\{\mathcal{C}_1, \mathcal{C}_2\}$.

We focus on the $(n-1)^{\text{th}}$ and n^{th} columns of $\mathbf{r}_2$ corresponding to the points in $\mathcal{C}_3$. For the measure $\widetilde{\Phi}$, at least one these two columns will be necessarily compared to one of the $n-1$ first columns of $\mathbf{r}_1$. In other words, the distances associated with a point in $\mathcal{C}_3$ will be compared to the distances associated with a point in $\mathcal{C}_1$,$\mathcal{C}_2$ or $\mathcal{C}_4$. As we can see in Table 1.1 and 1.2, such comparisons will lead to the lower bound $\widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2) \geq n-3$. The corresponding computation are done below, focusing on the two vectors of distances $[r_2(k, n-1)]_{k \leq n-2}$ and $[r_2(k,n)]_{k \leq n-2}$.

By definition, $\widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2)$ is based on the infimum over all permutations. Let σ be any permutation of $\{1, \ldots, n\}$ and prove the lower bound for σ, distinguishing three cases.

Case 1: if $\sigma(n) = n$, then $\sigma(j) \in \{1, \ldots, n-1\}$ for all $j \leq n-1$. For convenience, note $\mathcal{C}_{i,j}$ for a point in $\mathcal{C}_i \cup \mathcal{C}_j$. For all $j \leq n-2$, one has

$$\left| r_2^2(j, n-1) - r_1^2(\sigma(j), \sigma(n-1)) \right| = \left| r_W^2(\mathcal{C}_{1,2}, \mathcal{C}_3) - r_W^2(\mathcal{C}_{1,2,4}, \mathcal{C}_{1,2,4}) \right|$$

according to the chosen labelings (described above). It follows from Table 1.2 that:

$$\left| r_W^2(\mathcal{C}_{1,2}, \mathcal{C}_3) - r_W^2(\mathcal{C}_{1,2,4}, \mathcal{C}_{1,2,4}) \right| \geq \delta/4 - \max(2\eta, 5\delta^2)$$

which is equal to $\delta(1/4 - 5\delta)$ since $\eta = 2/(n-2)$ and $\delta^2 > 8/(n-2)$ by assumption. Hence, using the condition $\delta \leq 1/40$, it is larger than $\delta/8$, so that,

$$16 \left| r_2^2(j, n-1) - r_1^2(\sigma(j), \sigma(n-1)) \right| \geq 2\delta$$

for all $j \leq n-2$.

It remains to upper bound the bias terms by 2δ. The ones related to $\mathbf{r}_2$ are easily obtained: for all $j \leq n$,

$$r_2(j, m(j)) \leq r_W(\mathcal{C}_1, \mathcal{C}_2) \leq 2\eta \leq \delta$$

since on the event $\mathcal{R}_2$, a point ω_j is either in $\mathcal{C}_1 \cup \mathcal{C}_2$ and hence $r_2(j, m(j)) \leq r_W(\mathcal{C}_1, \mathcal{C}_2)$, or in $\mathcal{C}_3$ and thus $r_2(j, m(j)) = 0$ (because its nearest neighbor is in $\mathcal{C}_3$ too). This gives the bounds on the bias terms

$$r_2(i, m(i)) + r_2(j, m(j)) \leq 2\,\delta$$

for all i, j. The corresponding bounds for $\mathbf{r}_1$ are similarly obtained from Table 1.1, but with more calculations. It is therefore encapsulated in the following lemma.

Lemma 1.D.8 *If $\sigma(n) = n$, we have $r_1(\sigma(i), m(\sigma(i)) + r_1(\sigma(j), m(\sigma(j)) \leq 2\delta$ for all $j, i \leq n-1$ such that $i \neq j$.*

Combining the above displays, we obtain the lower bound

$$\mathrm{Card}\left\{ (i,j) \,:\, \begin{array}{l} 16\left|r_2^2(i,j) - r_1^2(\sigma(i)\,\sigma(j))\right| \geq 2\,\delta \\ r_2(i, m(i)) + r_2(j, m(j)) \leq 2\,\delta \\ r_1(\sigma(i), m(\sigma(i)) + r_1(\sigma(j), m(\sigma(j)) \leq 2\,\delta \end{array} \right\} \geq n-3 \tag{1.32}$$

for all permutations fulfilling $\sigma(n) = n$.

Case 2: if $\sigma(n-1) = n$, then $\sigma(n), \sigma(j) \in \{1, \ldots, n-1\}$ for all $j \leq n-2$. Following the same proof as above, we can show that $\left|r_2^2(n,j) - r_1^2(\sigma(n), \sigma(j))\right| \geq 2\delta$ for all $j \leq n-2$. Likewise, the bounds on the bias terms are obtained as before. The inequality (1.32) is therefore proved for all permutations fulfilling $\sigma(n-1) = n$.

Case 3: if $\sigma(n) \neq n$ and $\sigma(n-1) \neq n$. Following the same proof as above, we can show that $\left|r_2^2(n,j) - r_1^2(\sigma(n), \sigma(j))\right| \geq 2\delta$ for all $j \leq n-2$ such that $j \neq \sigma^{-1}(n)$. The inequality (1.32) is therefore proved for all permutations $\sigma(n) \neq n$ and $\sigma(n-1) \neq n$.

Finally, the lower bound (1.32) is true for all permutations σ, in particular for the infimum over all of them. Lemma 1.D.5 is proved. □

Proof of Lemma 1.D.8. Let us upper bound the bias terms for $\mathbf{r}_1$, in the case of an arbitrary permutation σ fulfilling $\sigma(n) = n$. On the event $\mathcal{R}_1$, one has

$$r_1(\sigma(i), m(\sigma(i)) + r_1(\sigma(j), m(\sigma(j)) \leq r_W(\mathcal{C}_1, \mathcal{C}_4) + r_W(\mathcal{C}_1, \mathcal{C}_2).$$

for all $j, i \leq n-1$ such that $j \neq i$. In Table 1.1 and Table 1.2, one observes that

$$\begin{aligned} r_W(\mathcal{C}_1, \mathcal{C}_2) &\leq \sqrt{2\eta} \\ r_W(\mathcal{C}_1, \mathcal{C}_4) &\leq \sqrt{\delta^2(1-2\eta) + (\delta/2)\eta + \delta^2(\eta/2) + (1/4)(\eta/2)}. \end{aligned}$$

The second bound is smaller than $\sqrt{\delta^2(1-(3\eta/2)) + \eta/4}$ since $(\delta/2)\eta \leq (1/4)(\eta/2)$ (using the assumption $\delta \leq 1/40$). Hence,

$$r_1(\sigma(i), m(\sigma(i)) + r_1(\sigma(j), m(\sigma(j)) \leq \sqrt{\delta^2 + \eta/4} + \sqrt{2\eta}$$

which is lower than $\delta + 2\sqrt{\eta}$, and again, lower than 2δ (since $\sqrt{\eta} = \sqrt{2/(n-2)}$ is smaller than $\delta/2$ by assumption). Lemma 1.D.8 is proved. □

Proof of Lemma 1.D.7. Given any matrix $B \in \{0,1\}^{n\times n}_{sym}$, let us show the following inequality $\Phi(B, \mathbf{r}_1) + \Phi(B, \mathbf{r}_2) \geq \widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2)$ where $\widetilde{\Phi}$ and Φ are respectively defined by (1.30) and (1.28).

For all permutations σ of $\{1, \dots, n\}$, the triangle inequality gives

$$\begin{aligned} 2\Big|B_{ij} - r_2^2(i,j)\Big| \vee 2\Big|B_{ij} - r_1^2(\sigma(i)\,\sigma(j))\Big| &\geq \Big|B_{ij} - r_2^2(i,j)\Big| + \Big|B_{ij} - r_1^2(\sigma(i)\,\sigma(j))\Big| \\ &\geq \Big|r_2^2(i,j) - r_1^2(\sigma(i)\,\sigma(j))\Big| \end{aligned}$$

so that $\mathrm{Card}\left\{(i,j) \,:\, \begin{array}{l} 16\Big|r_2^2(i,j) - r_1^2(\sigma(i)\,\sigma(j))\Big| \geq 2\,\delta \\ r_2(i,m(i)) + r_2(j,m(j)) \leq 2\,\delta \\ r_1(\sigma(i), m(\sigma(i))) + r_1(\sigma(j), m(\sigma(j))) \leq 2\,\delta \end{array}\right\}$ lower bounds the sum

of the two cardinal numbers

$$\mathrm{Card}\Big\{(i,j) \;:\; 32\,\big|B_{ij} - r_2^2(i,j)\big| \;\geq\; 2\;\delta \geq\; r_2(i,m(i)) + r_2(j,m(j))\Big\} \quad \text{and}$$

$$\mathrm{Card}\Big\{(i,j) \;:\; 32\,\big|B_{ij} - r_1^2(\sigma(i)\,\sigma(j))\big| \;\geq\; 2\;\delta \geq\; r_1(\sigma(i), m(\sigma(i))) + r_1(\sigma(j), m(\sigma(j)))\Big\}.$$

Taking a permutation that minimizes the latter cardinal, one has

$$\mathrm{Card}\Big\{(i,j) \,:\, 32\,\big|B_{ij} - r_2^2(i,j)\big| \;\geq\; 2\;\delta \geq\; r_2(i,m(i)) + r_2(j,m(j))\Big\} + \Phi(B, \mathbf{r}_1) \geq \widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2)$$

by definition of Φ and $\widetilde{\Phi}$. The above inequality holds for any matrix in $\{0,1\}^{n\times n}_{sym}$, in particular for B^σ defined by $B^\sigma_{ij} = B_{\sigma(i),\sigma(j)}$ (where $B \in \{0,1\}^{n\times n}_{sym}$ and any permutation σ). Using $\Phi(B^\sigma, \mathbf{r}_1) = \Phi(B, \mathbf{r}_1)$, the above display becomes

$$\mathrm{Card}\Big\{(i,j) \,:\, 32\,\big|B^\sigma_{ij} - r_2^2(i,j)\big| \;\geq\; 2\;\delta \geq\; r_2(i,m(i)) + r_2(j,m(j))\Big\} + \Phi(B, \mathbf{r}_1) \geq \widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2)$$

and thus, choosing the permutation that minimize the left term,

$$\Phi(B, \mathbf{r}_1) + \Phi(B, \mathbf{r}_2) \geq \widetilde{\Phi}(\mathbf{r}_1, \mathbf{r}_2).$$

This generalized triangle inequality holds for all $B \in \{0,1\}^{n\times n}_{sym}$. Lemma 1.D.7 is proved. □

Proof of Lemma 1.D.6. In the following, we write $\mathbb{P}$ for $\mathbb{P}_{(\Omega,\mu,W)}$, and $\mu^{\otimes n}$ for the product measure, and $\boldsymbol{\omega}$ for the n-tuple $(\omega_1, \dots, \omega_n)$. Lemma 1.D.6 states that for all $M \in \{0,1\}^{n\times n}_{sym}$,

$$\mathbb{P}[A = M | \boldsymbol{\omega} \in \mathcal{R}_1] = \mathbb{P}[A = M | \boldsymbol{\omega} \in \mathcal{R}_2]$$

which is equivalent to

$$p_{\mathcal{R}_1}(M)/\mathbb{P}(\mathcal{R}_1) = p_{\mathcal{R}_2}(M)/\mathbb{P}(\mathcal{R}_2) \tag{1.33}$$

where $p_{\mathcal{R}_1}(M)$ denotes

$$p_{\mathcal{R}_1}(M) := \mathbb{P}\left(\{A = M\} \cap \mathcal{R}_1\right) = \int_{\boldsymbol{\omega} \in \mathcal{R}_k} \mathbb{P}(A = M|\boldsymbol{\omega})\, \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega}).$$

Hence, we want to prove that

$$2p_{\mathcal{R}_1}(M) = p_{\mathcal{R}_2}(M)$$

since $2\mathbb{P}(\mathcal{R}_1) = \mathbb{P}(\mathcal{R}_2)$ by definition of the events $\mathcal{R}_1$ and $\mathcal{R}_2$.

Let $\mathcal{R}_1(k,l)$ be the the event defined by $\mathcal{R}_1 \cap \{(\omega_k, \omega_l) \in \mathcal{C}_4 \times \mathcal{C}_5\}$. Thus, the event $\mathcal{R}_1$ is the union $\cup_{1 \le k \ne l \le n} \mathcal{R}_1(k,l)$. For any matrix $M = [M_{ij}]_{i,j \le n}$ in $\{0,1\}^{n \times n}_{sym}$,

$$p_{\mathcal{R}_1}(M) = \int_{\boldsymbol{\omega} \in \mathcal{R}_1} \mathbb{P}(A = M|\boldsymbol{\omega})\, \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega}) = \sum_{1 \le k \ne l \le n} \int_{\boldsymbol{\omega} \in \mathcal{R}_1(k,l)} \mathbb{P}(A = M|\boldsymbol{\omega})\, \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega}).$$

Given a permutation σ of $\{1, \dots, n\}$, denote by M^σ the matrix $M^\sigma_{ij} = M_{\sigma(i),\sigma(j)}$ with $i, j \in \{1, \dots, n\}$. Write σ_{kl} for a permutation fulfilling $\sigma(n-1) = k$ and $\sigma(n) = l$. Then, the probability $p_{\mathcal{R}_1}(M)$ is equal to

$$\sum_{1 \le k \ne l \le n} \int_{\boldsymbol{\omega} \in \mathcal{R}_1(k,l)} \mathbb{P}(A^{\sigma_{kl}} = M^{\sigma_{kl}}|\boldsymbol{\omega}) \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega}) = \sum_{1 \le k \ne l \le n} \int_{\boldsymbol{\omega} \in \mathcal{R}_1(n-1,n)} \mathbb{P}(A = M^{\sigma_{kl}}|\boldsymbol{\omega}) \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega}).$$

Conditionally to $\boldsymbol{\omega}$, the entries of A for $i < j$ are independent Bernoulli variables, so that

$$p_{\mathcal{R}_1}(M) = \sum_{1 \le k \ne l \le n} \int_{\boldsymbol{\omega} \in \mathcal{R}_1(n-1,n)} \prod_{1 \le i < j \le n} \mathbb{P}(A_{ij} = M^{\sigma_{kl}}_{ij}|\omega_i, \omega_j) \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega}).$$

On the event $\mathcal{R}_1(n-1,n)$, the $\omega_1, \dots, \omega_{n-2}$ are in $\mathcal{C}_1 \cup \mathcal{C}_1$, and (ω_{n-1}, ω_n) are in $\mathcal{C}_4 \times \mathcal{C}_5$. As the function W_n of the SBM is equal to $1/2$ on the diagonal blocks $\{\mathcal{C}_1, \mathcal{C}_2\}^2$ and $\{\mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}^2$, one has $\mathbb{P}(A_{ij} = M^{\sigma_{kl}}_{ij}|\omega_i, \omega_j) = \frac{1}{2}$ for all (i,j) in the set $\{(i,j) : i < j \le n-2\} \cup \{(n-1, n)\}$ of cardinality $g_n = n(n-1)/2 - 2(n-2)$. Hence, the probability $p_{\mathcal{R}_1}(M)$ is equal to

$$\sum_{1 \le k \ne l \le n} \left(\frac{1}{2}\right)^{g_n} \int_{\boldsymbol{\omega} \in \mathcal{R}_1(n-1,n)} \prod_{1 \le i \le n-2} \mathbb{P}(A_{i,n-1} = M^{\sigma_{kl}}_{i,n-1}|\omega_i, \omega_{n-1}) \mathbb{P}(A_{i,n} = M^{\sigma_{kl}}_{i,n}|\omega_i, \omega_n) \mathrm{d}\mu^{\otimes n}(\boldsymbol{\omega})$$

or equivalently to

$$\sum_{1 \le k \ne l \le n} \left(\frac{1}{2}\right)^{g_n} \int_{(\omega_{n-1}, \omega_n) \in \mathcal{C}_4 \times \mathcal{C}_5} X_{M^{\sigma_{kl}}}(\omega_{n-1}, \omega_n) \mathrm{d}\mu^{\otimes 2}(\omega_{n-1}, \omega_n)$$

with

$$X_{M^{\sigma_{kl}}}(\omega_{n-1}, \omega_n) := \prod_{1 \le i \le n-2} \int_{\omega_i \in \mathcal{C}_1 \cup \mathcal{C}_2} \mathbb{P}(A_{i,n-1} = M^{\sigma_{kl}}_{i,n-1}|\omega_i, \omega_{n-1}) \mathbb{P}(A_{i,n} = M^{\sigma_{kl}}_{i,n}|\omega_i, \omega_n)\, \mathrm{d}\mu(\omega_i).$$

Likewise, $\mathcal{R}_2$ is the union $\cup_{1 \le k < l \le n} \mathcal{R}_2(k,l)$ where each $\mathcal{R}_2(k,l)$ is the event $\mathcal{R}_2 \cap \{\omega_k, \omega_l \in \mathcal{C}_3\}$. Following the same proof as for $\mathcal{R}_1$, one can show that

$$p_{\mathcal{R}_2}(M) = \sum_{1 \le k < l \le n} \left(\frac{1}{2}\right)^{g_n} \int_{(\omega_{n-1}, \omega_n) \in \mathcal{C}_3 \times \mathcal{C}_3} X_{M^{\sigma_{kl}}}(\omega_{n-1}, \omega_n) \mathrm{d}\mu^{\otimes 2}(\omega_{n-1}, \omega_n).$$

Lemma 1.D.9 *There exists a constant $X_{M^{\sigma_{kl}}}$ such that $X_{M^{\sigma_{kl}}}(\omega_{n-1},\omega_n) = X_{M^{\sigma_{kl}}}$ whether $\mathcal{R}_1(n-1,n)$ or $\mathcal{R}_2(n-1,n)$ holds.*

Using Lemma 1.D.9, one has

$$p_{\mathcal{R}_1}(M) = \left(\frac{1}{2}\right)^{g_n} \sum_{1\le k\neq l\le n} X_{M^{\sigma_{kl}}}\,\mu(\mathcal{C}_4)\,\mu(\mathcal{C}_5)$$

and

$$p_{\mathcal{R}_2}(M) = \left(\frac{1}{2}\right)^{g_n} \sum_{1\le k<l\le n} X_{M^{\sigma_{kl}}}\,\mu(\mathcal{C}_3)^2$$

so that $p_{\mathcal{R}_1}(M) = p_{\mathcal{R}_2}(M)/2$, since $\mu(\mathcal{C}_4) = \mu(\mathcal{C}_5) = \mu(\mathcal{C}_3)/2$ (by construction of the SBM). Lemma 1.D.6 is proved. □

Proof of Lemma 1.D.9. For brevity, write $\mathbb{P}$ for $\mathbb{P}_{(\Omega,\mu,W)}$ in the following. By definition, $X_{M^{\sigma_{kl}}}(\omega_{n-1},\omega_n)$ is the product of the $n-2$ following terms

$$\int_{\omega_i\in\mathcal{C}_1\cup\mathcal{C}_2} \mathbb{P}(A_{i,n-1} = M^{\sigma_{kl}}_{i,n-1}|\omega_i,\omega_{n-1})\,\mathbb{P}(A_{i,n} = M^{\sigma_{kl}}_{i,n}|\omega_i,\omega_n)\,\mathrm{d}\mu(\omega_i)$$

$i = 1,\dots,n-2$. The above display is equal to

$$\begin{aligned}&\int_{\omega_i\in\mathcal{C}_1} \mathbb{P}(A_{i,n-1} = M^{\sigma_{kl}}_{i,n-1}|\omega_i,\omega_{n-1})\mathbb{P}(A_{i,n} = M^{\sigma_{kl}}_{i,n}|\omega_i,\omega_n)\mathrm{d}\mu(\omega_i)\\ &+\int_{\omega_i\in\mathcal{C}_2} \mathbb{P}(A_{i,n-1} = M^{\sigma_{kl}}_{i,n-1}|\omega_i,\omega_{n-1})\mathbb{P}(A_{i,n} = M^{\sigma_{kl}}_{i,n}|\omega_i,\omega_n)\mathrm{d}\mu(\omega_i).\end{aligned}$$

If $(\omega_{n-1},\omega_n)\in\mathcal{C}_4\times\mathcal{C}_5$, then

$$\begin{aligned}&= \int_{\omega_i\in\mathcal{C}_1} [1/2 + (2M^{\sigma_{kl}}_{i,n-1} - 1)\,\delta]\,[1/2 + (2M^{\sigma_{kl}}_{i,n} - 1)\,(1/2)]\,\mathrm{d}\mu(\omega_i)\\ &+\int_{\omega_i\in\mathcal{C}_2} [1/2 - (2M^{\sigma_{kl}}_{i,n-1} - 1)\,\delta]\,[1/2 - (2M^{\sigma_{kl}}_{i,n} - 1)\,(1/2)]\,\mathrm{d}\mu(\omega_i)\end{aligned}$$

which is equal to $\left[1/2 + (2M^{\sigma_{kl}}_{i,n-1} - 1)(2M^{\sigma_{kl}}_{i,n} - 1)\delta\right]\mu(\mathcal{C}_1)$, since $\mu(\mathcal{C}_1) = \mu(\mathcal{C}_2)$.
If $(\omega_{n-1},\omega_n)\in\mathcal{C}_3\times\mathcal{C}_3$, then

$$\begin{aligned}&= \int_{\omega_i\in\mathcal{C}_1} [1/2 + (2M^{\sigma_{kl}}_{i,n-1} - 1)\,\sqrt{\delta/2}][1/2 + (2M^{\sigma_{kl}}_{i,n} - 1)\,\sqrt{\delta/2}]\,\mathrm{d}\mu(\omega_i)\\ &+\int_{\omega_i\in\mathcal{C}_2} [1/2 - (2M^{\sigma_{kl}}_{i,n-1} - 1)\,\sqrt{\delta/2}][1/2 - (2M^{\sigma_{kl}}_{i,n} - 1)\,\sqrt{\delta/2}]\,\mathrm{d}\mu(\omega_i)\end{aligned}$$

which is equal to $\left[1/2 + (2M^{kl}_{i,n-1} - 1)(2M^{\sigma_{kl}}_{i,n} - 1)\delta\right]\mu(\mathcal{C}_1)$.

Hence $X_{M^{\sigma_{kl}}}(\omega_{n-1},\omega_n)$ is equal to the same constant whether (ω_{n-1},ω_n) belongs to $\mathcal{C}_3\times\mathcal{C}_3$ or $\mathcal{C}_4\times\mathcal{C}_5$. Lemma 1.D.9 is proved. □

1.E Proofs for the estimation of the Minkowski dimension

1.E.1 Proof of Theorem 1.4.4: the upper bound

Theorem 1.4.4 is a corollary of Theorem 1.E.1, which gives non-asymptotic high-probability bounds for the risk of the data-function $-\log \widehat{N}_{\Omega}^{(c)}(\epsilon)/\log \epsilon$.

Theorem 1.E.1 *Assume the graphon (Ω, μ, W) satisfies $H_1^{\alpha,v}$ and $H_2^{m,M,v}$ and has a Minkowski dimension $d \in]0, \infty[$. If n is large enough to satisfy the below inequality*

$$2\log n/n \leq \alpha \, (v/14)^{2d} \wedge (v/14)^4 \, ,$$

then the following holds with probability at least $1-(2+4\alpha M)/n$ with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W)}$. The sum of the distance error bound (1.14) and the sampling error (1.13) is upper bounded as follows

$$\frac{b_{sup} + s_\omega}{6} \leq err_{n,d} := \left(\frac{\log n}{n}\right)^{1/4} + \left(\frac{2\log n}{\alpha \, n}\right)^{1/2d} . \tag{1.34}$$

For all $\epsilon \in]2(b_{sup} + s_\omega), v/7]$, the covering number estimator $\widehat{N}_{\Omega}^{(c)}(\epsilon)$ satisfies the following upper bound

$$\left| \frac{\log \widehat{N}_{\Omega}^{(c)}(\epsilon)}{-\log \epsilon} - d \right| \leq \frac{1}{-\log \epsilon} \left[\log \left(M \vee \frac{1}{m} \right) + 6d \frac{err_{n,d}}{\epsilon} \left(1 + \frac{err_{n,d}}{\epsilon} \right) \right]$$

Theorem 1.4.4 follows from Theorem 1.E.1 by choosing any radius radius that minimizes the above upper bound, that is, any radius ϵ_D of the order of $\sup_{\{d:\, d \leq D\}} err_{n,d} = err_{n,D}$. □

Comments on Theorem 1.E.1 : We first remark that the above theorem based on the covering number can also be adapted to the packing number (without ulties). We now comment on the two additive error terms in the upper bound. The term $-\log \left(M \vee (1/m)\right) / \log \epsilon$ stands for the gap between the Minkowski dimension and the quantity that we actually estimate, i.e. $-\log N_{\Omega}^{(c)}(\epsilon)/\log \epsilon$. This gap depends on the parameters of the assumption $H_2^{m,M,v}$. The second error term $-d \, err_{n,d}/(\epsilon \log \epsilon)$ represents the gap between the latter estimated quantity and the estimator $-\log \widehat{N}_{\Omega}^{(c)}(\epsilon)/\log \epsilon$. To control this gap, we need to estimate the covering number correctly, and thus to control the error sum $b_{sup} + s_\omega$ involved in Proposition 1.4.3. Actually, the theorem ensures that this error sum is smaller than $err_{n,d}$. This comes from the fact that the difference between the sample $\omega_1, \ldots, \omega_n$ and the latent space Ω is not too large, thanks to the assumption $H_1^{\alpha,v}$. See the proof below for details.

Finally, the upper bound holds with probability at least $1 - 2/n - 4\,\alpha\, M/n$. The first quantity $2/n$ corresponds to the event $\mathcal{E}_{dist}^c$ defined in Theorem 1.4.1, i.e. that the the distance estimator does not satisfy the distance error bound b_{sup}. The second quantity $4\,\alpha\, M/n$ corresponds to the probability of the event where the sampled points do not cover well the latent space, leading to a large sampling error s_ω and a large distance error bound b_{sup}. This event, denoted by $\mathcal{E}_{bad}$, is rigorously defined in the following proof.

Proof of Theorem 1.E.1. Assume the event $\mathcal{E}_{dist}$ of Theorem 1.4.1 holds, that is the errors of distance-estimator are uniformly bounded by b_{sup}. On this event, Proposition 1.4.3 gives

$$N_\Omega^{(c)}\left(\epsilon + b_{sup} + s_\omega\right) \leq \widehat{N}_\Omega^{(c)}\left(\epsilon\right) \leq N_\Omega^{(c)}\left(\epsilon - b_{sup} - s_\omega\right)$$

for all $\epsilon \in]b_{sup} + s_\omega, 1[$, so that

$$\frac{\log N_\Omega^{(c)}(\epsilon + s_\omega + b_{sup})}{-\log\epsilon} - d \leq \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon)}{-\log\epsilon} - d \leq \frac{\log N_\Omega^{(c)}(\epsilon - s_\omega - b_{sup})}{-\log\epsilon} - d.$$

As the assumption $H_2^{m,M,v}$ is valid in the neighborhood $]0, v]$, we need to check that $\epsilon + s_\omega + b_{sup} \in]0, v]$ to use this assumption. For clarity, we do this verification at the end of the proof. Hence, using $H_2^{m,M,v}$, one has

$$\frac{\log m}{-\log\epsilon} - d\left[\frac{\log(\epsilon + s_\omega + b_{sup})}{-\log\epsilon} + 1\right] \leq \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon)}{-\log\epsilon} - d \leq \frac{\log M}{-\log\epsilon} - d\left[\frac{\log(\epsilon - s_\omega - b_{sup})}{-\log\epsilon} + 1\right] \tag{1.35}$$

In the right hand side of (1.35), the right term is upper bounded by

$$-d\left[\frac{\log(\epsilon - s_\omega - b_{sup})}{-\log\epsilon} + 1\right] \leq -d\,\frac{\log\big(1 - (s_\omega + b_{sup})/\epsilon\big)}{-\log\epsilon}$$

which is again upper bounded by

$$d\,\frac{(s_\omega + b_{sup})/\epsilon + \left((s_\omega + b_{sup})/\epsilon\right)^2}{-\log\epsilon}$$

if $(s_\omega + b_{sup}) \leq \epsilon/2$. Similarly in the left hand side of (1.35), the right term is lower bounded by

$$-d\left[\frac{\log(\epsilon + s_\omega + b_{sup})}{-\log\epsilon} + 1\right] \geq -d\frac{(s_\omega + b_{sup})/\epsilon}{-\log\epsilon}.$$

Combining the above displays, one derive

$$\frac{\log m}{-\log\epsilon} - d\frac{(s_\omega + b_{sup})/\epsilon}{-\log\epsilon} \leq \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon)}{-\log\epsilon} - d \leq \frac{\log M}{-\log\epsilon} + d\,\frac{(s_\omega + b_{sup})/\epsilon + \left((s_\omega + b_{sup})/\epsilon\right)^2}{-\log\epsilon}. \tag{1.36}$$

It remains to upper bound the error sum $s_\omega + b_{sup}$ in (1.36). Given a cover of Ω, composed of $N_\Omega^{(c)}(\eta)$ balls B_j of radius η, one define the following event

$$\mathcal{E}_{bad}(\eta) := \Big\{\exists j : B_j \text{ contains exactly 0 or 1 sampled point among } \omega_1, \ldots, \omega_n\Big\}. \tag{1.37}$$

Assume the complementary event $\mathcal{E}_{bad}^c(\eta)$ holds. This means that each ball of the cover of Ω contains at least two sampled points. Hence, one has

$$\begin{aligned} &s_\omega \leq 2\eta, \\ &\sup_{i\in\{1,..,n\}} r_W(\omega_i, \omega_{m(i)}) \leq 2\eta. \end{aligned}$$

which directly implies the following upper bound

$$b_{sup} + s_\omega \leq 6\left(\frac{\log n}{n}\right)^{1/4} + 4\sqrt{\eta} + 2\eta$$

by definition of b_{sup} in (1.14). Thus, for the particular radius $\eta_n := [2\log(n)/(\alpha\, n)]^{1/d}$,

$$b_{sup} + s_\omega \leq 6\left(\frac{\log n}{n}\right)^{1/4} + 6\left(\frac{2\log n}{\alpha\, n}\right)^{1/2d}.$$

It follows from the definition (1.34) of $err_{n,d}$ that

$$s_\omega + b_{sup} \leq 6 err_{n,d}.$$

Combining the above upper bound with (1.36), one deduce the inequalities of the theorem.

The above displays hold conditionally to the event $\mathcal{E}^c_{bad}(\eta_n) \cap \mathcal{E}_{dist}$, which happens with probability at least $1 - (2 + 4\alpha M)/n$ (Lemma 1.E.2).

Lemma 1.E.2 *The probability $\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n) \cup \mathcal{E}^c_{dist})$ is smaller than $(2 + 4\alpha M)/n$.*

The condition $\epsilon + s_\omega + b_{sup} \in]0, v]$ (used at the beginning of the proof) is satisfied (Lemma 1.E.3).

Lemma 1.E.3 *On the event $\mathcal{E}^c_{bad}(\eta_n) \cap \mathcal{E}_{dist}$, one has $\epsilon + s_\omega + b_{sup} \in]0, v]$.*

Theorem 1.E.1 is proved. □

Proof of Lemma 1.E.3. We want to prove that $\epsilon + s_\omega + b_{sup} \in]0, v]$ on the event $\mathcal{E}^c_{bad}(\eta_n) \cap \mathcal{E}_{dist}$. We have already seen that $s_\omega + b_{sup} \leq 6 err_{n,d}$ on this event, so it is enough to prove that $\epsilon + 6 err_{n,d} \leq v$. By assumption in Theorem 1.E.1, one has

$$2\frac{\log n}{n} \leq \alpha\left(\frac{v}{14}\right)^{2d} \wedge \left(\frac{v}{14}\right)^4,$$

which implies

$$\left(\frac{2\log n}{\alpha\, n}\right)^{1/2d} \leq v/14 \quad \text{and} \quad \left(\frac{\log n}{n}\right)^{1/4} \leq v/14,$$

and thus

$$err_{n,d} \leq v/7.$$

Finally, one has $\epsilon + 6 err_{n,d} \leq v$, since $\epsilon \leq v/7$ by assumption. Hence, $\epsilon + s_\omega + b_{sup} \in]0, v]$ on the event $\mathcal{E}^c_{bad}(\eta_n) \cap \mathcal{E}_{dist}$. The lemma is proved. □

Proof of Lemma 1.E.2. Let us upper bound the probability $\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n) \cup \mathcal{E}^c_{dist})$. The union bound gives

$$\begin{aligned}\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n) \cup \mathcal{E}^c_{dist}) &\leq \mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n)) + \mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}^c_{dist}) \\ &\leq \mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n)) + \frac{2}{n}\end{aligned}$$

where the last inequality comes from Theorem 1.4.1. If the cover defined by $\mathcal{E}_{bad}(\eta_n)$ satisfies the condition (1.39), then Lemma 1.E.4 ensures that

$$\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n)) \leq 2N_{\Omega}^{(c)}(\eta_n)n\beta\exp[-\beta(n-1)]. \tag{1.38}$$

Lemma 1.E.4 *Let $B_1, \ldots, B_N$ be N balls in (Ω, r_W) of measure (strictly) larger than $1/n$, that is,*

$$\min_{j\leq N} \mu(B_j) \geq \beta > 1/n \tag{1.39}$$

for some real β. Then the probability that (at least) one ball contains exactly zero or one sampled point is smaller than

$$2Nn\beta\exp[-\beta(n-1)].$$

Assume that $\eta_n \in]0, v]$ to use the assumption $H_1^{\alpha,v}$. Then, one obtain the following lower bound for the cover defined by $\mathcal{E}_{bad}(\eta_n)$,

$$\mu(B_j) \geq \alpha\ \eta_n^d = 2\log(n)/n$$

so that assumption (1.39) is satisfied. Applying Lemma 1.E.4 for $\beta = \alpha\ \eta_n^d$, one has

$$\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n)) \leq 2N_{\Omega}^{(c)}(\eta)n\alpha\ \eta^d\exp\left[-\alpha\ \eta^d(n-1)\right].$$

Combining with the inequality $N_{\Omega}^{(c)}(\eta) \leq M\eta^{-d}$ from assumption $H_2^{m,M,v}$, one derive

$$\mathbb{P}_{(\Omega,\mu,W)}(\mathcal{E}_{bad}(\eta_n)) \leq 2Mn\alpha\ \exp\left[-\alpha\ \eta^d(n-1)\right]$$

and since $\alpha\,\eta_n^d = 2\log(n)/n$, one obtain the upper bound

$$2Mn\alpha\ \exp\left[-\frac{2\log n}{n}(n-1)\right].$$

The above display is finally smaller than

$$4Mn\alpha\exp\left[-2\log n\right] \leq (4M\alpha)/n.$$

To conclude the proof, it remains to check the condition $\eta_n \in]0, v]$ that we assume earlier. The following assumption of Theorem 1.E.1

$$2\frac{\log n}{n} \leq \alpha\left(\frac{v}{14}\right)^{2d} \wedge \left(\frac{v}{14}\right)^4$$

ensures that the radius $\eta_n = [2\log(n)/(\alpha\,n)]^{1/d}$ satisfies the condition $\eta_n \in]0, v]$. Lemma 1.E.2 is proved. □

Proof of Lemma 1.E.4. Given N balls $B_1, \ldots, B_N$, let us upper bound the probability that (at least) one of the balls contains exactly zero or one sampled point ω_i. With the union bound, this probability is lower than

$$\sum_{j=1}^{N} \mathbb{P}_{(\Omega,\mu,W)}\left\{B_j \text{ contains exactly 0 or 1 sampled point among } \omega_1, \ldots, \omega_n\right\}$$

which is again upper bounded with the union bound by

$$\sum_{j=1}^{N} \mathbb{P}_{(\Omega,\mu,W)} \{B_j \text{ contains exactly 0 point}\} + \sum_{j=1}^{N} \mathbb{P}_{(\Omega,\mu,W)} \{B_j \text{ contains exactly 1 point}\}.$$

Since the probability of the event $\{B_j$ contains exactly 0 point$\}$ is equal to $(1-\mu(B_j))^n$, and since the probability of $\{B_j$ contains exactly 1 point$\}$ is $n\mu(B_j)(1-\mu(B_j))^{n-1}$, the above sum is upper bounded by

$$\sum_{j=1}^{N}(1-\mu(B_j))^n + \sum_{j=1}^{N} n\mu(B_j)(1-\mu(B_j))^{n-1}.$$

Combining the assumption $\mu(B_j) \geq \beta > 1/n$ with the monotonicity of the functions $x \mapsto (1-x)^n$ and $x \mapsto nx(1-x)^{n-1}$ on $]1/n, 1[$, one has the following upper bound

$$N\left[(1-\beta)^n + n\beta(1-\beta)^{n-1}\right]$$

which is lower than $2Nn\beta(1-\beta)^{n-1} \leq 2Nn\beta \exp[-\beta(n-1)]$. Lemma 1.E.4 is proved. □

1.E.2 Lower bound and minimal conditions

Proof of Theorem 1.4.5. From [Falconer, chap.2], we deduce directly the following lemma.

Lemma 1.E.5 *Given $L > 1$ and $n \geq 2$, there exists a set $\Omega_0 \subset]0, 1/(Ln)[\times]0, 1/(Ln)[$ with Minkowski dimension $d_2 = 1+\log^{-1}(n)$ w.r.t the Euclidean distance of $[0,1]^2$, and a probability measure μ_0 on Ω_0.*

Based on (Ω_0, μ_0) described in Lemma 1.E.5, we construct two graphons that are ult to distinguish for any estimator.

- $\Omega_1 =]0, 1[\times\{0\} \subset [0,1]^2$ endowed with the uniform measure λ on $]0, 1[$. In particular, $\lambda(]0, 1[\times\{0\}) = 1$.
- $\Omega_2 = \Omega_1 \cup \Omega_0 \subset [0,1]^2$ endowed with the probability measure:

$$\mu_2 = (1-n^{-1})\lambda + n^{-1}\mu_0.$$

Consider a symmetric function $W : [0,1]^2 \times [0,1]^2 \to [0,1]$ satisfying a double Hölder condition (1.6) with Hölder exponent $\alpha = 1$. Then, Appendix 1.A.2 shows that the neighborhood distance (associated with such a W) behaves like the euclidean distance on $[0,1]^2$, i.e.:

$$r_W(\omega, \omega') \asymp ||\omega - \omega||_2$$

for all $\omega, \omega' \in [0,1]^2$. Hence, (Ω_1, λ, W) and (Ω_2, μ, W) satisfy $dim\,\Omega_2 = 1 + \log^{-1}(n)$ and $dim\,\Omega_1 = 1$, respectively. For brevity, we denote these dimensions by d_2 and d_1 in the following.

Let us check that all conditions of $\mathcal{W}_n(D, \alpha, m, M, v)$ are satisfied by both graphons (Ω_1, λ, W) and (Ω_2, μ_2, W). It is clear that (Ω_1, λ, W) belongs to the set $\mathcal{W}_n(D, \alpha, m, M, v)$ for large enough M and small enough α, m. For the graphon (Ω_2, μ_2, W), one has:

- Assumption $H_1^{\alpha,v}$: for any point $\omega \in \Omega_0$, note $\omega_{proj} \in \Omega_1$ its closest point in Ω_1. As $\Omega_0 \subset]0, 1/(Ln)[^2$, we have $r_W(\omega, \omega_{proj}) \leq 1/(2n)$ for large enough L. Then, for all $\epsilon > 1/n$ and all $\omega \in \Omega_0$, one has

$$\begin{aligned}\mu_2\left[B(\omega,\epsilon)\right] &\geq (1-n^{-1})\lambda\left[B(\omega,\epsilon)\right] \\ &\geq (1-n^{-1})\lambda\left[B(\omega_{proj},\epsilon-1/(2n))\right] \\ &\geq \frac{1}{2}\lambda\left[B(\omega_{proj},\epsilon/2)\right].\end{aligned}$$

  which is larger than ϵ (up to a numerical constant) since (Ω_1, λ, W) satisfies the condition $H_1^{\alpha,v}$ for all $\epsilon > 0$.

- Assumption $H_2^{m,M,v}$ lower bound: $N^{(c)}_{\Omega_2}(\epsilon) \gtrsim N^{(c)}_{\Omega_1}(\epsilon) \gtrsim \epsilon^{-d_1}$ which is larger than $\epsilon^{-d_2+\log^{-1}(n)} \gtrsim \epsilon^{-d_2}$ because $\epsilon^{\log^{-1}(n)} \asymp 1$ for all $\epsilon \in]1/n, 1[$.

- Assumption $H_2^{m,M,v}$ upper bound: $N^{(c)}_{\Omega_2}(\epsilon) \lesssim N^{(c)}_{\Omega_1}(\epsilon) + N^{(c)}_{\Omega_0}(\epsilon) \lesssim N^{(c)}_{\Omega_1}(\epsilon)$ since $N^{(c)}_{\Omega_0}(\epsilon) \lesssim N^{(c)}_{\Omega_0}(1/n) = 1$ for $\epsilon > 1/n$ and large enough L. Combining with the fact that (Ω_1, λ, W) satisfies $H_2^{m,M,v}$, one obtain $N^{(c)}_{\Omega_2}(\epsilon) \lesssim \epsilon^{-d_1}$.

Thus, both graphons (Ω_1, λ, W) and (Ω_2, μ_2, W) fulfill all conditions of $\mathcal{W}_n(D, \alpha, m, M, v)$ for large enough constants L, M and small enough constants α, m.

We define the event $\mathcal{E}_{\Omega_1}$ where the i.i.d. sample $\omega_1, \ldots, \omega_n$ is such that all points $\omega_1, \ldots, \omega_n$ belong to Ω_1. In particular, for the graphon (Ω_2, μ_2, W), the probability of this event is larger than

$$\mu_2[\mathcal{E}_{\Omega_1}] \geq (1-n^{-1})^n \geq \frac{1}{3}.$$

Then, for any estimator $\hat{d}$ based on the adjacency matrix A, one has

$$\begin{aligned}&\mathbb{P}_{(\Omega_2,\mu_2,W)}\left[|\hat{d} - dim\,\Omega_2| \geq \frac{1}{2}\log^{-1}(n)\right] \\ &\qquad \geq \mathbb{P}_{(\Omega_2,\mu_2,W)}\left[|\hat{d} - dim\,\Omega_2| \geq \frac{1}{2}\log^{-1}(n)\middle|\mathcal{E}_{\Omega_1}\right]\mu_2(\mathcal{E}_{\Omega_1})\end{aligned}$$

which is larger than

$$\frac{1}{3}\mathbb{P}_{(\Omega_1,\lambda,W)}\left[|\hat{d} - dim\,\Omega_1| \leq \frac{1}{2}\log^{-1}(n)\right]$$

since $|\hat{d} - dim\,\Omega_1| \leq \frac{1}{2}\log^{-1}(n)$ implies $|\hat{d} - dim\,\Omega_2| \geq \frac{1}{2}\log^{-1}(n)$. Thus, by writting

$$p := \mathbb{P}_{(\Omega_1,\lambda,W)}\left[|\hat{d} - dim\,\Omega_1| > \frac{1}{2}\log^{-1}(n)\right],$$

the above displays entail

$$\mathbb{P}_{(\Omega_2,\mu_2,W)}\left[|\hat{d} - dim\,\Omega_2| \geq \frac{1}{2}\log^{-1}(n)\right] \geq \frac{1}{3}(1-p)$$

which imply that

$$\sup_{\mathcal{W}_n(D,\alpha,m,M,v)} \mathbb{P}_{(\Omega,\mu,W)}\Big[|\hat{d} - dim\,\Omega| \geq \frac{1}{2}\log^{-1}(n)\Big]$$
$$\geq \max_{(\Omega_1,\lambda,W),(\Omega_2,\mu_2,W)} \mathbb{P}_{(\Omega,\mu,W)}\Big[|\hat{d} - dim\,\Omega| \geq \frac{1}{2}\log^{-1}(n)\Big]$$
$$\geq p \vee \frac{1-p}{3}$$

which is larger than $1/4$. Theorem 1.4.5 is proved. □

Proof of Theorem 1.4.6. There are two cases.

For the class $\mathcal{W}_n^{min(1)}(D,\alpha,m,M,v)$, the condition $H_1^{\alpha,v}$ is not imposed. Consider the two following graphons.

- (Ω_1,λ,W) where $\Omega_1 = [0,1]\times\{0\}^{D-1}$ is endowed with the uniform measure λ on $[0,1]$, with $\lambda(\Omega_1)=1$, and where $W:[0,1]^D\times[0,1]^D\to[0,1]$ is a symmetric function that satisfies a double Hölder condition (1.6) with Hölder exponent $\alpha = 1$.
- (Ω_2,μ_2,W) where $\Omega_2=[0,1]^D$ and $\mu_2 = (1-n^{-1})\lambda + n^{-1}\nu$, with ν the uniform measure on $[0,1]^D$.

Following the proof of Theorem 1.4.5, we can show that these two graphons belong to $\mathcal{W}_n^{min(1)}(D,\alpha,m,M,v)$, and that

$$\sup_{\mathcal{W}_n^{min(1)}(D,\alpha,m,M,v)} \mathbb{P}_{(\Omega,\mu,W)}\Big[|\hat{d} - dim\,\Omega| \geq \frac{D}{2}\Big] \geq \frac{1}{4} \tag{1.40}$$

which gives the error bound of Theorem 1.4.6.

For the class $\mathcal{W}_n^{min(2)}(D,\alpha,m,M,v)$, the assumption $H_2^{m,M,v}$ is not assumed. As in the proof of Theorem 1.4.5, we can see that the two following graphons belong to $\mathcal{W}_n^{min(2)}(D,\alpha,m,M,v)$.

- (Ω_1,λ,W) as defined in the above case.
- (Ω_2,μ_2,W) where $\Omega_2=[0,1/(Ln)]^D$ for some large enough (numerical) constant L, and $\mu_2 = (1-n^{-1})\lambda + n^{-1}\nu$, with ν the uniform measure on $[0,1/(Ln)]^D$.

Following the proof of Theorem 1.4.5, with the above two graphons, one obtain the error bound (1.40) over the class $\mathcal{W}_n^{min(2)}(D,\alpha,m,M,v)$.

Thus, (1.40) is proved for $\mathcal{W}_n^{min(j)}(D,\alpha,m,M,v)$, with $j\in\{1,2\}$, and Theorem 1.4.6 follows.

□

1.F Proof for the case of sparse observations

1.F.1 Proof of Corollary 1.6.1 : estimation of the distances

Corollary 1.6.1 is a reformulation of Theorem 1.4.1 in the sparse setting and their proofs are almost identical. In this appendix, denote by W_n the function $\rho_n W$. Accordingly, r_{W_n} denotes the neighborhood distance (1.4) where W has been replaced with W_n. Hence, $r_{W_n} = \rho_n r_W$.

Corollary 1.6.1 is a direct consequence of the two following Lemmas.

Lemma 1.F.1 *For $\rho_n \geq 2\sqrt{\frac{\log n}{n-2}}$ and $n \geq 5$, the following event*

$$\mathcal{E}_{in}^{sp} := \left\{ \forall i,j \in [n] : \ \left| \langle A_i, A_j \rangle_n - \langle W_n(\omega_i,.), W_n(\omega_j,.) \rangle \right| \leq 5\,\rho_n \sqrt{\frac{\log n}{n}} \right\}$$

holds with probability at least $1 - \frac{2}{n}$ with respect to the distribution $\mathbb{P}_{(\Omega,\mu,W),\rho_n}$.

Following the proof of Proposition 1.D.1, we show Lemma 1.F.1 below, by replacing Hoeffding inequality with Bernstein inequality, in order to benefit from the small variance of A_{ij} (which is now of the order of ρ_n).

Lemma 1.F.2 *Conditionally to the event $\mathcal{E}_{in}^{sp}$, the following inequalities*

$$\forall i \in [n] : \ \left| \langle A_i, A_{\widehat{m}(i)} \rangle_n - \langle W_n(\omega_i,.), W_n(\omega_i,.) \rangle \right| \leq 3\rho_n\, r_{W_n}(\omega_i, \omega_{m(i)}) \ + \ 25\,\rho_n \sqrt{\log(n)/n}$$

hold simultaneously.

The proof of lemma 1.F.2 is almost the same as for Proposition 1.D.2. It is omitted.

Proof of Lemma 1.F.1 Conditionally to ω_i, ω_j, $i \neq j$, the $n-2$ random variables $\{A_{ik}A_{kj} : k \in [n], k \neq i,j\}$ are independent with expectation $\mathbb{E}[A_{ik}A_{kj}] = \int_\Omega W_n(\omega_i,z)W_n(\omega_j,z)\mu(dz)$ for all $k \neq i,j$ (where $\mathbb{E}$ is taken w.r.t. the distribution $\mathbb{P}_{(\Omega,\mu,W),\rho_n}$). Using Bernstein inequality [see Sridharan, 2002, for instance], one has

$$\mathbb{P}_{(\Omega,\mu,W),\rho_n}\left(\frac{1}{n-2} \Big| \sum_{k \neq i,j} A_{ik}A_{kj} - (n-2) \int_\Omega W_n(\omega_i,z)W_n(\omega_j,z)\mu(dz) \Big| \geq \epsilon \ \Big| \omega_i, \omega_j \right)$$

smaller than

$$\leq 2\exp\left(\frac{-(n-2)\epsilon^2}{2\rho_n^2 + 2\epsilon/3} \right)$$

for $\epsilon > 0$. Since the above inequality is satisfied for almost every $\omega_i, \omega_j \in \Omega$, we have the same upper bound for the non-conditional probability. Then, setting $\epsilon = 3\rho_n\sqrt{\frac{\log n}{n-2}}$ gives

$$2\exp\left(\frac{-(n-2)\epsilon^2}{2\rho_n^2+2\epsilon/3}\right) \leq 2\exp\left(\frac{-9\log n}{2+\frac{2}{\rho_n}\sqrt{\frac{\log n}{n-2}}}\right) \leq \frac{2}{n^3}$$

since $\rho_n \geq 2\sqrt{\frac{\log n}{n-2}}$ by assumption. Thus, by using the union bound over all $i \neq j$, one obtain

$$\mathbb{P}_{(\Omega,\mu,W),\rho_n}\left(\bigcup_{i,j:i\neq j}\left\{\frac{1}{n-2}\left|\sum_{k\neq i,j} A_{ik}A_{kj} - (n-2)\int_\Omega W_n(\omega_i,z)W_n(\omega_j,z)\mu(dz)\right| \geq \epsilon\right\}\right) \leq \frac{2}{n}.$$

And finally, following the proof of Proposition 1.D.1 leads to

$$\max_{i,j:i\neq j}\left|\sum_k \frac{A_{ik}A_{kj}}{n} - \int_\Omega W_n(\omega_i,z)W_n(\omega_j,z)\mu(dz)\right| \leq 3\rho_n\sqrt{\frac{\log n}{n-2}} + \frac{4}{n}$$

with probability at least $1-\frac{2}{n}$. To conclude the proof, observe that above display is upper bounded by

$$\leq 5\rho_n\sqrt{\frac{\log n}{n}}$$

as soon as $n \geq 5$. □

1.F.2 Proof of of Corollary 1.6.2 : estimation of the dimension

In the proof of Theorem 1.E.1, one has seen

$$\frac{\log m}{-\log\epsilon} - d\frac{(s_\omega+b_{sup})/\epsilon}{-\log\epsilon} \leq \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon)}{-\log\epsilon} - d \leq \frac{\log M}{-\log\epsilon} + d\,\frac{(s_\omega+b_{sup})/\epsilon + ((s_\omega+b_{sup})/\epsilon)^2}{-\log\epsilon}.$$

The sampling error s_ω is not affected by the sparsification of the data through ρ_n, and thus takes the same value as in Theorem 1.E.1. On the other hand, the distance error bound b_{sup} changes, and is now defined as

$$b_{sup}^2 := 6\max_{1\leq i\leq n} r_W(\omega_i,\omega_{m(i)}) + \frac{60}{\rho_n}\sqrt{\log(n)/n}$$

according to Corollary 1.6.1. Following the proof of Theorem 1.E.1, one has

$$b_{sup}+s_\omega \leq 6\left(\frac{2\log n}{\alpha\, n}\right)^{1/2d} + \frac{8}{\sqrt{\rho_n}}\left(\frac{\log n}{n}\right)^{1/4}.$$

Define

$$err_{n,d,\rho_n} := \left(\frac{2\log n}{\alpha\, n}\right)^{1/2d} + \frac{1}{\sqrt{\rho_n}}\left(\frac{\log n}{n}\right)^{1/4} \tag{1.41}$$

so that

$$b_{sup} + s_\omega \leq 8\, err_{n,d,\rho_n}.$$

Following the proof of Theorem 1.E.1, one obtain the same error bound for the dimension estimation, after replacing $6\, err_{n,d}$ with $8\, err_{n,d,\rho_n}$. Indeed, one has

$$\left| \frac{\log \widehat{N}_\Omega^{(c)}(\epsilon)}{-\log \epsilon} - d \right| \leq \frac{1}{-\log \epsilon} \left[\log \left(M \vee \frac{1}{m} \right) + 8d \frac{err_{n,d}}{\epsilon} \left(1 + \frac{err_{n,d}}{\epsilon} \right) \right] \tag{1.42}$$

for all $\epsilon \in]0, v/9]$ and all n such that

$$2\log n/n \leq \alpha\, (v/18)^{2d} \wedge \rho_n^2\, (v/18)^4\,.$$

As in the proof of Theorem 1.4.4, one minimizes the error bound (1.42) by choosing a particular radius of the order of $\sup_{\{d:\, d \leq D\}} err_{n,d,\rho_n} = err_{n,D,\rho_n}$. This gives a radius that satisfies the following relation

$$\epsilon_{D,\rho_n} \asymp \left(\frac{\log n}{n} \right)^{1/(2D)} \vee \frac{1}{\sqrt{\rho_n}} \left(\frac{\log n}{n} \right)^{1/4}.$$

Corollary 1.6.2 follows from the plug-in of ϵ_{D,ρ_n} in (1.42). □

1.G Proofs for the type I and II errors

The current appendix is organized as follows. We first analyse the performance of the new distance estimator (1.19) and then deduce a control on the type I and II errors of the test.

1.G.1 Performance of the new distance estimator

Lemma 1.G.1 shows that the new distance-estimator $\widehat{r}_{new}$ does not over-estimate r_W in the sense of (1.44), without underestimating too much (1.45). Let U be the function defined by

$$U(i) = \text{argmax}_{t \in \{i, \widehat{m}(i)\}} \langle W(\omega_t, .), W(\omega_t, .) \rangle \tag{1.43}$$

for all $i \in [n]$. This means that $U(i)$ indicates which of the two functions $W(\omega_i, .)$ or $W(\omega_{\widehat{m}(i)}, .)$ has the largest l_2-norm $||.||_{2,\mu}$ (see Section 1.4.1 for the definitions of the inner product and the norm).

Lemma 1.G.1 *Consider $t_n = 12\sqrt{\frac{\log n}{n}}$ a fluctuation term and the function U introduced in (1.43). One has the following bounds on the new distance estimator (1.19)*

$$\widehat{r}^2_{new}(i,j) \leq r^2_W(\omega_{U(i)}, \omega_{U(j)}) + t_n \tag{1.44}$$

$$\widehat{r}^2_{new}(i,j) \geq r^2_W(\omega_i, \omega_j) - 5\, r_W(\omega_i, \omega_{m(i)}) - 5\, r_W(\omega_j, \omega_{m(j)}) - 5t_n \tag{1.45}$$

holding simultaneously for all $i, j \in [n]$ *with probability at least* $1 - \frac{2}{n}$ *with respect to the distribution* $\mathbb{P}_{(\Omega,\mu,W)}$.

Recall the useful Proposition 1.D.1 on the convergence of the inner products: the event $\mathcal{E}_{in}$ where the following inequalities hold simultaneously for all $i \neq j$

$$|\langle A_i, A_j\rangle_n - \langle W(\omega_i, .), W(\omega_j, .)\rangle\,| \leq 3\sqrt{\log n/n} \tag{1.46}$$

happens with probability at least $1 - 2/n$.

Proof of (1.44). Assume the above event $\mathcal{E}_{in}$ holds. For all $i, j \in [n]$ such that $\{i, \widehat{m}(i)\} \cap \{j, \widehat{m}(j)\} = \emptyset$, the line (1.46) gives

$$\widehat{r}^2_{new}(i,j) \leq \langle W(\omega_i, .), W(\omega_{\widehat{m}(i)}, .)\rangle + \langle W(\omega_j, .), W(\omega_{\widehat{m}(j)}, .)\rangle \\ - 2 \max_{v\in\{i,\widehat{m}(i)\}, w\in\{j,\widehat{m}(j)\}} \langle W(\omega_v, .), W(\omega_w, .)\rangle + t_n$$

with $t_n = 12\sqrt{\frac{\log n}{n}}$. Then, using the function U defined by (1.43), one has

$$\widehat{r}^2_{new}(i,j) \leq \langle W(\omega_{U(i)}, .), W(\omega_{U(i)}, .)\rangle + \langle W(\omega_{U(j)}, .), W(\omega_{U(j)}, .)\rangle \\ - 2\,\langle W(\omega_{U(i)}, .), W(\omega_{U(j)}, .)\rangle + t_n$$

which is upper bounded by

$$r^2_W(\omega_{U(i)}, \omega_{U(j)}) + t_n$$

with Cauchy-Schwarz inequality. The line (1.44) is proved in the case $\{i, \widehat{m}(i)\} \cap \{j, \widehat{m}(j)\} = \emptyset$.

If $\{i, \widehat{m}(i)\} \cap \{j, \widehat{m}(j)\} \neq \emptyset$, we can see that $\widehat{r}^2_{new}(i,j) \leq 0$. Thus (1.44) trivially holds in this case too. The inequalities (1.44) are proved. □

Proof of (1.45). Assume the event $\mathcal{E}_{in}$ of Proposition 1.D.1 holds.

If $i, j \in [n]$ such that $\{i, \widehat{m}(i)\} \cap \{j, \widehat{m}(j)\} = \emptyset$,

$$\left|r^2_W(\omega_i, \omega_j) - \widehat{r}^2_{new}(i,j)\right| \leq \left|r^2_W(\omega_i, \omega_j) - \widehat{r}^2(i,j)\right| + \left|\widehat{r}^2(i,j) - \widehat{r}^2_{new}(i,j)\right|$$

by triangle inequality. The left term is upper bounded by

$$3\, r_W(\omega_j, \omega_{m(j)}) + 3\, r_W(\omega_i, \omega_{m(i)}) + 36\sqrt{\log(n)/n}$$

thanks to Theorem 1.4.1. The right term is equal to

$$2\left|\langle A_i, A_j\rangle - \max_{k\in\{i,\widehat{m}(i)\}, l\in\{j,\widehat{m}(j)\}} \langle A_k, A_l\rangle\right|$$

which is upper bounded by

$$r_W(\omega_j, \omega_{m(j)}) + r_W(\omega_i, \omega_{m(i)}) + 12\sqrt{\log(n)/n}$$

using the same technique as in the proof of Theorem 1.4.1. Combining the above displays, one has

$$\left|r_W^2(\omega_i,\omega_j) - \widehat{r}_{new}^2(i,j)\right| \leq 5\, r_W(\omega_j,\omega_{m(j)}) + 5\, r_W(\omega_i,\omega_{m(i)}) + 60\,\sqrt{\log(n)/n},$$

which implies

$$\widehat{r}_{new}^2(i,j) \geq r_W^2(\omega_i,\omega_j) - 5\, r_W(\omega_j,\omega_{m(j)}) - 5\, r_W(\omega_i,\omega_{m(i)}) - 60\,\sqrt{\log(n)/n}. \tag{1.47}$$

The line (1.45) is therefore proved in the case $\{i,\widehat{m}(i)\}\cap\{j,\widehat{m}(j)\} = \emptyset$.

If $i,j\in[n]$ such that $\{i,\widehat{m}(i)\}\cap\{j,\widehat{m}(j)\}\neq\emptyset$,

$$\widehat{r}_{new}(i,j) = 0.$$

Hence, it is enough to show that the right hand side of (1.47) is non-positive. Consider the particular case where $\widehat{m}(i) = j$ and $i\neq\widehat{m}(j)$ for example. Then, one has

$$|\widehat{r}^2(i,j)| = \left|\langle A_{\widehat{m}(j)},A_j\rangle - \langle A_i,A_j\rangle\right| \leq |\langle A_i,A_j - A_{\widehat{m}(j)}\rangle| + |\langle A_i - A_j, A_{\widehat{m}(j)}\rangle|$$

which is upper bounded by

$$\widehat{f}(j,m(j)) + \widehat{f}(i,m(i))$$

where $\widehat{f}$ has been introduced in (1.9). As in the proof of Theorem 1.4.1, one can show that the above display is upper bounded by

$$r_W(\omega_j,\omega_{m(j)}) + r_W(\omega_i,\omega_{m(i)}) + 12\,\sqrt{\log(n)/n}$$

on the event $\mathcal{E}_{in}$. Combining this upper bound of $\widehat{r}$ with the following lower bound from Theorem 1.4.1

$$\widehat{r}^2(i,j) \geq r_W^2(\omega_i,\omega_j) - 3\, r_W(\omega_i,\omega_{m(i)}) - 3\, r_W(\omega_j,\omega_{m(j)}) - 36\,\sqrt{\log(n)/n}, \tag{1.48}$$

one derive

$$r_W^2(\omega_i,\omega_j) \leq 4\, r_W(\omega_i,\omega_{m(i)}) + 4\, r_W(\omega_j,\omega_{m(j)}) + 48\,\sqrt{\log(n)/n}.$$

This implies that the right hand side of (1.47) is non positive. Hence (1.45) is proved in the particular case $\widehat{m}(i) = j$ and $i\neq\widehat{m}(j)$. By symmetry, it remains only the case $\widehat{m}(i) = \widehat{m}(j)$ to do. Following the above proof, we can show taht (1.45) holds for this case too. The inequality (1.45) is therefore proved in the case $\{i,\widehat{m}(i)\}\cap\{j,\widehat{m}(j)\}\neq\emptyset$.

The line (1.45) is proved. □

1.G.2 Control on the type I and II errors

In Theorem 1.5.2 on the new packing number estimator, the left hand side of (1.20) is similar to Section 1.4.3 on the covering number estimator, and thus straightforward. The right hand side of (1.20) and Corollary 1.5.3 are proved together below.

Proof for the type I error. Assume the null-hypothesis $N_\Omega^{(p)}(\epsilon)\leq K$ holds. We want to show that the same inequality is satisfied by the statistic $\widehat{N}_\Omega^{(p.new)}(\widehat{\epsilon})$. Proof by contradiction:

assume the inequality $\widehat{N}_{\Omega}^{(p.new)}(\widehat{\epsilon}) \geq K+1$ holds. This means that there are $K+1$ indices $i_1, \ldots, i_{K+1} \in [n]$ such that the following inequalities hold

$$\forall s, t \in \{1, \ldots, K+1\} : \quad \widehat{\epsilon}^2 < \widehat{r}^2_{new}(i_s, i_t).$$

Combining the above inequalities with the under-estimation property (1.44), one has

$$\forall s, t \in \{1, \ldots, K+1\} : \quad \widehat{\epsilon}^2 < r^2_W(\omega_{U(i_s)}, \omega_{U(i_t)}) + t_n$$

with probability at least $1 - 2/n$. Replacing the radius $\widehat{\epsilon}^2$ by its value $\epsilon^2 + t_n$, it comes

$$\forall s, t \in \{1, \ldots, K+1\} : \quad \epsilon^2 < r^2_W(\omega_{U(i_s)}, \omega_{U(i_t)}).$$

Thus, $K+1$ sampled points are separated by at least a distance ϵ, which implies $N_{\Omega}^{(p)}(\epsilon) \geq K+1$. This contradicts the null-hypothesis. □

Corollary 1.5.3 and Theorem 1.5.2 are therefore proved.

Proof for the type II error (Theorem 1.5.4). Consider a graphon (Ω, μ, W) in the set $\mathcal{W}(\eta, \beta)$. By definition of $\mathcal{W}(\eta, \beta)$, there are $K+1$ balls in (Ω, r_W) whose centers are separated by at least a distance $\sqrt{\epsilon^2 + 10\eta + 6t_n} + \eta$. Label these balls by $s \in \{1, \ldots, K+1\}$. As in the proof for the dimension estimation, assume the complementary of the event $\mathcal{E}_{bad}$, i.e. assume that each of the $K+1$ balls contains at least two sampled points. Accordingly, denote by $i_1, j_1, \ldots, i_{K+1}, j_{K+1}$ the indices of the corresponding sampled points such that $\omega_{i_s}, \omega_{j_s}$ belong to the s^{th} ball with $s \in \{1, \ldots, K+1\}$. Since the radius of these ball is smaller than $\eta/2$, one has

$$r_W(\omega_{i_s}, \omega_{m(i_s)}) \leq r_W(\omega_{i_s}, \omega_{j_s}) \leq \eta \tag{1.49}$$

for all $s \in \{1, \ldots, K+1\}$.

On the event $\mathcal{E}_{in}$ of Proposition 1.D.1, Lemma 1.G.1 gives

$$\widehat{r}^2_{new}(i_s, i_t) \geq r^2_W(\omega_{i_s}, \omega_{i_t}) - 5\, r_W(\omega_{i_s}, \omega_{m(i_s)}) - 5\, r_W(\omega_{j_s}, \omega_{m(j_s)}) - 5t_n.$$

for all $s \neq t \in \{1, \ldots, K+1\}$. Using (1.49), one derive

$$\widehat{r}^2_{new}(i_s, i_t) \geq r^2_W(\omega_{i_s}, \omega_{i_t}) - 10\eta - 5t_n. \tag{1.50}$$

The ball centers are separated by at least a distance $\sqrt{\epsilon^2 + 10\eta + 6t_n} + \eta$ by assumption, which implies that the points in these balls are separated by

$$r_W(\omega_{i_s}, \omega_{i_t}) > \sqrt{\epsilon^2 + 10\eta + 6t_n}$$

for all $s \neq t \in \{1, \ldots, K+1\}$, since the ball radii are all smaller than $\eta/2$. Combining this inequality with the line (1.50), one obtain

$$\widehat{r}^2_{new}(i_s, i_t) > \epsilon^2 + t_n$$

for all $s \neq t \in \{1, \ldots, K+1\}$. Since $\widehat{\epsilon} = \sqrt{\epsilon^2 + t_n}$, this gives $\widehat{N}_{\Omega}^{(p.new)}(\widehat{\epsilon}) \geq K+1$. Thus, the alternative hypothesis is confirmed correctly.

The above displays hold on the event $\mathcal{E}_{in} \cap \mathcal{E}_{bad}^c$. Let us upper bound the probability of the complementary event. The union bound gives

$$\mathbb{P}(\mathcal{E}_{in}^c \cup \mathcal{E}_{bad}) \leq \frac{2}{n} + (K+1)2n\beta \, \exp[-\beta(n-1)]$$

thanks to Proposition 1.D.1 and Lemma 1.E.4. Theorem 1.5.4 is then proved. □

Proof for the improvement of the type II error (Theorem 1.A.1). We have seen that the type II error is upper bounded by the probability of the event $\mathcal{E}_{in}^c \cap \mathcal{E}_{bad}$. Here the only difference is that $\mathcal{E}_{bad}$ refers to the new event where, for each of the M collections of $K+1+K'$ balls, at least $K'+1$ balls contain strictly less than two sampled points. For clarity, label these collections by $\{1, \dots, M\}$, and denote by $\mathcal{C}_j$ the event where at least $K'+1$ balls of the j^{th} collection contain strictly less than two sampled points. Then, we have

$$\mathbb{P}[\mathcal{E}_{bad}] = \mathbb{P}[\mathcal{C}_1 \cap \ldots \cap \mathcal{C}_M]$$

where $\mathbb{P}$ denote the probability distribution $\mathbb{P}_{(\Omega,\mu,W)}$ of the W-random graph. The above display is equal to

$$\mathbb{P}[\,\mathcal{C}_1] \times \mathbb{P}[\mathcal{C}_2 \big| \mathcal{C}_1] \times \ldots \times \mathbb{P}[\mathcal{C}_M \big| \mathcal{C}_1, \ldots, \mathcal{C}_{M-1}]$$

which is upper bounded by

$$\mathbb{P}[\,\mathcal{C}_1] \times \mathbb{P}[\mathcal{C}_2] \times \ldots \times \mathbb{P}[\mathcal{C}_M]$$

since the events $\mathcal{C}_1, \dots \mathcal{C}_M$ are negatively associated (it is shown at the end of the proof). Finally, we have

$$\mathbb{P}[\mathcal{E}_{bad}] \leq \mathbb{P}[\mathcal{C}_1]^M. \tag{1.51}$$

Given the first collection of $K+1+K'$ balls, denote by $\mathcal{E}_j$ the event where the j^{th} ball of the collection contains strictly less than two sampled points. By definition of the event $\mathcal{C}_1$, we have

$$\mathbb{P}[\,\mathcal{C}_1] = \mathbb{P}[\,\exists\, i_1, \ldots, i_{K'+1} \in \{1, \ldots, K+1+K'\} : \mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{K'+1}}].$$

The union bound gives

$$\mathbb{P}[\,\mathcal{C}_1] \leq \sum_{i_1, \ldots, i_{K'+1}} \mathbb{P}[\,\mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{K'+1}}]$$

where the sum is taken over all possible $K'+1$ different indices. The above upper bound is equal to

$$\sum_{i_1, \ldots, i_{K'+1}} \mathbb{P}[\,\mathcal{E}_{i_1}] \times \mathbb{P}[\mathcal{E}_{i_2} \big| \mathcal{E}_{i_1}] \times \ldots \times \mathbb{P}[\mathcal{E}_{i_{K'+1}} \big| \mathcal{E}_{i_1}, \ldots, \mathcal{E}_{i_{K'}}].$$

which is smaller than

$$\sum_{i_1, \ldots, i_{K'+1}} \mathbb{P}[\,\mathcal{E}_{i_1}] \times \ldots \times P[\mathcal{E}_{i_{K'+1}}] \tag{1.52}$$

by negative association of the events $\mathcal{E}_k$ (this fact is proved at the end). Finally, Lemma 1.E.4 ensures that

$$\mathbb{P}[\,\mathcal{E}_k] \leq 2\beta n \, \exp[-\beta(n-1)]$$

for all k, which allows to upper bound (1.52) and have

$$\mathbb{P}[\mathcal{C}_1] \leq \binom{K+K'+1}{K'+1} \Big(2\beta n \, \exp[-\beta(n-1)]\Big)^{(K'+1)}. \tag{1.53}$$

Thus, setting $\widetilde{p}_n = \mathbb{P}[\mathcal{C}_1]$, we deduce from (1.51) that

$$\mathbb{P}(\mathcal{E}_{in}^c \cup \mathcal{E}_{bad}) \leq \mathbb{P}(\mathcal{E}_{in}^c) + \mathbb{P}(\mathcal{E}_{bad}) \leq \frac{2}{n} + \widetilde{p}_n^M,$$

where $\widetilde{p}_n$ is upper bounded by (1.53).

It remains to show the negative association that we use in the above proof. Given the first collection of $K+1+K'$ balls, let us show that the corresponding events $\mathcal{E}_1, \dots, \mathcal{E}_{K+1+K'}$ are negatively associated. For the n sampled points $\omega_1, \dots, \omega_n$, define n_j the number of points in the j^{th} ball of the collection. Theorem 13 of Dubhashi and Ranjan [1998] ensures that the variables $n_1, \dots, n_{K+1+K'}$ are negatively associated. Define the non-increasing function $h(n_j) = \mathbb{I}_{\mathcal{E}_j}$ where $\mathbb{I}_{\mathcal{E}_j}$ is the indicator function of $\mathcal{E}_j$. The second point of Proposition 7 of Dubhashi and Ranjan [1998] shows that $h(n_1), \dots, h(n_{K+1+K'})$ are negatively associated. This means that the events $\mathcal{E}_1, \dots, \mathcal{E}_{K+1+K'}$ are negatively associated.

Similarly, we show the negative association of the events $\mathcal{C}_1, \dots, \mathcal{C}_M$. Consider n_j^t the number of sampled points in the j^{th} ball of the t^{th} collection. These variables are negatively associated according to Theorem 13 of Dubhashi and Ranjan [1998]. Define the non-increasing functions $h_t(n_1^t, \dots, n_{K+1+K'}^t) = \mathbb{I}_{\mathcal{C}_j}$ for all $t \leq M$. Then, Proposition 7 of Dubhashi and Ranjan [1998] shows that $\mathbb{I}_{\mathcal{C}_1}, \dots, \mathbb{I}_{\mathcal{C}_M}$ are negatively associated.

Theorem 1.A.1 is proved. □

Chapter 2

Pair-Matching: Links Prediction with Adaptive Queries

The pair-matching problem appears in many applications where one wants to discover good matches between pairs of entities or individuals. Formally, the set of individuals is represented by the nodes of a graph where the edges, unobserved at first, represent the good matches. The algorithm queries pairs of nodes and observes the presence/absence of edges. Its goal is to discover as many edges as possible with a fixed budget of queries. Pair-matching is a particular instance of multi-armed bandit problem in which the arms are pairs of individuals and the rewards are edges linking these pairs. This bandit problem is non-standard though, as each arm can only be played once.

Given this last constraint, sublinear regret can be expected only if the graph presents some underlying structure. This paper shows that sublinear regret is achievable in the case where the graph is generated according to a Stochastic Block Model (SBM) with two communities. Optimal regret bounds are computed for this pair-matching problem. They exhibit a phase transition related to the Kesten-Stigum threshold for community detection in SBM. The pair-matching problem is considered in the case where each node is constrained to be sampled less than a given amount of times. We show how optimal regret rates depend on this constraint. The paper is concluded by a conjecture regarding the optimal regret when the number of communities is larger than 2. Contrary to the two communities case, we argue that a statistical-computational gap would appear in this problem.

Contents

2.1 Introduction

Many real world data can be represented as a graph of pairwise relationships. Examples include social networks connections, metabolic networks, protein-protein interaction networks, citations network, recommendations and so on. Matchmaking algorithms and link prediction algorithms are routinely used in many practical situations to discover biochemical interactions, new contacts, hidden connections between criminals, or to match players in online multiplayers video games and sport tournaments. As testing a link in biological networks, or discovering connections between criminals can be expansive, link prediction algorithms are useful to focus on the most relevant links. In social networks or online video games, they can help in finding relevant partners.

These applications raise the following mathematical problem that this paper intends to study. Suppose that there exists a graph whose nodes represent a set of entities or individuals and whose edges represent successful matches between entities or individuals. The nodes are known to the statistician while the edges are typically hidden at first. Matchmaking algorithms make queries on pairs of individuals, trying to discover as many edges as possible. For biological networks like protein-protein interaction networks, the individuals are proteins, an edge is an interaction between the two proteins and a query is an experiment to test whether the interaction exists. The goal of matchmaking algorithms is to discover as many edges of the graph as possible while minimizing the number of mismatches. To stress that the focus lies on discovering graph structures, the problem at hand is called hereafter pair-matching rather than matchmaking.

The pair-matching algorithm is forced to explore the graph as it cannot make queries on edges that have already been observed. To learn interesting features on unobserved edges from previous observations, it is necessary to make assumptions on the structure of the hidden graph. This paper considers the arguably simplest situation where the graph has been generated according to an assortative conditional stochastic block model (SBM) [Holland et al., 1983] with two balanced communities, see Section 2.2.1 for a formal presentation. In this model, individuals are grouped into two (unobserved) communities and the probability of successful match (edge) between two individuals is larger if they belong to the same

community than to different ones. In this context, the set of pairs is partitioned into good and bad ones, good pairs contain two individuals from the same community and bad pairs two individuals from different communities. A pair-matching algorithm samples pairs and should sample as many good pairs as possible. Of course, the partition into good and bad pairs is unknown.

When the graph is fully observed, communities are recovered using clustering algorithms, which have been extensively studied over the past few years, see for example [Abbe, 2017, Moore, 2017, Can et al., 2018] for recent overviews. A key parameter in the analysis of clustering algorithms, called here *scaling parameter* s, is the ratio

$$s = \frac{(p-q)^2}{p+q},$$

where p is the probability of connection within a community and q the probability of connection between communities. This parameter measures the difficulty of clustering, see Section 2.2.1 for details. The quality of a pair-matching algorithm is evaluated by the expected number of discovered edges after T queries. Equivalently, the performance can be measured by the expected number of pairs sampled that do not contain edges, which should be as small as possible, see Section 2.2.3 for details. This last quantity is proportional to the expected number of bad pairs sampled, which is called *sampling regret* in this paper. As in practical situations, individuals may not be solicited too many times, we consider algorithms constrained to sample each individual less than a certain amount of times B_T before T queries have been made.

Our main contribution of the paper is that the sampling regret of any strategy that cannot sample pairs more than once, that is invariant to nodes labelling and which satisfies the above constraint (see Assumptions **(NR)**, **(IL)** and **(SpS)** in Section 2.2.2 for details) is larger than

$$T \wedge \frac{\sqrt{T} \vee (T/B_T)}{s},$$

up to multiplicative constants. Moreover, a polynomial-time algorithm with sampling regret bounded from above by a constant times $T \wedge \frac{\sqrt{T} \vee (T/B_T)}{s}$ is described and analysed, see Theorem 2.4.1. These results show that no strategy can achieve sub-linear sampling regret before $T = O(1/s^2)$ pairs have been sampled and that, on the other hand, there exist strategies with sub-linear regret scaling as the optimal rate $(\sqrt{T} \vee (T/B_T))/s$ once $T \gtrsim 1/s^2$. It transpires from this result that the constraint has no substantial effect as long as $B_T \gtrsim \sqrt{T}$. On the other hand, strong constraints such as $B_T = O(1/s)$ induce unavoidable linear regret.

The following problem, related to matchmaking, has recently attracted attention, in particular in Bradley-Terry models [Bradley and Terry, 1952, Zermelo, 1929]. The task is to infer, from the observation of pairs, a vector of parameters characterizing the strength of players. Most results considered the case where all the graph is observed, see [Hunter, 2004, Caron and Doucet, 2012]. Recent contributions dealing in particular with ranking issues also consider the case of partially observed graphs, see [Shah and Wainwright, 2017, Shah et al., 2016, Jang et al., 2016] for example and the references therein. In all cases, the list of observed pairs is given as input to the algorithm evaluating the strength of all players. The choice of a relevant list of successive observed pairs, independent of the observation of the edges is sometimes called a scheduling problem, see [Le Corff et al., 2018]. Scheduling problems are

different from matchmaking problems considered here where the algorithm should choose the observed pairs and can use preliminary observations to make its choice. For online video games, classical algorithms used to evaluate strength of players are ELO or TRUESKILLS [Herbrich et al., 2007, Minka et al., 2018]. Matchmaking algorithms such as EOMM [Chen et al., 2017] (used with TRUESKILLS see [Minka et al., 2018]) are then used to pair players, taking as inputs these estimated strengths. In this approach, the number of mismatches during the learning phase is not controlled. It is an important conceptual difference with this paper where the matchmaking problem is considered together with the problem of discovering the strength (communities here). Here, pair-matching algorithms have to simultaneously explore the graph to evaluate the strength and sample as many "good" pairs as possible to optimize the number of successful matches. Closer to our setting is the active ranking literature [Jamieson and Nowak, 2011, Szörényi et al., 2015, Heckel et al., 2019], where the goal is to discover adaptively the rank or strength of players with a minimal amount of queries. Contrary to our problem, only the exploration matters in adaptive ranking and no notion of regret is investigated.

Pair-matching algorithms take sequential decisions to explore new pairs exploiting previous observations. This kind of exploration and exploitation dilemma is typical in multi-armed bandit problems [Thompson, 1933, Robbins, 1952, Lai and Robbins, 1985, Burnetas and Katehakis, 1996]. In stochastic multi-armed bandit problems, a set of actions, called *arms* is proposed to a player who chooses one of these actions at each time step and receives a payoff. The payoffs are independent random variables with unknown distribution. For any arm, payoffs are identically distributed. The player wants to maximize its total payoff after T queries. The pair-matching problem introduced above can be seen as a non-standard instance of stochastic multi-armed bandit problems. In this interpretation, each pair of nodes is an arm and the associated payoff is 1 if an edge links these nodes and 0 otherwise. The payoffs hence follow a Bernoulli distribution with parameter p for good pairs and parameter q for bad pairs. The unusual feature is that each arm can only be played once, so the pair-matcher must choose a new arm at each time step. For this reason, optimal strategies differ in spirit from classical strategies in bandit problems, see Section 2.2.4 for more details. On the other hand, useful inequalities are borrowed from the classical bandit literature [Kaufmann et al., 2016, Garivier et al., 2018] to prove lower bounds.

Forgetting the constraint that a node cannot be sampled more than B_T times, the pair-matching bandit problem could be seen as an extreme version of mortal or rotting bandit problems [Chakrabarti et al., 2009, Levine et al., 2017, Seznec et al., 2019], where every arm would systematically die or have zero pay-off after the first sampling. Without additional assumptions, the regret would be inexorably linear in the querying budget T. Here, an important difference with classical mortal or rotting multi-armed bandits is that payoffs are structured by the underlying stochastic block model (SBM). Stochastic block models have attracted a lot of attention in the recent years, with a focus on the determination of optimal strategies for clustering and for parameter estimation, see [Abbe, 2017, Moore, 2017]. In this prolific literature, the graph is fully observed and the question is to identify precisely the weakest separation between the probabilities of connection necessary to perfectly or partially recover the communities, or to estimate the parameters of the SBM. Closer to our setting, the paper [Yun and Proutière, 2014b] investigates the question of recovering communities from a minimal number of observed pairs, sampled sequentially. In this problem, the question is to assign a community to all nodes after a minimal number T of time steps and try to

minimize the number of misclassified nodes. This is quite different from the minimization of the sampling regret considered here, where we seek to find on a budget as many good pairs as possible and not to classify all nodes. As discussed in Section 2.3.3, applying the algorithm of [Yun and Proutière, 2014b] would lead to a suboptimal regret in our problem.

The formalization of the pair-matching problem considered in this paper may be restrictive in some applications. Section 2.5 presents some conjectures that seem reasonable for K classes SBMs. Other graph structures would also be interesting, such as Bradley-Terry models [Bradley and Terry, 1952, Zermelo, 1929] which have been used for sport tournaments [Sire and Redner, 2009], chess ranking [Joe, 1990] and predictions of animal behaviors [Whiting et al., 2006]. Various constraints dealing with first discoveries for example may be interesting depending on the applications: the first match of a node is the most important in some situations[1], and, for the search of a life partner, discovering a match with a node already connected in the observed graph is (for most nodes at least) less interesting than a match with an isolated node. These constraints naturally induce different versions of the pair-matching problem and raise mathematical questions of interest. Multiplayer video games suggest the extension to hypergraphs of the pair-matching problem. Indeed, the value of a player could be evaluated as part of a team and with respect to a possible team of opponents rather than simply as part of a pair. Finally, in many practical situations, additional information on individuals is available and could be used to improve pair-matching algorithms. It is clear from our first results that this information is necessary to avoid linear regret in applications such as life partner research. These extensions are postponed to follow-up works. This paper should be seen as a first step to formalize and study the important pair-matching problem. It focuses on a toy example but opens several interesting questions that arise when dealing with natural constraints in practical applications of interest.

The remainder of the paper is decomposed as follows. Section 2.2 introduces the formal setting and objectives. As a warm-up, Section 2.3 focuses on the case where the algorithms are not constrained to sample nodes more than a certain amount of times. Section 2.4 presents the main results where the algorithm are constrained. Section 2.5 gives conjectures for K-classes SBMs. The proofs of the main results are postponed to the appendix.

Notation: we write $x_n \lesssim y_n$ and $x_n = O(y_n)$, if there exist numerical constants such that $x_n \leqslant C y_n$ for all $n \geq n_0$; and we write $x_n \asymp y_n$ and $x_n = \Theta(y_n)$, if $x_n = O(y_n)$ and $y_n = O(x_n)$ that is, if there exist numerical constants $c, c' > 0$ and n_0 such that $c x_n \leq y_n \leq c' x_n$ for all $n \geq n_0$. We denote by $\lceil x \rceil$ (respectively $\lfloor x \rfloor$) the upper (resp. lower) integer part of x; by $|A|$ the cardinal of a set A; and by $A \Delta B$ the symmetric difference between two sets A and B.

2.2 Setting and Problem Formalization

2.2.1 Two-Classes SBM

The n individuals are represented by the set $V = \{1, \dots, n\}$. Successful matches are represented by a set of edges E between nodes in V: there is a successful match between a and b in V if and only if the pair $\{a, b\}$ belongs to E. Hereafter, a set of two distinct elements in V is called

[1] Richard III in Shakespeare's play offers his "kingdom for a horse!", he would certainly propose less for a second one!

a *pair* and an element of E is called an *edge*. The graph (V, E) is conveniently represented by its adjacency matrix $A \in \mathbb{R}^{n\times n}$, with entries $A_{ab} = 1$ if $\{a, b\} \in E$ and $A_{ab} = 0$ otherwise. In the following, any graph (V, E) is identified with its adjacency matrix $(A_{ab})_{a,b\in V}$. For any pair $e = \{a, b\}$, the notations A_e and A_{ab} are used indifferently. Since the graph is undirected, the adjacency matrix A is symmetric, and since there is no self-matching (no self-loop in the graph), the diagonal of A is equal to zero.

Individuals are grouped into two (unknown) communities according to their affinity. To model this situation, the graph (V, E) is random and distributed as a two-classes conditional stochastic block model. Let $0 < q, p < 1$, and let n_1 denote an integer $n_1 \geq n - n_1 \geq 1$. The collection $\mathrm{cSBM}(n_1, n-n_1, p, q)$ of two-classes conditional stochastic block model distributions on graphs is defined as follows. Let $G = \{G_1, G_2\}$ be a partition of $\{1, \dots, n\}$ into two groups, with $|G_1| = n_1$ and $|G_2| = n - n_1$. The partition G represents the communities of individuals. Let μ_G denotes the distribution on graphs with nodes $\{1, \dots, n\}$, such that the adjacency matrix is symmetric, null on the diagonal and with lower diagonal entries $(A_{ab})_{a<b}$ sampled as independent Bernoulli random variables with $\mu_G(A_{ab} = 1) = p$ when a and b belong to the same group G_i, and $\mu_G(A_{ab} = 1) = q$ when a and b belong to different groups. In other words, two individuals are successfully matched with probability p if they belong to the same community, and with probability q otherwise. The class $\mathrm{cSBM}(n_1, n - n_1, p, q)$ is defined as the set of all distributions μ_G defined above, where $G = \{G_1, G_2\}$ describes the set of partitions of $\{1, \dots, n\}$ satisfying $|G_1| = n_1$ and $|G_2| = n - n_1$:

$$\mathrm{cSBM}(n_1, n - n_1, p, q) = \{\mu_G : G = \{G_1, G_2\} \text{ partition satisfying } |G_1| = n_1,\ |G_2| = n - n_1\}.$$

In the following, the communities are balanced and successful matches happen with higher probability if individuals belong to the same community. Formally, n is even and the graph (V, E) has been generated according to a distribution μ in $\mathrm{cSBM}(n/2, n/2, p, q)$, for some unknown parameters p and q such that $0 < q < p \leq 1/2$. As $q < p$, the distribution of (V, E) is called an *assortative* $\mathrm{cSBM}(n/2, n/2, p, q)$. All along the paper, the ratio p/q is also assumed bounded from above. To sum up, p and q are smaller than $1/2$ and satisfy

$$1 < p/q \leq \rho^*. \tag{2.1}$$

Given p and q, the following scaling parameter plays a central role

$$s = \frac{(p-q)^2}{p+q}. \tag{2.2}$$

This parameter appears in various results in the literature on SBM. The following property, proved for example in [Yun and Proutière, 2014a, Chin et al., 2015, Abbe and Sandon, 2015, Lu and Zhou, 2016, Gao et al., 2017, Fei and Chen, 2019, Giraud and Verzelen, 2019], will be used repeatedly in the paper. When the graph $(V, E) \sim \mathrm{cSBM}(n_1, n - n_1, p, q)$, there exist polynomial-time clustering algorithms that return a partition of $\{1, \dots, n\}$ such that, with large probability, the proportion of misclassified nodes decreases exponentially:

$$\text{Proportion of misclassified nodes} \leq \exp(-cns), \quad \text{when } ns \geq c',$$

where $c, c' > 0$ are numerical constants. The rate ns of exponential decay in this result is optimal (up to a constant) when (2.1) is met. Hence, the scaling parameter s drives the

difficulty of clustering. To stress the importance of s, the following parametrization will be used henceforth

$$p = s\left(\alpha + \sqrt{\alpha}\right)/2, \quad q = s\left(\alpha - \sqrt{\alpha}\right)/2,$$

with $\alpha = (p+q)^2/(p-q)^2$. In this parametrization, Assumption (2.1) is met if and only if α is bounded from below by $(\rho^*+1)^2/(\rho^*-1)^2$. Another useful property is that there exist numerical constants $c_1, c_2 > 0$ such that non-trivial community recovery is possible as soon as $s \geq c_1/n$, see [Decelle et al., 2011, Massoulié, 2014, Chin et al., 2015, Abbe and Sandon, 2015, Bordenave et al., 2018, Fei and Chen, 2019, Giraud and Verzelen, 2019] and perfect community recovery is possible as soon as $s \geq c_2 \log(n)/n$, see [Abbe and Sandon, 2015, Chen and Xu, 2016, Mossel et al., 2016].

The reader familiar with SBM literature may be more comfortable with the parametrization $p = a_n/n$ and $q = b_n/n$ for a SBM distribution with two communities. For a comfortable translation of the results, the following relations between s, α and a_n, b_n are provided:

$$s = \frac{(a_n - b_n)^2}{n(a_n + b_n)}, \qquad \alpha = \frac{(a_n + b_n)^2}{(a_n - b_n)^2},$$
$$\frac{a_n}{b_n} = \frac{\alpha + \sqrt{\alpha}}{\alpha - \sqrt{\alpha}} \quad \text{and} \quad a_n + b_n = n\alpha s.$$

2.2.2 Sequential Matching strategies

Denote by $\mathcal{E}$ the set of all pairs of nodes, that is the set of all subsets of V containing two distinct elements. Heuristically, a sequential matching strategy samples at each time t a new pair $\widehat{e}_t \in \mathcal{E}$, using only past observations $(\widehat{e}_1, \dots, \widehat{e}_{t-1}, A_{\widehat{e}_1}, \dots, A_{\widehat{e}_{t-1}})$ and an internal randomness of the algorithm.

Formally, let $U_0, U_1, \dots$ be i.i.d uniform random variables in $[0,1]$, independent of A and representing the sequence of internal randomness for the algorithm. A sequential matching strategy ψ on $\mathcal{E}$ (shortened *strategy* in the following) is a sequence $\psi = (\psi_t)_{0 \leq t \leq \binom{n}{2}-1}$ of measurable functions $\psi_t : \mathcal{E}^t \times \{0,1\}^t \times [0,1]^{t+1} \to \mathcal{E}$. Any sequential matching strategy ψ defines a matching algorithm as follows. The first pair is sampled as $\widehat{e}_1 = \psi_0(U_0)$. Then, at each time $t \geq 0$, the pair $\widehat{e}_{t+1}$ is defined by

$$\widehat{e}_{t+1} = \psi_t(\widehat{\mathcal{E}}_t, (A_e)_{e \in \widehat{\mathcal{E}}_t}, U_0, \dots, U_t) \quad \text{with} \quad \widehat{\mathcal{E}}_t = \{\widehat{e}_1, \dots, \widehat{e}_t\}.$$

The strategy takes as input the observed graph $(A_e)_{e \in \widehat{\mathcal{E}}_t}$ and possibly an internal independent randomness U_t to output the new observed pair $\widehat{e}_{t+1}$.

In the following, strategies are assumed to satisfy the following constraints: a pair can only be sampled once and strategies are invariant to labelling of the nodes. These constraints can be formalized as follows.

Non-redundancy (NR). *The strategy ψ samples any pair at most once, that is, for any $0 \leq t \leq \binom{n}{2} - 1$ and $e_1, \dots, e_t \in \mathcal{E}$, the map ψ_t fulfills $\psi_t(\{e_1, \dots, e_t\}, \dots) \notin \{e_1, \dots, e_t\}$.*

Invariance to labelling requires some notation. For any pair $e \in \mathcal{E}$ and any strategy ψ, let

$$N_e(\psi, t) := \mathbf{1}_{e \in \widehat{\mathcal{E}}_t} \tag{2.3}$$

indicate if the pair e has been sampled or not before time t by the strategy ψ. For any non-redundant strategy ψ (i.e. satisfying **(NR)**), pairs are sampled at most once and the observation of $\{N_e(\psi,t) : e \in \mathcal{E}\}$ is equivalent to that of $\widehat{\mathcal{E}}_t$.

Let μ be a distribution in cSBM$(n/2, n/2, p, q)$ and σ be a permutation of V. For any pair $\{a,b\} \in \mathcal{E}$, let $\sigma(\{a,b\}) := \{\sigma(a), \sigma(b)\}$. Let μ^σ denote the distribution of $(A_{\sigma(e)})_{e\in\mathcal{E}}$, where $(A_e)_{e\in\mathcal{E}}$ is distributed according to μ.

Invariance to labelling (IL). *The distribution of the outcomes of the strategy ψ is invariant by permutations of the nodes labels: For any $\mu \in cSBM(n/2, n/2, p, q)$ and any permutation σ on V, the distribution of $\left(N_e(\psi,t) : e \in \mathcal{E}, 1 \le t \le \binom{n}{2}\right)$ under μ^σ is the same as the distribution of $\left(N_{\sigma(e)}(\psi,t) : e \in \mathcal{E}, 1 \le t \le \binom{n}{2}\right)$ under μ.*

Besides **(NR)** and **(IL)**, we consider strategies that do not sample a node more than B times before time T. This constraint appears naturally in practical situations. For example, if the algorithm matches biological entities or individuals, one may not want to query too many times each individual for logistic or acceptability reasons. To stress that the constraint B typically grows with the time horizon T, it is denoted B_T. Formally, for any $a \in V$, let

$$N_a(\psi,t) = \sum_{b\in V : b\neq a} N_{\{a,b\}}(\psi,t) \tag{2.4}$$

denote the number of times the node a has been sampled in a pair $\{a,b\}$ after t queries.

Sparse sampling (SpS). *Let T and B_T denote two integers. The strategy ψ is called B_T-sparse up to time T if it satisfies*

$$\forall a \in V, \quad N_a(\psi,T) \le B_T. \tag{2.5}$$

Since $N_a(\psi,T) \le (n-1) \wedge T$ for all nodes a, choosing $B_T \ge (n-1) \wedge T$ corresponds to the unconstrained case.

2.2.3 Objectives of the Pair-matcher

Let $\mu \in$ cSBM$(n/2, n/2, p, q)$ be the distribution of an assortative conditional stochastic block model with associated partition $G = \{G_1, G_2\}$. Define $\mathcal{E}^{good}(\mu)$ (or simply $\mathcal{E}^{good}$) as the set of pairs $\{a,b\}$ with a and b from the same community, and $\mathcal{E}^{bad}(\mu)$ (or simply $\mathcal{E}^{bad}$) as the set of pairs $\{a,b\}$ with a and b from two different communities.

The objective of the pair-matcher is to discover as many edges (i.e. successful matches between individuals) as possible with T queries. Its strategy ψ should maximize the number of discovered edges, in expectation with respect to the randomness of the SBM and the strategy. Optimal strategies should therefore sample as many pairs in $\mathcal{E}^{\text{good}}$ as possible. Formally, consider a time horizon T smaller than $|\mathcal{E}^{\text{good}}| = 2\binom{n/2}{2} \sim n^2/4$. Any strategy ψ has an expected number of discoveries equal to

$$\mathbb{E}_\mu\left[\sum_{t=1}^{T} A_{\widehat{e}_t}\right] = pT - (p-q)\,\mathbb{E}_\mu\left[N^{bad}(\psi,T)\right],$$

where $N^{bad}(\psi, T) = \sum_{e \in \mathcal{E}^{bad}} N_e(\psi, T)$ is the number of pairs in $\mathcal{E}^{bad}$ sampled up to time T. Since $p > q$, the maximal expected value of discoveries is achieved by any oracle strategy ψ^* sampling only edges in $\mathcal{E}^{\text{good}}$. In that case, $N^{bad}(\psi^*, T) = 0$ and the maximal expected number of discoveries is equal to pT. The regret of the strategy ψ is defined as the difference between pT and its expected number of discoveries:

$$R_T(\psi) = pT - \mathbb{E}_\mu \left[\sum_{t=1}^{T} A_{\widehat{e}_t} \right] = (p-q)\, \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right].$$

As long as $T \leqslant |\mathcal{E}_{\text{good}}|$, the regret is proportional to the expected number of sampled between-group pairs $\mathbb{E}_\mu \left[N^{bad}(\psi, T) \right]$. Therefore, the main results analyse this last quantity rather than the regret. The expected number of bad sampled pairs $\mathbb{E}_\mu \left[N^{bad}(\psi, T) \right]$ is called hereafter *sampling-regret*.

Remark. Without assumption on ψ, the distribution of $N^{bad}(\psi, T)$ may depend on the distribution μ of the cSBM. On the other hand, when the strategy ψ fulfils **(IL)**, the distribution of $N^{bad}(\psi, T)$ does not depend on the distribution μ in cSBM$(n/2, n/2, p, q)$. Indeed, let μ, μ' be two distributions in cSBM$(n/2, n/2, p, q)$. By definition, there exists a permutation σ on $\{1, \dots, n\}$ such that $\mu' = \mu^\sigma$, where μ^σ has been defined page 92. Since $\mathcal{E}^{bad}(\mu^\sigma) = \sigma^{-1}(\mathcal{E}^{bad}(\mu))$, it follows from **(IL)** that the distribution under μ^σ of $\sum_{e \in \mathcal{E}^{bad}(\mu^\sigma)} N_e(\psi, T)$ is the same as the distribution under μ of $\sum_{e \in \mathcal{E}^{bad}(\mu)} N_e(\psi, T)$.

2.2.4 A Special Bandit Problem

The pair-matching problem described above can be interpreted as a non-standard multi-armed bandit problem. Actually, each pair $\{a, b\}$ can be seen as an arm and the discovery of a successful match as a payoff. The payoff of the arm $\{a, b\} \in \mathcal{E}$ follows a Bernoulli distribution with parameter p if $\{a, b\} \in \mathcal{E}^{good}$ and with parameter q if $\{a, b\} \in \mathcal{E}^{bad}$. This bandit problem is non-standard, as arms cannot be sampled more than once and payoffs have a structure inherited from the SBM distribution.

To sum up, the main differences with the standard multi-armed bandit problem are:

1. the arms are sampled at most once,
2. at most B_T arms involving a given node can be sampled up to time T,
3. the distribution of the payoffs have a hidden structure inherited from the SBM setup.

Compared with the standard multi-armed bandit problem, points 1 and 2 make this problem harder, while point 3 is a strong structural property that gives hope to find regimes with sub-linear regret.

These special features make this problem quite different from classical bandit problems. In classical bandit problems, optimal strategies have to identify the best arm (or some of the best arms) and each arm is played many times to reach this goal. Here, half the arms are "optimal" but one cannot play an arm more than once. Therefore, instead of identifying one of these, optimal strategies should avoid bad arms, possibly disregarding a non-negligible proportion of good arms in the process.

The constraint **(SpS)** also induces a specific exploration / exploitation trade-off. When the community of a node is identified, we wish to pair it with a maximum of nodes of the same community in order to maximise the rewards (exploitation). Yet, we also need to pair this node to some new nodes in order identify the community of new nodes (exploration). Since a node can only be paired to B_T other nodes, we need to trade-off between these two strategies.

The bandit literature is mainly used to establish our lower bounds which involve inequalities from [Garivier et al., 2018, Kaufmann et al., 2016].

2.3 Warm-up: Unconstrained Optimal Pair-Matching

2.3.1 Optimal Rates for Unconstrained Pair-Matching

As a warm-up, we focus first on the simplest case, where $B_T = +\infty$, which amounts to remove the constraint **(SpS)**. Let Ψ_∞ denote the set of strategies ψ fulfilling **(NR)** and **(IL)**. The first main result describes the best sampling-regret that can be achieved by a strategy in Ψ_∞, as a function of s and T.

Theorem 2.3.1 *Let T and n be positive integers with $T \leq |\mathcal{E}^{good}| = 2\binom{n/2}{2}$. Let $p, q \in [0, 1/2]$ be two parameters fulfilling (2.1) and such that*

$$s \leq \frac{1}{32(1+\rho^*)},$$

where the scaling parameter s is defined in (2.2). Then, for any $\mu \in cSBM(n/2, n/2, p, q)$,

$$\inf_{\psi \in \Psi_\infty} \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \geq \frac{1}{32} \left[\frac{\sqrt{T}}{32(1+\rho^*)s} \wedge T \right]. \tag{2.6}$$

Moreover, there exist two numerical constants $c_1, c_2 > 0$, and a strategy $\psi \in \Psi_\infty$ corresponding to a polynomial-time algorithm described in Section 2.3.2, taking s as input, such that, for any p, q satisfying (2.1), any $\mu \in cSBM(n/2, n/2, p, q)$ and any time horizon $1 \leq T \leq c_2 n^2$

$$\mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \leq c_1 \left[\frac{\sqrt{T}}{s} \wedge T \right].$$

The proof of Theorem 2.3.1 is provided in the appendix. The lower bound is proved in Section 2.A and the upper bound in Section 2.B. The upper bound derives from a stronger result showing that similar bounds hold with high probability, see Theorem 2.B.1 for a precise statement. Theorem 2.3.1 provides only the upper bound in expectation for clarity.

Theorem 2.3.1 states that, when (2.1) holds, for any $\mu \in \text{cSBM}(n/2, n/2, p, q)$ and any time horizon $1 \leq T \leq c_2 n^2$, the optimal sampling-regret

$$\inf_{\psi \in \Psi_\infty} \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \asymp \frac{\sqrt{T}}{s} \wedge T,$$

grows linearly with T as long as $T \lesssim 1/s^2$ and becomes sub-linear, of order $\sqrt{T}/s$, when $T \gtrsim 1/s^2$.

This result can be understood intuitively. As long as communities cannot be recovered better than random, there is no hope of getting better sampling-regret than with purely random sampling of the pairs. In this regime, the sampling-regret grows linearly with T. To identify when this occurs, consider the situation where pairs are sampled at random among N nodes and $T = \beta N^2/2$ (with $\beta \leqslant 1$). Then the *observed* edges at time T are approximately distributed as in a SBM with N nodes, within-group connection probability $p_\beta = \beta p$, and between-group connection probability $q_\beta = \beta q$. It follows from [Decelle et al., 2011, Massoulié, 2014, Mossel et al., 2015, Bordenave et al., 2018] that weak recovery of the communities is possible if and only if $N(p_\beta - q_\beta)^2 \geq 2(p_\beta + q_\beta)$, which is equivalent to $\sqrt{\beta T}s \geq \sqrt{2}$ or $T \geq 2/(\beta s^2)$. Since $\beta \leq 1$ by definition, no information about the communities can be recovered when $T \leq 2/s^2$. Hence, the sampling-regret is expected to grow linearly with T for $T = O(1/s^2)$. This intuition is confirmed by Eq. (2.6).

When $T \gg 1/s^2$, the situation is different. Classical results, such as [Yun and Proutière, 2014a, Chin et al., 2015, Abbe and Sandon, 2015, Lu and Zhou, 2016, Mossel et al., 2016, Fei and Chen, 2019, Giraud and Verzelen, 2019] among others, ensure that the communities of N nodes can be recovered almost perfectly if $N \gg 1/s$ and all edges between these nodes are observed. Therefore, when $1/s \ll N = \left(\sqrt{T}/s\right)^{1/2} \ll \sqrt{T}$, one can sample all the edges between N nodes and recover almost perfectly their community with a sampling regret smaller than $N^2 = \sqrt{T}/s$.

A recipe in order to get a sublinear regret is the following. If we are able to find the community of $\Theta(\sqrt{T})$ nodes, then we can spend a budget of T queries without further regret by sampling pairs among these $\Theta(\sqrt{T})$ nodes. To do so, we need to identify the community of $\Theta(\sqrt{T})$ nodes from the N clustered nodes, with a regret smaller than $\sqrt{T}/s$. Given the N clustered nodes, it is possible to identify the community of a new node with a sampling regret of order $O(1/s)$. Proceeding recursively, $\Theta(\sqrt{T})$ new nodes can be identified with a sampling-regret of order $O(\sqrt{T}/s)$. The remaining budget of T queries can then be spent by sampling pairs among these $\Theta(\sqrt{T})$ nodes without further regret if there were no errors in the community assignment. This informal reasoning suggests that the optimal sampling-regret grows like $\sqrt{T}/s$ when $T \gg 1/s^2$. Again, this intuition is confirmed by Eq. (2.6). An algorithm achieving the optimal upper bound in Theorem 2.3.1 and taking as input s and the time horizon T is provided in Section 2.3.2. It essentially proceeds as in the informal strategy outlined above, even if some steps have to be refined. In particular, the identification of the community of $\Theta(\sqrt{T})$ nodes has to be conducted with care in order to balance the regret and the community assignment errors. The dependency of the algorithm of Section 2.3.2 on the time horizon T, can be easily dropped out with a classical doubling trick, see Section 2.B in the appendix.

To sum up the discussion: in the early stage where $T = O(1/s^2)$, one cannot do better than random guessing, up to multiplicative constant factors. In the second stage where $T \geq 1/s^2$, the rate $\sqrt{T}/s$ can be interpreted as follows. A total of $O(\sqrt{T})$ nodes are involved at time T and, for each of them, $O(1/s)$ observations are necessary to obtain an educated guess of their community.

Finally, Theorem 2.3.1 can be equivalently stated in terms of the regret $R_T(\psi)$: for any

time horizon $1 \leq T \leq c_2 n^2$, the minimal regret satisfies

$$\inf_{\psi \in \Psi_\infty} R_T(\psi) \asymp \sqrt{\alpha}\left(\sqrt{T} \wedge (sT)\right),$$

when the assumptions of Theorem 2.3.1 are met.

2.3.2 Algorithm with Specified Horizon T

This section presents an algorithm achieving the upper bound in Theorem 2.3.1. This algorithm takes as input the scaling parameter s *and* the time horizon T. This dependency on the time-horizon can be avoided with the classical doubling trick, see Section 2.B in the appendix. We discuss in Section 2.5.1 a heuristic for the preliminary estimation of s involving less than $O(1/s^2)$ edges.

When the horizon T is $O(1/s^2)$, any strategy achieves a regret of order $O(T)$. Hence, without loss of generality, it is assumed in the remaining of the section that $T \geq c_{th}/s^2$ for some numerical constant c_{th}. Moreover, as Theorem 2.3.1 holds for $T \leq c_2 n^2$, it is also assumed that this condition is fulfilled for a sufficiently small constant c_2.

The algorithm proceeds in three steps. In the first step, a kernel $\mathcal{N}$ of $|\mathcal{N}| = \Theta(\sqrt{T}/\log(s\sqrt{T}))$ vertices is chosen uniformly at random and each pair within this kernel is sampled with probability $\Theta\big((\log(s\sqrt{T}))^2/(s\sqrt{T})\big)$. Hence, an average of $\Theta(\sqrt{T}/s)$ pairs are sampled within this kernel. A community recovery algorithm is run on this observed graph that outputs two estimated communities with a fraction of misclassified nodes vanishing as $O(\log(s\sqrt{T})/(s\sqrt{T}))$ with high probability.

The second step identifies with high probability $\Theta(\sqrt{T})$ vertices from the same community, say community 1. To do so, it picks uniformly at random a set $\mathcal{A}_0$ of $8\sqrt{2T}$ vertices outside of the kernel $\mathcal{N}$ (this is possible thanks to the condition $T \leq c_2 n^2$) and samples pairs between this set and the estimated community 1 of the kernel. This set of edges is used to estimate the connectivity between these vertices and community 1. Vertices with low connectivity, that seem to belong to community 2, are removed online to keep the sampling regret under control. The goal of this screening is not to classify perfectly the $8\sqrt{2T}$ picked vertices, but instead to sift out vertices of community 2 with a low sampling regret. In particular, a price to pay to achieve this goal is to possibly remove a non-negligible proportion of vertices of community 1 from the $8\sqrt{2T}$ picked vertices. This second step of the algorithm is crucial for getting the optimal regret rate $O(\sqrt{T}/s)$. A simplified version of this second step can be connected to a particular k out of m best arms identification problem. This connection is discussed in Section 2.3.3 below.

The third step samples all pairs $\{a, b\}$ such that a and b belong to the $\Theta(\sqrt{T})$ vertices isolated in the second step of the algorithm, until the remaining budget of T queries is expended.

The pair-matching algorithm calls an external clustering algorithm (generically denoted by `GOODCLUST` in the following). `GOODCLUST` takes as input a graph (V, E) and outputs a partition $\widehat{G} = (\widehat{G}_1, \widehat{G}_2)$. We require that `GOODCLUST` fulfills the following recovery property: There exist numerical constants $c^{\mathrm{GC}}, c_1^{\mathrm{GC}} > 0$ such that, for all $N = N_1 + N_2$ and all $\tilde{p}, \tilde{q} \in [0, 1]$, if

$(V, E) \sim \mathrm{cSBM}(N_1, N_2, \tilde{p}, \tilde{q})$, the proportion of misclassified nodes

$$\varepsilon_N = \frac{|\widehat{G_1} \Delta G_1| + |\widehat{G_2} \Delta G_2|}{2N},$$

with Δ the symmetric difference, satisfies

$$\varepsilon_N \leq \exp\left(-c_1^{\mathrm{GC}} N \frac{(\widetilde{p} - \widetilde{q})^2}{\widetilde{p}}\right), \tag{2.7}$$

with probability at least $1 - c^{\mathrm{GC}}/N^3$. Algorithms achieving this proportion of misclassification can be found e.g. in [Giraud and Verzelen, 2019], see also [Yun and Proutière, 2014a, Chin et al., 2015, Abbe and Sandon, 2015, Lu and Zhou, 2016, Gao et al., 2017, Fei and Chen, 2019] for similar results.

Unconstrained Algorithm

Inputs: s scaling parameter, T time horizon, V set of nodes.

Internal constants: $c_{\mathcal{O}_0} = 2 \vee (1/c_1^{\mathrm{GC}})$, $C_k = 2200$ and $C_I = 4$.

Step 1: finding communities in a kernel

1. Sample uniformly at random a set $\mathcal{N} \subset V$ of $N = \left\lceil \sqrt{T}/\log(s\sqrt{T}) \right\rceil$ nodes.
2. Sample each pair of $\mathcal{N}$ with probability $c_{\mathcal{O}_0} \dfrac{\sqrt{T}}{s\binom{N}{2}}$, call $\mathcal{O}_0 \subset \mathcal{E}$ the output.
3. Estimate global connectivity $\tau = (p+q)/2$ by $\hat{\tau} = \dfrac{1}{|\mathcal{O}_0|} \displaystyle\sum_{e \in \mathcal{O}_0} A_e$.
4. Run `GOODCLUST` on the graph with nodes set $\mathcal{N}$ and edges present in $\mathcal{O}_0$. Output, for any $x \in \mathcal{N}$, $\widehat{Z}_x$ the estimated community of x. Choose the label $\hat{Z} = 1$ for the largest estimated community.

Step 2: expanding the communities

5. Sample uniformly at random a set $\mathcal{A}_0$ of $|\mathcal{A}_0| = \left\lceil 8\sqrt{2T} \right\rceil$ nodes in $V \setminus \mathcal{N}$.
6. Set $k = \lceil C_k/s \rceil$ and $I = \left\lceil C_I \log(s\sqrt{T}) \right\rceil$
7. **For** $i = 1, \dots, I$, **do**
   (a) **For** $x \in \mathcal{A}_{i-1}$, sample k nodes $(y^x_{k(i-1)+a})_{a=1,\dots,k}$ uniformly at random in $\mathcal{N} \cap \left\{\hat{Z} = 1\right\} \setminus \{y^x_a\}_{a=1,\dots,k(i-1)}$.
   (b) Sample the pairs $(\{x, y^x_{k(i-1)+a}\})_{a=1,\dots,k}$ and let $\hat{p}_{x,i} = \frac{1}{ki} \sum_{a=1}^{ki} A_{x y^x_a}$.
   (c) Select $\mathcal{A}_i = \{x \in \mathcal{A}_{i-1} : \hat{p}_{x,i} \geq \hat{\tau}\}$.
   (d) **In case**[a] where $\mathcal{A}_i = \emptyset$, **then** set $\mathcal{A}_I = \emptyset$ and **BREAK**.

Step 3: sampling pairs within estimated communities

8. Sample uniformly at random pairs within the set $\mathcal{A}_I$ until T pairs have been sampled overall. If the number of sampled pairs is smaller than T after all pairs in $\mathcal{A}_I$ have been sampled[a], then sample the remaining pairs at random.

Output: T pairs sampled at steps 2., 7.(b) and 8. of the algorithm.

[a] with high probability, this undesirable case does not happen

2.3.3 Community Expansion versus k out of m Best Arm Identification

As proved in Lemma 2.B.2 in the Appendix 2.B, after Step 1, with high probability, we end up with a set of N classified nodes, where at most $O(1/s)$ of them are misclassified, and the empirical connectivity $\hat{\tau}$ does not deviate from the population one $\tau = (p+q)/2$ by more than $(p-q)/4$. The goal of Step 2 is then to identify $\sqrt{2T}$ new nodes of community 1, with at

most $O(1/s)$ misclassified nodes and a regret at most $O(\sqrt{T}/s)$. Let us connect this problem to a k out of m best arms identification problem.

Let us consider a simplified version of the problem of Step 2. Assume that we have identified $N_1 = N/2$ nodes of community 1 with no error, that we have access to the population connectivity τ and that among the $M = 8\sqrt{2T}$ nodes in $\mathcal{A}_0$, half of them are of community 1. Then, each node $a \in \mathcal{A}_0$ can be seen as an arm, and pulling the arm a amounts to query a pair $\{a, b\}$ with b one of the N_1 nodes of community 1 identified at Step 1. The mean reward of the arm a is p if it belongs to community 1, and q otherwise. Hence, a simplified version of the problem in Step 2 amounts to identify $k = \sqrt{2T}$ out of $m = M/2 = 4\sqrt{2T}$ best arms, with at most $O(1/s)$ errors, and a cumulated regret $O(k/s)$. We have the additional constraint that an arm can be pulled at most N_1 times, but we will forget this additional feature in this discussion, for simplicity of the comparison.

The problem of identifying k out of m best arms with a tolerance ϵ has been investigated in [Goschin et al., 2013, Ren et al., 2019]. The focus on these papers is on the minimal sample size needed to identify k arms whose expected reward is larger than the m^{th} largest expected reward minus ϵ. The main results of [Ren et al., 2019] states that, with probability at least $1 - k^{-2}$, the algorithm AL-Q-FK can recover with a sample size

$$O\left(\frac{1}{(p-q)^2}\left(M \log\left(\frac{m+1}{m+1-k}\right) + k\log(k)\right)\right)$$

k out of the m best arms with a tolerance $\epsilon = (p-q)/2$. The sampling regret is not considered and it can be as large as the sample size. In the same setting, the screening algorithm of Step 2 achieves the following performance. For $m \geq ck \geq c'/s$, with probability at least $1 - c''k^{-2}$ a budget of at most $O(ks^{-1}\log(sk))$ queries, and a sampling regret at most $O(k/s)$, the algorithm identifies a set of arms with at least k out of m best arms and at most $O(1/s)$ arms not in the m best ones. As $s = (p-q)^2/(p+q)$, the sampling regret achieved by the screening algorithm of Step 2 is at least $(p+q)/\log(k)$ times smaller. We can explain this gain by several reasons. The $p+q$ improvement comes from the fact that we explicitly take into account the fact that the rewards have a Bernoulli distribution. The $1/\log(k)$ improvement is obtained by a careful design of the algorithm to keep the regret low, at the price of possibly $O(1/s)$ identification errors.

Specified to the simplified version of the problem in Step 2 depicted above, the AL-Q-FK algorithm would return $k = \sqrt{2T}$ nodes out of the $m = M/2 = 4\sqrt{2T}$ nodes of community 1 with a sampling regret

$$O\left(\frac{\sqrt{T}}{(p+q)s}\ \log(\sqrt{T})\right).$$

This sampling regret is larger than the $O(\sqrt{T}/s)$ regret needed for our Step 2, so the AL-Q-FK cannot be used as a black-box for Step 2.

We emphasize also that the expansion of the communities in Step 2 is somewhat more complex than the simplified version described above: at Step 1, up to $O(1/s)$ nodes are misclassified, we only have access to the empirical connectivity $\hat{\tau}$, an arm can only be pulled N_1 times and the number of best arms is random.

We also emphasize that we cannot use the algorithm of [Yun and Proutière, 2014b] as a black-box to identify $\sqrt{2T}$ nodes of community 1 within $\mathcal{A}_0$ with at most $1/s$ errors and with

a sampling regret $O(\sqrt{T}/s)$. Indeed, if we take $\Theta(\sqrt{T})$ nodes and apply the procedure of [Yun and Proutière, 2014b] to classify them with a sampling regret at most $O(\sqrt{T}/s)$, then a *fixed* proportion of the nodes are misclassified and pairing them together at Step 3 would generate a final regret of order $\Theta(T)$. In addition, from the lower bounds in [Yun and Proutière, 2014b], we observe that the above phenomenon occurs, whatever the algorithm, if we try to classify *all* the nodes in $\mathcal{A}_0$. To overcome this issue, the algorithm of Step 2 recovers the class for a *fraction* only of the nodes in $\mathcal{A}_0$ with a sampling regret at most $O(\sqrt{T}/s)$ and at most $1/s$ errors. When recovering the class of $\sqrt{2T}$ nodes within $\mathcal{A}_0$, we do not sample pairs at random, but we carefully select them in order to avoid as much as possible the sampling of bad pairs.

2.4 Constrained Optimal Pair-Matching

2.4.1 Main Results

Let us now consider the general problem, where sparse sampling **(SpS)** is enforced. The algorithm described in Section 2.3.2 for unconstrained pairs-matching uses extensively the opportunity to make "localized" queries: At time T, a small number of $\Theta(\sqrt{T})$ nodes has been queried a large number of $\Theta(\sqrt{T})$ times, while other nodes have been queried less than $O(\log(s\sqrt{T})^2/s)$ times. So, the strategy has to be adapted to fulfills **(SpS)**.

For a sparsity bound B_T, denote by $\Psi_{B_T,T}$ the set of strategies ψ fulfilling the Non-redundancy **(NR)**, Invariance to labelling **(IL)** and Sparse sampling **(SpS)** properties at time T.

Theorem 2.4.1 *Let T and n be positive integers with $T \leq |\mathcal{E}^{good}| = 2\binom{n/2}{2}$. Let $p, q \in [0, 1/2]$ be two parameters fulfilling (2.1) and such that the parameter s, defined in (2.2), fulfills*

$$s \leq \frac{1}{32(1+\rho^*)}.$$

Then, for any $\mu \in cSBM(n/2, n/2, p, q)$,

$$\inf_{\psi \in \Psi_{B_T,T}} \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \geq \frac{1}{32} \left[\frac{\sqrt{T} \vee (T/B_T)}{32(1+\rho^*)s} \wedge T \right].$$

Conversely, there exist two numerical constants $c_1, c_2 > 0$ such that, for any time horizon T and constraint B_T satisfying $1 \leq T \leq c_1 n(B_T \wedge n)$, there exist a strategy $\psi \in \Psi_{B_T,T}$ corresponding to a polynomial-time algorithm, described in Section 2.4.2, such that

$$\mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \leq c_2 \left[\frac{\sqrt{T} \vee (T/B_T)}{s} \wedge T \right]. \tag{2.8}$$

We refer to the appendix for a proof of this theorem. The lower bound is proved in Section 2.A and the upper bound in Section 2.C.

Compared with Theorem 2.3.1, Theorem 2.4.1 shows that the sparse sampling constraint **(SpS)** amounts to replace $\sqrt{T}$ by $\sqrt{T} \vee (T/B_T)$ in the optimal sampling-regret. In particular,

the sparse sampling constraint downgrades optimal rates only when B_T is smaller than $\sqrt{T}$. Actually, a close look at the unconstrained algorithm page 98 reveals that, by construction, it satisfies assumption **(SpS)** with $B_T = 17\sqrt{T}$. So, in the regime where $B_T \geq 17\sqrt{T}$, the lower bound cannot be worse than the upper-bound of the unconstrained setting of Theorem 2.3.1.

When $B_T \lesssim \sqrt{T}$, the optimal sampling-regret is of order $(T/(B_T s)) \wedge T$. This rate can be understood as follows. If $B_T \leq 1/s$, there is not enough observations per node to infer their community better than at random, which induces an unavoidable linear regret. When $B_T \gg 1/s$, to proceed as in Step 3 of the constrained case, one needs to identify a sufficiently large set of nodes of the same community, among which one can sample up to T pairs without adding regret. As each node can now be paired with at most B_T others, this set should be of size $\Theta(T/B_T)$ instead of $\Theta(\sqrt{T})$ in the unconstrained case. As the identification of the community of a node requires at least $\Theta(1/s)$ queries, the sampling-regret expected to identify this large set of nodes is $\Theta(T/(B_T s))$.

The previous informal discussion suggests to extend the algorithm described in Section 2.3.2 for the unconstrained case. This extension, fully described and commented in Section 2.4.2, still proceeds in 3 steps and goes as follows. The first step of the constrained algorithm is essentially the same as the first step of the unconstrained algorithm, with $\sqrt{T}$ replaced by $B = (B_T \wedge \sqrt{T})/2$. In this first step, all pairs are sampled among a set of $B/\log(sB) \leq B_T$ nodes, so the constraint cannot be violated. Then, to keep the sampling-regret under control while not violating the **(SpS)** contraint, the trick is to apply recursively a variant of the screening algorithm in Step 2 and repeat these screenings until a total number of $\Theta(T/(B_T \wedge \sqrt{T}))$ nodes are correctly classified, with a small proportion of error. Finally, one can sample at most $B_T \wedge \sqrt{T}$ pairs for each of these nodes in Step 3 with a controlled regret. The resulting algorithm extends the unconstrained one of Section 2.3.2 where $B_T \wedge \sqrt{T} = \sqrt{T}$ and where the screening step is only applied once. This extension is fully described in Section 2.4.2.

To illustrate the theorem, one can discuss the results with the constraint $B_T = T^\gamma$, where $0 < \gamma \leq 1/2$. In this case, the optimal sampling-regret is of order $T \wedge (T^{1-\gamma}/s)$. It follows that any pair-matching algorithm that is T^γ-sparse up to time T (besides satisfying **(NR)** and **(IL)**) has linear sampling-regret up to time $s^{-1/\gamma}$. On the other hand, there exist strategies with optimal sampling-regret of order $T^{1-\gamma}/s$ after time $s^{-1/\gamma}$.

Notice that the sparse sampling property $N_a(\psi, T) \leq B_T$ only constrains the algorithm at the time horizon T. This time horizon has therefore to be specified beforehand for this constraint to be defined. In many practical situations, this specification is not reasonable and a more realistic constraint takes the form: $N_a(\psi, t) \leq B_t$ at any time $t \in \{1, \dots, T\}$. In the case where $B_t = \Theta(t^\gamma/(\log t)^\tau)$, the constraint can be enforced using a doubling trick, without enlarging the regret by more than a multiplicative numerical constante. This doubling trick is discussed in detail in Section 2.4.4.

2.4.2 Algorithm with Sparse Sampling

The algorithm described in page 98, that achieves optimal regret in the unconstrained case, identifies first a set of $\Theta(\sqrt{T})$ nodes from one community with $O(1/s)$ misclassified nodes and a regret of order $O(\sqrt{T}/s)$ in Steps 1 and 2. Then, it pairs these nodes together in Step 3 with a $O(\sqrt{T}/s)$ regret (due to the misclassified nodes).

The algorithm described in this section follows essentially the same steps, identifying first a set of nodes from the same community (with small error) and then sampling pairs among them. It has to be adapted to fulfill the **(SpS)** constraint. As the unconstrained algorithm fulfills the **(SpS)** constraint for any $B_T \geq 17\sqrt{T}$, it is assumed in the remaining of this section that $B_T = O(\sqrt{T})$. Moreover, as the result holds for $T \leq c_1 n(B_T \wedge n)$, this assumption is granted in the remaining of the section.

To respect the constraint **(SpS)**, no node may be sampled in more than B_T pairs. Hence, to perform the last step, the algorithm has to identify $\Theta(T/B_T)$ nodes from one community. It should achieve this identification with a sampling-regret smaller than $O(T/(sB_T))$ while respecting the **(SpS)** constraint. To respect the **(SpS)** constraint in the first step of the algorithm, a kernel $\mathcal{N}_{init}$ of cardinality smaller than B_T is chosen. Formally, in points 1. and 2. of Step 1 in the algorithm page 98, $\sqrt{T}$ is replaced by $(B_T \wedge \sqrt{T})/2$. Then, as in the unconstrained case, Step 2 expands the communities in order to identify, with high probability and up to a small error, $\Theta(T/B_T)$ nodes from one community. The main difference with the unconstrained case is that this expansion cannot be achieved in a single step of screening. Actually,

(i) $\Theta(N/s)$ pairs are required to identify the community of $\Theta(N)$ new nodes.

(ii) Any node from the kernel $\mathcal{N}_{init}$ cannot be sampled more than B_T times.

By (ii), one cannot sample more than $O(|\mathcal{N}_{init}|B_T)$ pairs and by (i), it follows that at most $O(|\mathcal{N}_{init}|B_T s) = O(B_T^2 s)$ nodes can be classified with a single screening step based on $\mathcal{N}_{init}$. The main idea of the new algorithm is to iterate the screening step, expanding progressively the communities. Along these iterations, to satisfy the **(SpS)** constraint, the screening has to be conducted with more care than in step 2 of the unconstrained algorithm page 98. The trick is to apply the `SCREENING` function described page 104, which compartmentalizes the nodes in order to enforce the condition **(SpS)**. This iterative process outputs a set of $\Theta(T/B_T)$ nodes from one community (with a small proportion of error with high probability). The algorithm finally pairs nodes among this subset while respecting the **(SpS)** constraint in Step 3 of the algorithm.

Constrained Algorithm

Inputs: s scaling parameter, T time horizon, V_{init} the set of the n nodes of the whole graph, B_T constraint.

Internal constants: set $c_{\mathcal{O}_0} = 8 \vee (1/c_1^{\mathrm{GC}})$ and $B = (B_T \wedge \sqrt{T})/2$.

Step 1: finding communities in a kernel

1. Sample uniformly at random an initial set $\mathcal{N}_{init} \subset V_{init}$ of $N_{init} = \left\lceil \frac{B}{\log(sB)} \right\rceil$ nodes.
2. Sample each pair of $\mathcal{N}_{init}$ with probability $c_{\mathcal{O}_0} \frac{B}{s} / \binom{N_{init}}{2}$, call $\mathcal{O}_0 \subset \mathcal{E}$ the output.
3. Estimate mean connectivity $\tau = \frac{p+q}{2}$ by $\hat{\tau} = \frac{1}{|\mathcal{O}_0|} \sum_{(x,x') \in \mathcal{O}_0} A_{x,x'}$.
4. Run `GOODCLUST` on the graph $(\mathcal{N}_{init}, \mathcal{O}_0)$ and output, for any $x \in \mathcal{N}_{init}$, $\widehat{Z}_x$ the estimated community of x (with the convention that the largest estimated community is labelled by 1).

Step 2: iteratively expanding the communities

Internal constants: set $N^{(0)} = \lceil N_{init}/2 \rceil$,

$$t_f = \left\lceil \frac{\log(\lceil T/B \rceil / N^{(0)})}{\log \lfloor \log(sB) \rfloor} \right\rceil \tag{2.9}$$

and for all $t \in \{0, \dots, t_f\}$,

$$N^{(t)} = N^{(0)} \lfloor \log(sB) \rfloor^t \wedge \left\lceil \frac{T}{B} \right\rceil . \tag{2.10}$$

5. Let $\mathcal{N}^{(0)}$ be a set of $N^{(0)}$ nodes in $\mathcal{N}_{init} \cap \{\widehat{Z} = 1\}$ sampled uniformly at random, and let $V^{(0)} = V_{init} \setminus \mathcal{N}_{init}$.
6. **For** $t = 1, \dots, t_f$, **set**

$$(\mathcal{N}^{(t)}, V^{(t)}) = \texttt{SCREENING}\left(\mathcal{N}^{(t-1)}, N^{(t)}, B, \hat{\tau}, V^{(t-1)}\right). \tag{2.11}$$

Step 3: sampling pairs within estimated communities

7. Sample pairs within the set $\mathcal{N}^{(t_f)}$ while respecting the constraint **(SpS)** with B_T, until T pairs have been sampled overall (the sampling method does not matter).

Function SCREENING$(\mathcal{N}, N', B, \nu, V) = (\mathcal{N}', V')$

Inputs: a reference kernel $\mathcal{N}$ of cardinality N, a target number of nodes N', a constraint $B \in \mathbb{R}_+$, a threshold $\nu \in [0,1]$, a set of "new" nodes V.

Output: a set of nodes $\mathcal{N}' \subset V$ of cardinality at most N' and the set of nodes $V' \subset V$ that are still "new" after running SCREENING. (Most of the nodes of $\mathcal{N}'$ will belong to the most represented community in $\mathcal{N}$.)

Internal constants: a number of pairs per step $k = \lceil \frac{C_k}{s} \rceil$ and a number of steps $I = \lceil C_I \log(sB) \rceil$, with $C_k = 2500$ and $C_I = 1026$.

1. Sample uniformly at random a set $\mathcal{A}_0$ of $|\mathcal{A}_0| = 4N'$ nodes in V.
2. Let $m = \lfloor N/(kI) \rfloor$. Take a uniform partition of $\mathcal{N}$ into m sets $(\mathcal{V}_j)_{1 \le j \le m}$ of cardinality kI and one set of cardinality smaller than kI.

   Likewise, take a uniform partition of $\mathcal{A}_0$ into m sets $(\mathcal{A}_0^{(j)})_{1 \le j \le m}$ with cardinality in $\{\lfloor 4N'/m \rfloor, \lceil 4N'/m \rceil\}$.
3. **For** $j = 1, \dots, m$ **and** $i = 1, \dots, I$, **do**

   For each $x \in \mathcal{A}_{i-1}^{(j)}$, **do**

   i. Sample k nodes $(y^x_{k(i-1)+a})_{a=1,\dots,k}$ uniformly at random in $\mathcal{V}_j \setminus \{y^x_a\}_{a=1,\dots,k(i-1)}$.

   ii. Sample pairs $(\{x, y^x_{k(i-1)+a}\})_{a=1,\dots,k}$ and compute

   $$\hat{p}_{x,i} = \frac{1}{ki} \sum_{a=1}^{ki} A_{x y^x_a}. \tag{2.12}$$

   iii. Select $\mathcal{A}_i^{(j)} = \left\{ x \in \mathcal{A}_{i-1}^{(j)} : \hat{p}_{x,i} \ge \nu \right\}$.
4. Set $\mathcal{N}'$ a set of N' nodes sampled uniformly at random from $\bigcup_{1 \le j \le m} \mathcal{A}_I^{(j)}$.

   In case[a] $|\bigcup_{1 \le j \le m} \mathcal{A}_I^{(j)}| < N'$, **then** sample at random N' nodes in $\mathcal{A}_0$.
5. Set $V' = V \setminus \mathcal{A}_0$.

Return $(\mathcal{N}', V')$.

[a] with high probability, this undesirable case does not happen

2.4.3 Screening versus k out of m Best Arms Identification

Similarly as in Section 2.3.3, let us compare the screening step to a k out of m best arms identification problem. The main additional feature compared to the situation discussed in Section 2.3.3, is that an arm a cannot be sampled more than B times. Hence, a simplified version of the screening problem amounts to identify k out of m best arms with tolerance $\epsilon = (p-q)/2$, with the constraint that each arm cannot be sampled more than B times. In these

simplified setting, the screening function achieves the following performance. Assume that $M \geq ck$ and $k, B \geq c'/s$. With probability $1-c(sk)^{-1}$, with a budget of $O(ks^{-1}\log(s(B\wedge k)))$ queries, and with a sampling regret at most $O(k/s)$, the screening function identifies at least k arms of community 1 with at most $O((k(sB)^{-1}) \vee s^{-1})$ errors.

The situation handled by the screening function is actually somewhat more complex than the stylized bandit problem depicted above. Actually, among the initial set of N classified nodes, we have up to $cN/(sB)$ misclassified nodes. At the same time, we cannot query more than B times any of these classified nodes. Hence, we need a careful querying policy in order to avoid the misclassified nodes to generate errors, while keeping the **(SpS)** condition enforced. Fulfilling together these two conditions is the main hurdle in the design and analysis of the screening function.

2.4.4 Pathwise Sparse Sampling Algorithm

The algorithm presented above fulfills the sparse sampling condition **(SpS)** at time horizon T. In many practical situations, it is more natural to consider Condition **(SpS)** at all times $t = 1, 2, \ldots$ rather than only at a predefined time horizon $t = T$. Formally, Condition **(SpS)** would be replaced by $N_a(\psi, t) \leq B_t$, for all $t = 1, 2, \ldots$. It is possible to modify the previous algorithm to build a strategy ψ such that, when $B_t = \Theta(t^\gamma \log^{-\tau}(t))$, the sampling regret $\mathbb{E}_\mu\left[N^{bad}(\psi, t)\right]$ fulfills

$$\mathbb{E}_\mu\left[N^{bad}(\psi, t)\right] = O\left(\frac{\sqrt{t} \vee (t/B_t)}{s} \wedge t\right), \quad \text{for } t = 1, 2, \ldots.$$

Assume that there exists $\gamma \in (0, 1/2]$ and $\tau \in [0, +\infty)$ such that $B_t = t^\gamma/(\log t)^\tau$, so $\sqrt{t} \vee (t/B_t) = t^{1-\gamma}\log^\tau(t)$. In this case, a pathwise sampling condition can be enforced using the simple doubling trick.For any positive integer l, let $t_l = 2^l$. At each time t_l, the new algorithm discards all nodes and pairs previously sampled and starts the algorithm of Section 2.4.2 with the remaining nodes, time horizon $T = t_{l+1} - t_l$ and terminal sparse sampling constraint $N_a(\psi, t_{l+1} - t_l) \leq \min_{t_l \leq t \leq t_{l+1}} B_t$. The resulting strategy does not depend on any time horizon and it fulfills the condition $N_a(\psi, t) \leq B_t$, for all $t = 1, 2, \ldots$.

Moreover, for any l such that $t_l \geq e^{\tau/\gamma}$, $\min_{t_l \leq t \leq t_{l+1}} B_t = B_{t_l}$. Hence, for any l such that $t_l < c_1 n(B_t \wedge n)$ and for any t such that $t_{l-1} \leq t \leq t_l < c_1 n(B_t \wedge n)$,

$$\begin{aligned}
\mathbb{E}\left[N^{bad}(\psi, t)\right] &= O\left(1 + \sum_{k=1}^{l} \frac{(t_k - t_{k-1})^{1-\gamma}\log^\tau(t_k - t_{k-1})}{s} \wedge (t_k - t_{k-1})\right) \\
&= O\left(\left(\frac{1}{s}\sum_{r=0}^{l-1} 2^{r(1-\gamma)}(r\log(2))^\tau\right) \wedge t_l\right) \\
&= O\left(\frac{t_l^{1-\gamma}\log^\tau(t_l)}{s} \wedge t_l\right) = O\left(\frac{t^{1-\gamma}\log^\tau(t)}{s} \wedge t\right).
\end{aligned}$$

According to Theorem 2.4.1, the sampling-regret of the algorithm derived from the doubling trick is then rate optimal.

2.5 Discussion

The present paper provides the optimal sampling-regret for pair-matching in the case where $G = (E, V)$ is a conditional SBM with a number of groups $K = 2$, where the groups have n/K elements, with intra class probability of connection p and inter-class q. The algorithm depicted p.98 in Section 2.3.2 runs in polynomial time and has optimal sampling-regret given in Theorem 2.3.1, up to a multiplicative constant. Let us discuss the two following questions: How can we estimate the scaling parameter s? How does the rates depend on the number K of groups?

2.5.1 A Heuristic to Estimate the Scaling Parameter s

The algorithms described p.98 and p.103 in Sections 2.3.2 and 2.4.2 take the scaling parameter s as input. This parameter is typically unknown in practice and an estimated value $\widehat{s}$ has to be plugged in the algorithm. To guarantee a sampling-regret smaller than $O(T \wedge (\sqrt{T}/s))$, the estimator $\widehat{s}$ should use at most $O(1/s^2)$ edges and satisfy $\widehat{s} \asymp s$ with high probability. The following heuristic builds a possible estimator $\widehat{s}$.

Pick uniformly at random N nodes in V and sample all $N(N-1)/2$ pairs between these N nodes. When $Ns > 2$, $p = a/N$ and $q = b/N$, [Mossel et al., 2015] ensures that, as $N \to \infty$, a and b can be consistently estimated. Therefore, $Ns = (a-b)^2/(a+b)$ can also be consistently estimated from these $T = N(N-1)/2 = O(1/s^2)$ observations. Yet, this estimator requires Ns larger than 2 and cannot therefore be used directly when s is unknown.

However, when $p = a/N$ and $q = b/N$ and $N \to \infty$, it is theoretically possible to detect whether $Ns = (a-b)^2/(a+b)$ is smaller or larger than 2. To proceed, denote by $\mathcal{B}$ the non-backtracking matrix associated to the graph (see [Bordenave et al., 2018] for a definition of the non-backtracking matrix). Let $\lambda_1, \lambda_2, \dots$ be the eigenvalues of $\mathcal{B}$ ranked in decreasing order of their moduli. The main result of [Bordenave et al., 2018] shows that, when $p = a/N$ and $q = b/N$, with $a, b > 0$ fixed, except on an event of vanishing probability as $N \to \infty$,

$$|\lambda_2|^2 < \lambda_1 \quad \text{when} \quad Ns < 2 \ ,$$
$$|\lambda_2|^2 > \lambda_1 \quad \text{when} \quad Ns > 2 \ .$$

In addition, when $Ns > 2$, the ratio $2|\lambda_2|^2/\lambda_1$ consistently estimates $(a-b)^2/(a+b)$.

This result suggests the following recursive algorithm to estimate s: start with a set V_1 of 2 nodes i and j picked uniformly at random in V. Query the pair $\{i, j\}$ and let E_1 denote the set of edges in $E \cap \{i, j\}$. At each step $k \geqslant 2$, pick at random a set V_k of 2^k nodes in $V \setminus \cup_{\ell \leqslant k-1} V_\ell$. Sample all pairs in V_k, and denote by E_k the set of edges among these pairs. Build the non-backtracking matrix $\mathcal{B}_k$ of the graph (V_k, E_k) and compute $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ the eigenvalues of this matrix with largest moduli. If $|\lambda_2^{(k)}|^2 < \lambda_1^{(k)}$ iterate. If $|\lambda_2^{(k)}|^2 > \lambda_1^{(k)}$ stop, denote by $\widehat{k}$ the stopping iteration time and $\widehat{N} = \binom{2^{\widehat{k}}}{2}$ the number of edges sampled in the last graph $(V_{\widehat{k}}, E_{\widehat{k}})$. Output $\widehat{s} = 2|\lambda_2^{(\widehat{k})}|^2/(\widehat{N}\lambda_1^{(\widehat{k})})$.

Assume that $p = a/N$ and $q = b/N$ with $a, b \in \mathbb{R}^+$ fulfilling $(a-b)^2/(a+b) > 2$. Let Ω_N denote the event where simultaneously $2 \leq \widehat{N}s \leq 4$ and $s/2 \leq \widehat{s} \leq 2s$. Then the results of [Bordenave et al., 2018] suggest that the event Ω_N holds with probability tending to 1 as

$N \to \infty$. In addition, the total number of sampled edges is $\cup_{k=1}^{\widehat{k}} \binom{2^k}{2} = O(\widehat{N}^2) = O(1/s^2)$ on this event. We emphasize yet that the results of [Bordenave et al., 2018] only hold in a setting where $p = a/N$, $q = b/N$, with a, b fixed and $N \to \infty$, and we cannot turn them into a theoretical guarantee that Ω_N holds with probability close to 1.

2.5.2 Case with $K > 2$ Groups

Let us discuss the case where the number of groups K is larger than 2, still assuming that all the groups have n/K elements, with intra class probability of connection p and inter-class q. Contrary to $K = 2$, we expect in this case an information-computation gap and conjecture the following optimal rates for pair-matching.

Conjecture 1 *Define Ψ_∞^{poly} as the intersection of Ψ_∞ defined page 94, with polynomial-time algorithms. Let*

$$s_K = \frac{(p-q)^2}{q + (p-q)/K}. \tag{2.13}$$

Under Assumption (2.1), without computational constraint:

$$\inf_{\psi \in \Psi_\infty} \mathbb{E}\left[N^{bad}(\psi, T)\right] \asymp \left(\left(\frac{K\log(K)}{s_K}\right)^2 \vee \frac{K\sqrt{T}}{s_K}\right) \wedge T. \tag{2.14}$$

With polynomial time constraint:

$$\inf_{\psi \in \Psi_\infty^{poly}} \mathbb{E}\left[N^{bad}(\psi, T)\right] \asymp \left(\left(\frac{K^2}{s_K}\right)^2 \vee \frac{K\sqrt{T}}{s_K}\right) \wedge T. \tag{2.15}$$

Let us explain the heuristics leading to these rates.

For $K = 2$, a central tool to design the rate-optimal polynomial-time algorithm p.98 is the existence of polynomial-time algorithms (called `GOODCLUST` p.98) achieving non trivial classification for a cSBM$(N/2, N/2, p, q)$ when Ns is larger than some constant. When $K > 2$ and the number of nodes $N \to \infty$, for p, q scaling as $1/N$, [Bordenave et al., 2018, Abbe and Sandon, 2015, Stephan and Massoulié, 2018] provide polynomial-time algorithms $\texttt{GOODCLUST}_K^{\text{poly}}$ achieving a non trivial classification for

$$Ns_K > K^2 =: \lambda_K^{poly}.$$

Furthermore, it is conjectured [Decelle et al., 2011] that there does not exist any polynomial-time algorithm achieving non-trivial classification when $Ns_K < K^2$. The threshold λ_K^{poly} is known as the Kesten-Stigum (KS) threshold. The information theoretic threshold λ_K^{inf} for non-trivial classification is below λ_K^{poly} for $K \geq 5$. Actually, [Banks et al., 2016] have proved that $\lambda_K^{inf} \asymp K\log(K)$ and $\lambda_K^{inf} < \lambda_K^{poly}$ for $K \geq 5$, so, if the conjecture of [Decelle et al., 2011] holds, there is an information-computation gap for $K \geq 5$. A consequence of the result of [Banks et al., 2016] is that there exist algorithms $\texttt{GOODCLUST}_K^{\text{inf}}$, with exponential complexity, achieving non-trivial classification for $Ns_K = O(K\log(K))$.

Theorem 2.3.1 requires that GOODCLUST has more than non-trivial classification, it should have vanishing classification error. Several papers have established, under Assumption (2.1), the existence of algorithms $\texttt{GOODCLUST}_K^{\text{poly}}$ and $\texttt{GOODCLUST}_K^{\text{inf}}$ with misclassification proportion smaller than $\exp(-cNs_K/K)$, for some positive constant c. This result is obtained for $Ns_K \geq c'\lambda_K^{poly}$ for $\texttt{GOODCLUST}_K^{\text{poly}}$, see for example [Chin et al., 2015, Gao et al., 2017, Fei and Chen, 2019, Giraud and Verzelen, 2019] and for $Ns_K \gg \lambda_K^{inf}$ for $\texttt{GOODCLUST}_K^{\text{inf}}$, see [Zhang et al., 2016].

As a consequence, without computational constraint, a linear sampling regret is expected for any algorithm as long as the time horizon satisfies $\sqrt{2T}s_K < \lambda_K^{inf}$, or equivalently

$$T < 0.5(\lambda_K^{inf}/s_K)^2 = 0.5(K\log K/s_K)^2.$$

On the other hand, when $T \gg (K(\log K)^2/s_K)^2$, one can choose N fulfilling $\lambda_K^{inf}/s_K \ll N \leq (K\sqrt{T}/s_K)^{1/2} \ll \sqrt{T}$. Selecting N nodes uniformly at random and observing all pairs of these N nodes, $\texttt{GOODCLUST}_K^{inf}$ classifies correctly the N nodes, but a proportion at most $\exp(-cNs_K/K)$ of them. The sampling-regret for this step does not exceed the number $O(N^2) = O(K\sqrt{T}/s_K)$ of pairs sampled. Since $Ns_K/K \gg \log(K)$, the proportion of misclassified nodes among these N nodes is small and a screening procedure as in Step 2 of the algorithm p.98 can be applied in order to classify correctly $\sqrt{T}$ nodes. As an average of K/s_K queries is necessary to classify one new node, this step will have a regret scaling as $K\sqrt{T}/s_K$. Then, we can pair all nodes of the same group until the budget of T queries is spent. Hence, in the regime where $T \gg (K(\log K)^2/s_K)^2$, the final regret should be proportional to $N^2 + K\sqrt{T}/s_K \asymp K\sqrt{T}/s_K$. To sum-up the discussion, without computational constraints, one can expect a sampling-regret of order

$$\Big((K\log(K)/s_K)^2 \vee K\sqrt{T}/s_K\Big) \wedge T\,,$$

which is the conjectured rate (2.14).

Using polynomial time algorithms for clustering, the information-theoretic threshold λ_K^{inf} should be replaced by the KS-threshold λ_K^{poly}. Following the same reasoning as before, linear regret is expected as long as

$$T < 0.5(\lambda_K^{poly}/s_K)^2 = 0.5(K^2/s_K)^2.$$

On the other hand, when $\sqrt{T} \gg K^3/s_K$, one can pick N nodes at random with N fulfilling $\lambda_K^{poly}/s_K \ll N \leq (K\sqrt{T}/s_K)^{1/2} \ll \sqrt{T}$. A polynomial time algorithm $\texttt{GOODCLUST}_K^{\text{poly}}$ run with all pairs based on these nodes classifies correctly these N nodes, except for a proportion at most $\exp(-cNs_K/K)$ of them. The sampling-regret associated to this classification step is smaller than $N^2 \leq K\sqrt{T}/s_K$. The screening step classifies correctly $\sqrt{T}$ nodes with a regret $K\sqrt{T}/s_K$. The remaining budget until sampling T pairs is spent by pairing together nodes in a same estimated group. Ultimately, taking into account the computational constraint, one can expect a sampling-regret of order $((K^2/s_K)^2 \vee K\sqrt{T}/s_K) \wedge T$, which is the conjectured rate (2.15).

Appendices

2.A Proof of the Lower Bounds

2.A.1 Distributional Properties under Assumption (IL)

Recall that $\mathcal{E}$ denotes the set of all pairs in $\{1,\dots,n\}$. The invariance to labelling property enforces some invariances on the distribution of the $(N_e(\psi,T) : e \in \mathcal{E})$, with $N_e(\psi,T)$ defined by (2.3) and on the distribution of the $(N_a(\psi,T) : a = 1,\dots,n)$ with $N_a(\psi,T)$ defined by (2.4).

Let μ be a distribution in cSBM$(n/2, n/2, p, q)$ associated to a partition $G = \{G_1, G_2\}$ of $\{1,\dots,n\}$. Consider a permutation σ which leaves the partition G invariant, that is such that, either $\sigma(G_1) = G_1$ and hence $\sigma(G_2) = G_2$, or $\sigma(G_1) = G_2$ and thus $\sigma(G_2) = G_1$. Then, the distribution μ^σ defined page 92 is equal to the distribution μ. Hence the invariance to labelling property ensures that for any permutation σ leaving G invariant, the vectors $(N_e(\psi,t) : e \in \mathcal{E}; t = 1,\dots,\binom{n}{2})$ and $(N_{\sigma(e)}(\psi,T) : e \in \mathcal{E}; t = 1,\dots,\binom{n}{2})$ have the same distribution. As a consequence, the following properties holds.

Lemma 2.A.1 *When the strategy ψ fulfills the invariance to labelling property, then the random variables $(N_e(\psi,T) : e \in \mathcal{E}^{good})$ are pair-wise exchangeable. The same property holds for $(N_e(\psi,T) : e \in \mathcal{E}^{bad})$ and $(N_a(\psi,T) : a = 1,\dots,n)$.*

Proof. Let $\{a,b\}, \{a',b'\}$ denote two pairs in $\mathcal{E}^{good}$ and let σ be a G-invariant permutation such that $\sigma(\{a,b\}) = \{a',b'\}$, and $\sigma(\{a',b'\}) = \{a,b\}$. Since $\mu = \mu^\sigma$ and ψ is invariant to labelling, the random variables $(N_{\{a,b\}}, N_{\{a',b'\}})$ and $(N_{\{a',b'\}}, N_{\{a,b\}})$ have the same distribution. The same reasoning applies for pairs in $\mathcal{E}^{bad}$.

Consider now two nodes $a, b \in \{1,\dots,n\}$. Let σ be a G-invariant permutation on $\{1,\dots,n\}$ such that $\sigma(a) = b$ and $\sigma(b) = a$. Since $\mu = \mu^\sigma$ and ψ is invariant to labelling, the random variables $(N_a(\psi,T), N_b(\psi,T))$ and $(N_b(\psi,T), N_a(\psi,T))$ have the same distribution. □

2.A.2 Proof of the Lower Bound in Theorems 2.4.1 and 2.3.1

This section contains the proof of the first part of Theorem 2.4.1. The first part of Theorem 2.3.1 follows by taking $B_T = T$.

We actually prove the following stronger lower bound: when $\tilde{s} = kl(p,q) \vee kl(q,p)$ satisfies $\tilde{s} \le 1/16$, for any $\mu \in$ cSBM$(n/2, n/2, p, q)$,

$$\inf_{\psi \in \Psi_{B_T,T}} \mathbb{E}_\mu \left[N^{bad}(\psi,T) \right] \ge \frac{1}{32} \left[\frac{\sqrt{T} \vee (T/B_T)}{16\tilde{s}} \wedge T \right]. \tag{2.16}$$

The first part of Theorem 2.4.1 follows from this bound and from Lemma 2.D.3 which ensures that $s \le \tilde{s} \le 2(1+\rho^*)s$ when (2.1) holds.

Recall that $N_a(\psi,T)$ denotes the number of pairs involving the node a sampled by the strategy ψ up to time T. Let $N_a^{bad}(\psi,T)$ be the number of pairs $\{a,b\}$ with b not in the community of a sampled up to time T. Hereafter in the proof, the strategy ψ is fixed and,

to simplify notations, the dependency of N_a and N_a^{bad} on ψ is dropped out: $N_a^{bad}(\psi, T)$ is denoted $N_a(T)$ and $N_a^{bad}(\psi, T)$ is denoted $N_a^{bad}(T)$. Let also $N_a^{good}(T) = N_a(T) - N_a^{bad}(T)$. The number of between-group sampled pairs is

$$N^{bad}(T) = \frac{1}{2}\sum_{a=1}^{n} N_a^{bad}(T).$$

Let us also recall that $N_{\{a,b\}}(\psi, T) \in \{0, 1\}$ (denoted $N_{\{a,b\}}(T)$), is the number of times the pair $\{a, b\}$ has been sampled before time T. Likewise, let $N_{aB}(T) = \sum_{b\in B} N_{\{a,b\}}(T)$ be the number of times a pair between node a and the set of nodes B has been sampled before time T. For $t \geq 0$, let $\mathcal{F}_t$ be the σ-algebra gathering information available up to time t: $\mathcal{F}_t$ is the σ-algebra generated by $(\widehat{\mathcal{E}}_t, (A_e)_{e\in\widehat{\mathcal{E}}_t}, U_0, \dots, U_t)$.

The main tools for proving Equation (2.16) are the next two lemmas. The first lemma is directly adapted from [Garivier et al., 2018, Kaufmann et al., 2016].

Lemma 2.A.2 *Let $\tilde{T}$ be a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$. Let $\mu, \mu' \in cSBM(n/2, n/2, p, q)$ and let $\nu = (\nu_{ab})_{a<b}$ and $\nu' = (\nu'_{ab})_{a<b}$ denote their connection probabilities, that is $\nu_{ab} = \mu(\{a, b\} \in E)$ and $\nu_{ab} = \mu'(\{a, b\} \in E)$ for all $a, b \in V$. If $\tilde{T} \leq T$ a.s., then for any $\mathcal{F}_{\tilde{T}}$-measurable random variable $\mathcal{Z}$ taking values in $[0, 1]$,*

$$\sum_{a<b} \mathbb{E}_\mu[N_{\{a,b\}}(\tilde{T})]kl(\nu_{ab}, \nu'_{ab}) \geq kl(\mathbb{E}_\mu[\mathcal{Z}], \mathbb{E}_{\mu'}[\mathcal{Z}]), \tag{2.17}$$

where $kl(p, q) = p\log(p/q) + (1 - p)\log((1 - p)/(1 - q))$ is the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q.

Proof. The lemma follows directly from Lemma 1 in [Kaufmann et al., 2016] and Lemma 1 in [Garivier et al., 2018]. As discussed in Section 2.2.4, the pair-matching problem can be seen as a bandit problem with restrictions on the set of admissible strategies. Since Lemma 1 in [Kaufmann et al., 2016] and Lemma 1 in [Garivier et al., 2018] hold for any strategy, Inequality (2.17) holds in particular for any strategy ψ satisfying the constraints $\psi_t(\widehat{\mathcal{E}}_t, \dots) \notin \widehat{\mathcal{E}}_t$ and $N_a(t) \leq B_T$. □

While the previous lemma is only based on the bandit nature of the problem, the next lemma is based on the constraint that arms can only be sampled once.

Lemma 2.A.3 *Let M be a positive real number and consider $T \geq 1$. Then*

$$\sum_{a=1}^{n}(N_a(T) \wedge M) \geq \left((M\sqrt{T}) \vee \frac{MT}{B_T}\right) \wedge \frac{T}{2}.$$

Proof of Lemma 2.A.3. Let $S_1 = \{a : N_a(T) \leq M\}$ and $S_2 = \{a : N_a(T) > M\}$.
If $\sum_{a\in S_1} N_a(T) \geq T/2$ then $\sum_{a=1}^{n}(N_a(T) \wedge M) \geq \sum_{a\in S_1} N_a(T) \geq T/2$.

Assume now that $\sum_{a\in S_1} N_a(T) < T/2$. Since $2T = \sum_{a=1}^n N_a(T)$,

$$\begin{aligned} 2T \le T/2 + \sum_{a\in S_2} N_a(T) &= T/2 + \sum_{a\in S_2} N_{aS_1}(T) + \sum_{a\in S_2} N_{aS_2}(T) \\ &= T/2 + \sum_{a\in S_1} N_{aS_2}(T) + \sum_{a\in S_2} N_{aS_2}(T) \\ &\le T + |S_2|(B_T \wedge |S_2|). \end{aligned}$$

Hence, $|S_2| \ge \sqrt{T} \vee (T/B_T)$ and

$$\sum_{a=1}^n (N_a(T) \wedge M) \ge |S_2| M \ge (M\sqrt{T}) \vee (MT/B_T).$$

The proof is complete. □

With these two lemmas, the core inequality of the proof can be established. This inequality shows that if $N_a(t) = O(1/\tilde{s})$, then $N_a^{bad}(t)$ is of the same order of magnitude than $N_a(t)$.

Let $G = (G_1, G_2)$ be a partition of $\{1, \dots, n\}$ with $G_1 = \{1, \dots, n/2\}$ and $G_2 = \{n/2+1, \dots, n\}$. Let $\mu \in cSBM(n/2, n/2, p, q)$ be the distribution of a conditional SBM with classes G_1 and G_2, within-group connection probability p and between-group connection probability q. Unless specified, $\mathbb{E} = \mathbb{E}_\mu$ in the following.

Lemma 2.A.4 *Let M be a positive integer such that $16M\tilde{s} \le 1$ and define the stopping time $\tilde{T} = T \wedge \inf\{t : \max(N_1(t), N_n(t)) \ge M\}$. Setting $N_{1+n}(T) = N_1(T) + N_n(T)$ and $N_{1+n}^{bad}(T) = N_{1G_2}(T) + N_{nG_1}(T)$,*

$$\mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right] \ge \frac{1}{4}\mathbb{E}\left[N_{1+n}(\tilde{T})\right] \ge \frac{1}{4}\mathbb{E}\left[N_1(T) \wedge M\right]. \tag{2.18}$$

Proof of Lemma 2.A.4. The last inequality in (2.18) follows directly from

$$N_{1+n}(\tilde{T}) \ge N_1(T)\mathbf{1}_{T<\tilde{T}} + M\mathbf{1}_{T\ge\tilde{T}} \ge N_1(T) \wedge M.$$

It remains to show the first inequality. Consider the transposition $\sigma = (1, n)$ of 1 and n which switches the labels 1 and n while keeping other nodes unchanged. Let μ^σ be the distribution of $(A_{\sigma(a),\sigma(b)})_{ab}$. The partition $G^\sigma = \{G_1^\sigma, G_2^\sigma\}$ associated to μ^σ, corresponds to G with 1 and n switched, that is $G_1^\sigma = \{n, 2, \dots, n/2\}$ and $G_2^\sigma = \{n/2+1, \dots, n-1, 1\}$.

Let M be a positive integer and set

$$\mathcal{Z} = \frac{N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T})}{2(M \wedge B_T)} \in [0, 1].$$

By invariance to labelling,

$$\begin{aligned} \mathbb{E}_{\mu^\sigma}\left[N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T})\right] &= \mathbb{E}_{\mu^\sigma}\left[N_{1G_2^\sigma}(\tilde{T}) + N_{nG_1^\sigma}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right] \\ &= \mathbb{E}_{\mu}\left[N_{1G_1}(\tilde{T}) + N_{nG_2}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right]. \end{aligned}$$

Hence, setting $\tilde{M} = M \wedge B_T$, Lemma 2.A.2 ensures that,

$$\begin{aligned}
&(kl(p,q) \vee kl(q,p))\, \mathbb{E}_\mu \left[N_1(\tilde{T}) + N_n(\tilde{T}) \right] \\
&\geq kl\left(\mathbb{E}_\mu \left[N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T}) \right] / (2\tilde{M}), \mathbb{E}_{\mu^\sigma} \left[N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T}) \right] / (2\tilde{M}) \right) \\
&= kl\left(\mathbb{E}_\mu \left[N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T}) \right] / (2\tilde{M}), \mathbb{E}_\mu \left[N_{1G_1}(\tilde{T}) + N_{nG_2}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right] / (2\tilde{M}) \right) \\
&\geq \frac{1}{2(M \wedge B_T)} \frac{\left(\mathbb{E}_\mu \left[N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T}) \right] - \mathbb{E}_\mu \left[N_{1G_1}(\tilde{T}) + N_{nG_2}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right] \right)^2}{\mathbb{E}_\mu \left[N_{1G_2}(\tilde{T}) + N_{nG_1}(\tilde{T}) \right] \vee \mathbb{E}_\mu \left[N_{1G_1}(\tilde{T}) + N_{nG_2}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right]},
\end{aligned}$$

where the last line follows from Lemma 2.D.3. Setting $N_{1+n}^{good}(T) = N_{1G_1}(T) + N_{nG_2}(T)$, the last inequality can be written as

$$\begin{aligned}
2(M \wedge B_T)\tilde{s}\, \mathbb{E}\left[N_{1+n}(\tilde{T}) \right] \left(\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right] \vee \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] \right) \\
\geq \left(\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right] - \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] \right)^2. \quad (2.19)
\end{aligned}$$

If $\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) \right] \leq \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right]$, then

$$2\,\mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] \geq \mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) \right] + \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] = \mathbb{E}\left[N_{1+n}(\tilde{T}) \right]$$

and Lemma 2.A.4 follows.

Assume therefore that $\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) \right] \geq \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right]$. It follows that

$$2\,\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) \right] \geq \mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) \right] + \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] = \mathbb{E}\left[N_{1+n}(\tilde{T}) \right],$$

so Inequality (2.19) implies

$$4(M \wedge B_T)\tilde{s}\, \mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right]^2 \geq \left(\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right] - \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] \right)^2.$$

Rearranging the expression gives

$$\begin{aligned}
\mathbb{E}\left[N_{1+n}^{bad}(\tilde{T}) \right] &\geq \mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T}) \right] \left(1 - \sqrt{4(M \wedge B_T)\tilde{s}} \right) \\
&\geq \frac{1}{2}\,\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) \right] \geq \frac{1}{4}\,\mathbb{E}\left[N_{1+n}(\tilde{T}) \right],
\end{aligned}$$

since $M \wedge B_T \leq 1/(16\tilde{s})$ by assumption. The proof is complete. □

The lower bound in Theorem 2.4.1 can now be proved. Recall that for any strategy $\psi \in \Psi_{B_T,T}$, Assumption **(IL)** implies that the sampling-regret $\mathbb{E}_\mu \left[N^{bad}(\psi, T) \right]$ does not depend on $\mu \in \text{cSBM}(n/2, n/2, p, q)$, see the remark page 93. Therefore, it is sufficient to prove (2.16) for any strategy ψ invariant by labelling, with the distribution μ defined above Lemma 2.A.4.

Let M be a positive integer such that

$$1 \leq M \wedge B_T \leq \frac{1}{16\tilde{s}}.$$

First, Lemma 2.A.1 ensures that, for any pair $\{a,b\} \in \mathcal{E}^{bad}$, $\mathbb{E}\left[N_{\{a,b\}}(T)\right] = \mathbb{E}\left[N_{\{1,n\}}(T)\right]$ and hence

$$\mathbb{E}\left[N^{bad}(T)\right] = \frac{n^2}{4}\,\mathbb{E}\left[N_{\{1,n\}}(T)\right] = \frac{n}{4}\,\mathbb{E}\left[N^{bad}_{1+n}(T)\right].$$

Lemma 2.A.1 also ensures that $\mathbb{E}\left[N_a(T) \wedge M\right] = \mathbb{E}\left[N_1(T) \wedge M\right]$ for all $a \in \{1,\dots,n\}$. By Lemma 2.A.4, it follows that

$$\begin{aligned} 16\,\mathbb{E}\left[N^{bad}(T)\right] = 4n\,\mathbb{E}\left[N^{bad}_{1+n}(T)\right] &\geq 4n\,\mathbb{E}\left[N^{bad}_{1+n}(\tilde{T})\right] \\ &\geq n\,\mathbb{E}\left[N_1(T) \wedge M\right] = \sum_{a=1}^{n} \mathbb{E}\left[N_a(T) \wedge M\right]. \end{aligned}$$

Hence, by Lemma 2.A.3

$$16\,\mathbb{E}\left[N^{bad}(T)\right] \geq \left((M\sqrt{T}) \vee \frac{MT}{B_T}\right) \wedge \frac{T}{2}.$$

For $\tilde{s} \leq 1/16$, taking M equal to the integer part of $1/(16\tilde{s})$ gives

$$16\,\mathbb{E}\left[N^{bad}(T)\right] \geq \left(\frac{\sqrt{T}}{32\tilde{s}} \vee \frac{T}{32\tilde{s}B_T}\right) \wedge \frac{T}{2}.$$

Since the sampling-regret does not depend on the choice of μ, the proof is complete.

2.B Proof of the Unconstrained Upper Bound

This section proves the following result, from which follows the upper bound of Theorem 2.3.1, as explained below Theorem 2.B.1.

Theorem 2.B.1 *There exist numerical constants $c_1, c_2 > 0$, such that, for any $T \leq c_2 n^2$, with probability at least $1 - 13/T$, the algorithm described in Section 2.3.2 fulfills*

$$N^{bad}(\psi, T) \leq c_1 \left(T \wedge \frac{\sqrt{T}}{s}\right).$$

Let us explain how the upper bound of Theorem 2.3.1 follows from Theorem 2.B.1. First, let us note that the upper bound of Theorem 2.B.1 also holds in expectation. Indeed, since $N^{bad}(\psi, T) \leq T$, the algorithm described in Section 2.3.2 fulfills

$$\mathbb{E}\left[N^{bad}(\psi, T)\right] \leq c_1 \left(T \wedge \frac{\sqrt{T}}{s}\right) + 13 \leq c_1' \left(T \wedge \frac{\sqrt{T}}{s}\right). \tag{2.20}$$

Second, we can get an horizon free algorithm by applying a doubling trick. For any integer l, let $t_l = 2^l$. At each time t_l, discard all nodes and pairs involved in the previous iterations of the algorithm and restart the algorithm described in Section 2.3.2 with time horizon $t_{l+1} - t_l$. The resulting strategy does not depend on any time horizon. Let us prove that this horizon-free algorithm also has a $O\left(T \wedge (\sqrt{T}/s)\right)$ sampling regret. The argument for this proof is classical: according to the upper bound (2.20), for any $t_{l-1} \leq T \leq t_l < c_2 n^2$,

$$\begin{aligned}
\mathbb{E}\left[N^{bad}(\psi, T)\right] &\leq c_1 \left(\frac{\sqrt{t_0}}{s} \wedge t_0 + \frac{\sqrt{t_1 - t_0}}{s} \wedge (t_1 - t_0) + ... + \frac{\sqrt{t_l - t_{l-1}}}{s} \wedge (t_l - t_{l-1}) \right) \\
&\leq c_1 \left(\frac{1}{s} + \frac{1}{s} \sum_{r=0}^{l-1} 2^{r/2} \right) \wedge t_l \\
&\leq c_1 \left(\frac{\sqrt{t_l}}{(\sqrt{2}-1)s} \wedge t_l \right) \leq 4c_1 \left(\frac{\sqrt{T}}{s} \wedge T \right).
\end{aligned}$$

Hence, we have proved that the upper bound of Theorem 2.3.1 is a consequence of Theorem 2.B.1.

The proof of Theorem 2.B.1 is quite lengthy. To help the reader to understand the organization of this demonstration, the section starts with a sketch of proof.

2.B.1 Outline of the Proof of Theorem 2.B.1

As any strategy has at most linear regret, it is sufficient to prove that there exist two positive numerical constants c_{thresh} and c_1 such that, for any $T \geq c_{\text{thresh}}/s^2$, the number $N^{bad}(\psi, T)$ of pairs sampled among $\mathcal{E}^{\text{bad}}$ by the strategy ψ described in the algorithm p.98 in Section 2.3.2 is smaller than $c_1\sqrt{T}/s$ with probability at least $1 - 13/T$. As a consequence, in the proof, without loss of generality, it is assumed that $T \geq c_{\text{thresh}}/s^2$, for a sufficiently large constant c_{thresh}. To prove the theorem, it is sufficient to show that neither Steps 1., 2. nor 3. of the algorithm sample more than $O(\sqrt{T}/s)$ "bad" pairs, where a bad pair involves one node from community 1 and one from community 2.

Step 1. In the first step, the algorithm samples at random a kernel $\mathcal{N}$ of $N = \lceil \sqrt{T}/\log(s\sqrt{T}) \rceil$ nodes (point 1. in the algorithm). In this kernel, with large probability, at least $\lceil N/4 \rceil$ nodes from each communities are sampled. This result follows from Hoeffding's concentration inequality for hypergeometric random variables, it is rigorously established in point 2 of Lemma 2.B.2.

Each pair of the kernel is sampled with probability proportional to $\sqrt{T}/\left(s\binom{N}{2}\right)$ (point 2. of the algorithm). With high probability, the set of sampled pairs $\mathcal{O}_0$ has cardinality $|\mathcal{O}_0| \asymp \sqrt{T}/s$, see point 3. of Lemma 2.B.2. At this point, the observed graph follows a cSBM with connection probabilities $\widetilde{p} \asymp p\sqrt{T}/\left(s\binom{N}{2}\right)$ and $\widetilde{q} \asymp q\sqrt{T}/\left(s\binom{N}{2}\right)$. By (2.7), setting $\widetilde{s} = (\widetilde{p} - \widetilde{q})^2/(\widetilde{p} + \widetilde{q})$ the proportion of misclassified nodes by `GOODCLUST` is upper bounded by

$$\exp(-c_1^{\text{GC}} N\widetilde{s}) = \exp\left(-c\frac{\sqrt{T}/s}{N}s\right) = \exp(-\log(s\sqrt{T})) \leq \frac{1}{Ns},$$

with probability at least $1 - c_2^{\mathrm{GC}}/N^3$. In particular, at most $1/s$ nodes of the kernel are misclassified. A rigorous proof of this last statement is provided in point 4 of Lemma 2.B.2.

Let us comment briefly the choice of the cardinalities N of the kernel and $|\mathcal{O}_0|$ of the sampled pairs in this first step of the algorithm. These are chosen to guarantee the following properties.

(1.i) N is sufficiently large to make the probability c^{GC}/N^3 small and, on the other hand, N is sufficiently small so that one can classify a large proportion of $\mathcal{N}$ with less than $|\mathcal{O}_0| = O(\sqrt{T}/s)$ observed pairs.

(1.ii) $|\mathcal{O}_0|$ is large enough to ensure that the proportion of misclassified nodes in $\mathcal{N}$ satisfies $\exp\big(-c_1^{\mathrm{GC}}\,\mathbb{E}[|\mathcal{O}_0|]s/N\big) \leq 1/(Ns)$.

(1.iii) On the other hand, $|\mathcal{O}_0|$ is small enough, namely $|\mathcal{O}_0| = O(\sqrt{T}/s)$, to ensure a regret $O(\sqrt{T}/s)$ in this exploratory phase of the algorithm.

Before moving to the screening step 2 of the algorithm, the estimator

$$\hat{\tau} = \frac{1}{|\mathcal{O}_0|} \sum_{\{x,x'\}\in\mathcal{O}_0} A_{\{x,x'\}}$$

of $\tau = (p+q)/2$ is shown to satisfy, with large probability,

$$|\hat{\tau} - p| \wedge |\hat{\tau} - q| \geq |\hat{\tau} - \tau|.$$

This property is obtained by a careful application of Bernstein inequality for hypergeometric random variables in point 5 of Lemma 2.B.2. This estimation of τ is sufficient for the screening step.

Step 2. The second step of the algorithm samples uniformly at random a set $\mathcal{A}_0$ of $\lceil 8\sqrt{2T}\rceil$ nodes. These nodes are screened with the following objectives.

(2.i) A set of at least $\lceil\sqrt{2T}\rceil$ nodes among $\mathcal{A}_0$ are selected containing at most $1/s$ members of community 2.

(2.ii) A set of at most $O(\sqrt{T}/s)$ bad pairs is sampled during this screening.

Claims (2.i) and (2.ii) are formally established in Lemma 2.B.3, Claim (2.i) in points 8 and 10 and Claim (2.ii) at point 9.

The main tool for proving these two properties is Lemma 2.B.4. It ensures that the probability that a node from community 2 is not removed after i steps of screening decreases exponentially fast with i. Therefore, after $I \asymp \log(s\sqrt{T})$ screening steps, each node from community 2 remains with probability at most $e^{-c\log(s\sqrt{T})}$. Since there are $O(\sqrt{T})$ nodes in $\mathcal{A}_0$, the expected number of remaining nodes from community 2 is upper bounded, when $T \gtrsim 1/s^2$, by

$$O\big(\sqrt{T}e^{-c\log(s\sqrt{T})}\big) \lesssim \frac{1}{s}.$$

The same bound holds with high probability. Similar arguments are used to obtain that, with large probability, less than $\lceil 8\sqrt{2T} \rceil - \lceil \sqrt{2T} \rceil$ nodes are removed during the screening step, which shows property (2.i).

The proof of Property (2.ii) is more involved. At step 7(b) of the algorithm, a bad pair is sampled when it involves either

(2.ii.a) a node of community 2 and a well classified node of the kernel,

(2.ii.b) a node of community 1 and a misclassified node of the kernel.

The number of pairs in the case (2.ii.a) is simply bounded from above by $|\mathcal{A}_0| = O(\sqrt{T})$ multiplied by the number of misclassified nodes in the kernel. We have checked in step 1, that the number of misclassified nodes in the kernel is bounded from above by $O(1/s)$. So, on this event, the number of such bad pairs is at most $O(\sqrt{T} \times 1/s)$.

The number of pairs in the case (2.ii.b) is bounded from above as follows. During each screening step (point 7.), a node is queried $k = O(1/s)$ times. Thus, the number of queries of a node from community 2 during this screening step is k times the number of screening steps before it is removed. Recall that, from Lemma 2.B.4, the probability that a node of community 2 remains after i screening steps decreases exponentially fast with i. Hence, the expected number of queries of a node from community 2 is bounded from above by

$$k \times \sum_{i \geq 1} e^{-ci} = O(k) = O(1/s)\,.$$

The number of sampled pairs in case (2.ii.b) is smaller than the total number of queries on nodes from community 2 in $\mathcal{A}_0$, which is smaller than $O(|\mathcal{A}_0|k) = O(\sqrt{T}/s)$. This bound also holds with high probability, which proves property (2.ii.b).

Step 3. During Step 3. of the algorithm, pairs within $\mathcal{A}_I$ are sampled until T pairs have been sampled overall. On the event where $|\mathcal{A}_I|$ is larger than $\sqrt{2T}$, this sampling is possible. In addition, on the event where the number of nodes from community 2 in $\mathcal{A}_I$ is upper bounded by $1/s$, the number of bad pairs in $\mathcal{A}_I$ is smaller than

$$O(|\mathcal{A}_I|/s) = O(|\mathcal{A}_0|/s) = O(\sqrt{T}/s).$$

2.B.2 Proof of Theorem 2.B.1

All we need is to prove that there exists a numerical constant $c_{\text{thresh}} \geq 1$, such that, for any $T \geq c_{\text{thresh}}/s^2$, the upper bound $N^{bad}(\psi, T) \leq c_1\sqrt{T}/s$ holds with probability at least $1 - 13/T$. We focus then on the case where $T \geq c_{\text{thresh}}/s^2$.

Denote by $\hat{G} = \{\hat{G}_1, \hat{G}_2\}$ the partition of $\mathcal{N}$ output by the `GOODCLUST` algorithm and by $S\Delta S'$ the symmetric difference between two sets S, S'. Define the community labelling vectors Z and $\hat{Z}$ by $Z_x = j$ for all $x \in G_j$ and $\hat{Z}_x = j$ for all $x \in \hat{G}_j$. The following lemma controls the first step of the algorithm.

Lemma 2.B.2 *There exists numerical constants $c_{thresh} \geq e$ and $T_0 \geq 1$ such that, if $T_0 \leq T \leq n^2/16$ and $s\sqrt{T} \geq c_{thresh}$, then with probability at least $1 - 9/T$:*

1. *only a small part of the nodes has been sampled: $N \leq \frac{n}{4}$;*
2. *the two communities of the sampled nodes are approximately balanced, that is $N_1 \wedge N_2 \geq N/4$, where $N_j := |\{Z = j\} \cap \mathcal{N}|$ is the number of nodes from community j in $\mathcal{N}$;*
3. *the cardinality of the sample pairs fulfills $\frac{c_{\mathcal{O}_0}\sqrt{T}}{2s} \leq |\mathcal{O}_0| \leq \frac{3\, c_{\mathcal{O}_0}\sqrt{T}}{2s}$;*
4. *the fraction of misclassified nodes is upper bounded by*

$$\varepsilon_N = \inf_{\pi \text{ permutation on } \{1,2\}} \frac{1}{2N} \sum_{k=1}^{2} |\{Z = k\} \Delta \{\hat{Z} = \pi(k)\}| \leq \frac{1}{sN};$$

5. $|\hat{\tau} - \frac{p+q}{2}| \leq \frac{p-q}{4}$.

We refer to Section 2.B.3.1 for a proof of this lemma.

At the end of the first step, $|\mathcal{O}_0| = O(\frac{\sqrt{T}}{s})$ pairs have been sampled according to point 3 of Lemma 2.B.2, thus resulting in a number of sampled bad pairs $O(\frac{\sqrt{T}}{s})$. Let us now turn to the second step of the algorithm.

Assume without loss of generality that the community labelling $\hat{Z}$ of the nodes in $\mathcal{N}$ is mostly in agreement with Z, i.e. the infimum in the definition of ε_N is achieved for the identity permutation:

$$\varepsilon_N = \frac{1}{2N} \sum_{k=1}^{2} |\{Z = k\} \Delta \{\hat{Z} = k\}|.$$

If it is not the case, the remaining of the proof still holds but with $\{Z = 1\}$ replaced by $\{Z = 2\}$.

For each $x \in \mathcal{A}_0$, the (distinct) nodes $\{y_1^x, \dots, y_{kI}^x\}$ are sampled uniformly at random in $\mathcal{N} \cap \left\{\hat{Z} = 1\right\}$. Let $\mathcal{V}_{x,0} = \emptyset$ for all $x \in \mathcal{A}_0$ and $\mathcal{V}_{x,i} = \{y_1^x, \dots, y_{ki}^x\}$ for $i = 1, \dots, I$. Note that $|\mathcal{V}_{x,j}| = kj$ for all $x \in \mathcal{A}_0$. By induction, construct the sequences of sets $(\mathcal{A}_i)_{0 \leq i \leq I}$, which contain the "active" nodes remaining at each iteration, and $(\mathcal{O}_i)_{0 \leq i \leq I}$, which contain the sampled pairs.

More formally, for $i \geq 1$ and all $x \in \mathcal{A}_{i-1}$, the pairs $\{\{x, y^x_{(i-1)k+a}\}, 1 \leq a \leq k\}$ are observed at iteration i, so that

$$\mathcal{O}_i = \mathcal{O}_{i-1} \cup \bigcup_{x \in \mathcal{A}_{i-1}} \{\{x, y^x_{(i-1)k+a}\}, 1 \leq a \leq k\}.$$

We remind the reader that we estimate the connectivity between x and community 1 by

$$\hat{p}_{x,i} = \frac{1}{ki} \sum_{y \in \mathcal{V}_{x,i}} A_{x,y}$$

and only keep the nodes whose estimated connectivity is large enough in the active set:

$$\mathcal{A}_i = \{x \in \mathcal{A}_{i-1} : \hat{p}_{x,i} \geq \hat{\tau}\}\,. \tag{2.21}$$

After I iterations, the total number of sampled pairs is

$$|\mathcal{O}_I| = |\mathcal{O}_0| + k\sum_{i=0}^{I-1} |\mathcal{A}_i|$$

and the number of sampled bad pairs from this step is upper bounded by

$$k\sum_{i=0}^{I-1} |\mathcal{A}_i \cap \{Z \neq 1\}| + |\mathcal{A}_0 \cap \{Z = 1\}| \times |\mathcal{N} \cap \{\hat{Z} \neq Z\}| \tag{2.22}$$

where the first term comes from the pairs connecting community 2 to the kernel and the second term comes from the pairs connecting community 1 to a misclassified vertex of the kernel.

The following lemma controls this screening step.

Lemma 2.B.3 *There exists numerical constants T_0', c'_{thresh} larger than 1 such that if $T_0' \leq T \leq (\frac{3n}{64\sqrt{2}})^2$ and $s\sqrt{T} \geq c'_{thresh}$, then with probability at least $1 - 13/T$, Lemma 2.B.2 holds and*

6. *the algorithm does not run out of connections with the kernel of the first step: $kI \leq N$;*
7. *it is possible to take $|\mathcal{A}_0|$ new vertices: $|\mathcal{A}_0| \leq \frac{3n}{4} \leq n - N$;*
8. *few vertices from the wrong community remain: $|\mathcal{A}_I \cap \{Z \neq 1\}| \leq \frac{1}{s}$;*
9. *the number of sampled bad pairs from nodes in the wrong community is controlled: $k\sum_{i=0}^{I-1} |\mathcal{A}_i \cap \{Z \neq 1\}| \leq C_{fail}\frac{\sqrt{T}}{s}$ for a numerical constant C_{fail};*
10. *enough vertices from community 1 remain for the next step: $|\mathcal{A}_I \cap \{Z = 1\}| \geq \sqrt{2T}$.*

We refer to Section 2.B.3.2 for a proof of this lemma.

Equation (2.22) together with point 9 of Lemma 2.B.3 and point 4 of Lemma 2.B.2 entail that the number of sampled bad pairs during the screening step is again $O(\sqrt{T}/s)$.

Finally, during the last step, the algorithm uses the remaining budget to observe pairs uniformly at random between vertices of $\mathcal{A}_I$. Point 10 of Theorem 2.B.3 ensures that the number of possible pairs is larger than $T - \sqrt{T/2}$, which allows to spend the whole budget (since at least $\lceil\sqrt{T/2}\rceil$ pairs have been observed in the previous steps), and point 8 ensures that the number of sampled bad pairs of this step is again $O(\frac{\sqrt{T}}{s})$.

Hence, the total number of bad pairs sampled during the whole process is $O(\sqrt{T}/s)$.

2.B.3 Proofs of the Technical Lemmas

2.B.3.1 Proof of Lemma 2.B.2

The proof of point 1 is straightforward: since $\sqrt{T} \leq n/4$ by assumption, the condition $N \leq n/4$ holds as soon as $N \leq \sqrt{T}$, that is $\lceil \frac{\sqrt{T}}{\log(s\sqrt{T})} \rceil \leq \sqrt{T}$ by definition of N. Therefore point 1 holds true as soon as $s\sqrt{T} \geq c_{thresh}$ for some numerical constant c_{thresh}.

Proof of point 2. There are only two communities, so it is enough to consider the first one. Since the communities are balanced, the number N_1 of nodes from community 1 in the kernel follows an hypergeometric distribution with parameters $(N, 1/2, n)$. Therefore,

$$\mathbb{P}\left(\left|N_1 - \frac{N}{2}\right| \geq \sqrt{2N\log N}\right) \leq \frac{2}{N^4}$$

using Equation (2.41). Since $N = \left\lceil \frac{\sqrt{T}}{\log(s\sqrt{T})} \right\rceil$,

$$\frac{2}{N^4} \leq \frac{2\log(s\sqrt{T})^4}{T^2} \leq \frac{(\log T)^4}{8T^2}$$

using $s \leq 1$, which is upper bounded by $1/T$ for all $T \geq 1$. Assuming $s\sqrt{T} \geq c_{thresh}$ for some numerical constant $c_{thresh} \geq e$, one has $\frac{\sqrt{T}}{\log\sqrt{T}} \leq N \leq \sqrt{T}$, so that

$$\begin{aligned}\frac{\sqrt{2N\log N}}{N/4} &\leq 4\sqrt{2}\sqrt{\frac{\log(\sqrt{T})}{\sqrt{T}/\log\sqrt{T}}} \\ &\leq 4\sqrt{2}\sqrt{\frac{\log(\sqrt{T})^2}{\sqrt{T}}}.\end{aligned}$$

Therefore, it is smaller than 1 as soon as $T \geq T_{0,2}$ for some numerical constant $T_{0,2}$, which entails

$$\mathbb{P}\left(\left|N_1 - \frac{N}{2}\right| \geq \frac{N}{4}\right) \leq \frac{1}{T},$$

and the same for N_2.

Proof of point 3. The number $|\mathcal{O}_0|$ of sampled pairs in the kernel $\mathcal{N}$ follows a binomial distribution with parameters $\left(\binom{N}{2}, c_{\mathcal{O}_0}\sqrt{T}/s\binom{N}{2}\right)$. Therefore,

$$\mathbb{P}\left(\left||\mathcal{O}_0| - c_{\mathcal{O}_0}\frac{\sqrt{T}}{s}\right| \geq \sqrt{2c_{\mathcal{O}_0}\frac{\sqrt{T}}{s}\log(2T) + \log(2T)}\right) \leq \frac{1}{T}$$

using Bernstein's inequality (2.43). This implies that

$$\frac{1}{2}c_{\mathcal{O}_0}\frac{\sqrt{T}}{s} \leq |\mathcal{O}_0| \leq \frac{3}{2}c_{\mathcal{O}_0}\frac{\sqrt{T}}{s} \tag{2.23}$$

as soon as $T \geq T_{0,3}$ for some numerical constant $T_{0,3}$.

Let us check that the probability parameter of the binomial distribution is well defined, that is, the condition $c_{\mathcal{O}_0}\sqrt{T}/s\binom{N}{2} \in [0,1]$ is satisfied. One can show that $N \geq 8$ as soon as $T \geq T_{0,3}$ for some numerical constant $T_{0,3}$. Then

$$\binom{N}{2} \geq \frac{N^2}{4}$$

so that the condition holds as soon as $c_{\mathcal{O}_0}\sqrt{T}/s \leq N^2/4$, which is implied by

$$c_{\mathcal{O}_0}\sqrt{T}/s \leq \frac{1}{4}\frac{T}{(\log(s\sqrt{T}))^2},$$

or equivalently

$$\frac{s\sqrt{T}}{(\log(s\sqrt{T}))^2} \geq 4c_{\mathcal{O}_0}.$$

Hence, the condition $c_{\mathcal{O}_0}\sqrt{T}/s\binom{N}{2} \in [0,1]$ holds as soon as $s\sqrt{T} \geq c_{thresh}$ for some numerical c_{thresh}. This, together with (2.23), concludes the proof of point 3.

Proof of point 4. Since each pair of $\mathcal{N}$ is sampled with probability $c_{\mathcal{O}_0}\sqrt{T}/s\binom{N}{2}$, the matrix $\widetilde{A}$ defined by $\tilde{A}_{x,x'} = A_{x,x'}$ if the pair $\{x,x'\}$ has been sampled and zero otherwise has the same distribution as the adjacency matrix of a fully observed SBM with connection probabilities $\widetilde{p} = p\, c_{\mathcal{O}_0}\sqrt{T}/s\binom{N}{2}$ and $\widetilde{q} = q\, c_{\mathcal{O}_0}\sqrt{T}/s\binom{N}{2}$. Therefore, the proportion ε_N of misclassified nodes in $\mathcal{N}$ by the `GOODCLUST` algorithm is upper bounded by

$$\varepsilon_N \leq \exp\left(-c_1^{\mathrm{GC}} N \frac{(\widetilde{p}-\widetilde{q})^2}{\widetilde{p}+\widetilde{q}}\right) \tag{2.24}$$

with probability at least $1 - c^{\mathrm{GC}}/N^3$. Hence with probability at least $1 - 1/T$

$$\varepsilon_N \leq \exp\left(-2\, c_1^{\mathrm{GC}} c_{\mathcal{O}_0} \frac{\log(s\sqrt{T})}{2}\right)$$

using $N := \lceil \frac{\sqrt{T}}{\log(s\sqrt{T})} \rceil \leq 2\frac{\sqrt{T}}{\log(s\sqrt{T})}$ as soon as $T \geq T_{0,4}$ for some numerical constant $T_{0,4}$. Hence, by taking $c_{\mathcal{O}_0} \geq 1/(c_1^{\mathrm{GC}})$, one has with probability at least $1 - 1/T$

$$\varepsilon_N \leq \exp\left(-\log(s\sqrt{T})\right) = \frac{1}{s\sqrt{T}}$$

so that

$$\varepsilon_N \leq \frac{1}{sN}$$

as soon as $N \leq \sqrt{T}$, which holds true when $s\sqrt{T} \geq c_{thresh}$ for some numerical constant c_{thresh}.

Proof of point 5. Let $\mathcal{O}_{\text{within}} := \mathcal{O}_0 \cap \mathcal{E}^{good}$ be the subset of within-group pairs, and $\mathcal{O}_{\text{out}} := \mathcal{O}_0 \setminus \mathcal{O}_{\text{within}}$ the subset of pairs between two different communities. Then

$$\hat{\tau} = \frac{|\mathcal{O}_{\text{within}}|}{|\mathcal{O}_0|} \frac{1}{|\mathcal{O}_{\text{within}}|} \sum_{(x,x')\in\mathcal{O}_{\text{within}}} A_{x,x'} + \frac{|\mathcal{O}_{\text{out}}|}{|\mathcal{O}_0|} \frac{1}{|\mathcal{O}_{\text{out}}|} \sum_{(x,x')\in\mathcal{O}_{\text{out}}} A_{x,x'}.$$

Conditionally to the number of sampled pairs $|\mathcal{O}_0|$ and the number of within-group pairs $|\mathcal{O}_{\text{within}}|$, the sum $\sum_{(x,x')\in\mathcal{O}_{\text{within}}} A_{x,x'}$ (resp. $\sum_{(x,x')\in\mathcal{O}_{\text{out}}} A_{x,x'}$) is independent of $\mathcal{O}_0$, and is a sum of i.i.d. Bernoulli random variables with parameter p (resp. q). Therefore, Bernstein's inequality (2.42) ensures that with probability at least $1 - 4/T$

$$\frac{|\mathcal{O}_{\text{within}}|}{|\mathcal{O}_0|} \left| \frac{1}{|\mathcal{O}_{\text{within}}|} \sum_{(x,x')\in\mathcal{O}_{\text{within}}} A_{x,x'} - p \right| \le \sqrt{2p\frac{\log T}{|\mathcal{O}_0|}} + \frac{\log T}{|\mathcal{O}_0|}$$

$$\text{and} \quad \frac{|\mathcal{O}_{\text{out}}|}{|\mathcal{O}_0|} \left| \frac{1}{|\mathcal{O}_{\text{out}}|} \sum_{(x,x')\in\mathcal{O}_{\text{out}}} A_{x,x'} - q \right| \le \sqrt{2q\frac{\log T}{|\mathcal{O}_0|}} + \frac{\log T}{|\mathcal{O}_0|}.$$

Using point 3, one has $|\mathcal{O}_0| \ge c_{\mathcal{O}_0}\sqrt{T}/(2s)$ with probability at least $1 - 1/T$, so that

$$\begin{aligned}\left| \hat{\tau} - \left(\frac{|\mathcal{O}_{\text{within}}|}{|\mathcal{O}_0|} p + \frac{|\mathcal{O}_{\text{out}}|}{|\mathcal{O}_0|} q \right) \right| &\le 2\sqrt{2ps\frac{\log T}{c_{\mathcal{O}_0}\sqrt{T}/2}} + 2s\frac{\log T}{c_{\mathcal{O}_0}\sqrt{T}/2} \\ &\le 2(p-q)\sqrt{2\frac{\log T}{c_{\mathcal{O}_0}\sqrt{T}/2}} + 2(p-q)\frac{\log T}{c_{\mathcal{O}_0}\sqrt{T}/2}\end{aligned}$$

with probability at least $1 - 5/T$, using $s = (p-q)^2/p \le p - q$. Finally, since $c_{\mathcal{O}_0}/2 \ge 1$ and $|\mathcal{O}_{\text{within}}| = |\mathcal{O}_0| - |\mathcal{O}_{\text{out}}|$,

$$\begin{aligned}\left| \hat{\tau} - \frac{p+q}{2} \right| &\le \left| \frac{|\mathcal{O}_{\text{within}}|}{|\mathcal{O}_0|} p + \frac{|\mathcal{O}_{\text{out}}|}{|\mathcal{O}_0|} q - \frac{p+q}{2} \right| + 2(p-q)\sqrt{2\frac{\log T}{\sqrt{T}}} + 2(p-q)\frac{\log T}{\sqrt{T}} \\ &\le \left| \left(2\frac{|\mathcal{O}_{\text{out}}|}{|\mathcal{O}_0|} - 1 \right) \frac{p-q}{2} \right| + \frac{|p-q|}{16} \qquad (2.25)\end{aligned}$$

as soon as $T \ge T_{0,4}$ for some numerical constant $T_{0,4}$.

Conditionally to the number of pairs $|\mathcal{O}_0|$ and the sizes N_1 and N_2 of the two communities sampled in $\mathcal{N}$, the number $|\mathcal{O}_{\text{out}}|$ of between group pairs follows an hypergeometric distribution with parameters $\left(|\mathcal{O}_0|, r, \binom{N}{2}\right)$ with $r = N_1N_2/\binom{N}{2}$. Conditionally to $|\mathcal{O}_0|$ and the event $\frac{3}{8} \le r \le \frac{5}{8}$, the random variable $|\mathcal{O}_{\text{out}}|$ dominates stochastically an hypergeometric random variable with parameters $(|\mathcal{O}_0|, \frac{3}{8}, \binom{N}{2})$ and it is stochastically dominated by an hypergeometric random variable with parameters $(|\mathcal{O}_0|, \frac{5}{8}, \binom{N}{2})$. There exists a real $\gamma > 0$ such that $N_1 = \gamma N$ and $N_2 = (1-\gamma)N$ so that

$$r = \frac{\gamma N(1-\gamma)N}{N(N-1)/2} = 2\gamma(1-\gamma)(1 + \frac{1}{N-1}) = 2\gamma(1-\gamma)(1 + \frac{1}{\frac{\sqrt{T}}{\log(s\sqrt{T})} - 1})$$

Using point 2, one has with probability at least $1 - 1/T$ that $\frac{1}{4} \leq \gamma \leq \frac{3}{4}$ which entails $\frac{3}{8} \leq r \leq \frac{5}{8}$ as soon as $T \geq T_{0,5}$ for some numerical constant $T_{0,5}$. Therefore,

$$\mathbb{P}\left(|\mathcal{O}_{\text{out}}| \leq \frac{3|\mathcal{O}_0|}{8} - \sqrt{\frac{|\mathcal{O}_0| \log T}{2}}\right) \leq \frac{1}{T} + \frac{1}{T}$$

using Equation (2.41), and similarly

$$\mathbb{P}\left(|\mathcal{O}_{\text{out}}| \geq \frac{5|\mathcal{O}_0|}{8} + \sqrt{\frac{|\mathcal{O}_0| \log T}{2}}\right) \leq \frac{1}{T} + \frac{1}{T}.$$

Using point 3, one has with probability at least $1-1/T$ that $|\mathcal{O}_0| \geq c_{\mathcal{O}_0}\sqrt{T}/(2s)$ which entails $\sqrt{\frac{|\mathcal{O}_0| \log T}{2}} \leq \frac{|\mathcal{O}_0|}{16}$ as soon as $T \geq T_{0,6}$ for some numerical constant $T_{0,6}$. Hence

$$\frac{5}{16} \leq \frac{|\mathcal{O}_{\text{out}}|}{|\mathcal{O}_0|} \leq \frac{11}{16}$$

with probability $1 - \frac{5}{T}$. This, together with Equation (2.25), concludes the proof of point 5 (which holds with probability $1 - \frac{8}{T}$).

2.B.3.2 Proof of Lemma 2.B.3

Proof of point 6 and point 7. There exists a constant c'_{thresh} such that $4C_I C_k \leq \frac{s\sqrt{T}}{(\log(s\sqrt{T}))^2}$ as soon as $s\sqrt{T} \geq c'_{\text{thresh}}$. It follows that $kI \leq N$.

Point 7 follows from straightforward algebra.

Proof of point 8. For all $x \in \mathcal{A}_0$, denote by $T_x = \max\{i : x \in \mathcal{A}_i\}$ the index of the last iteration where the vertex x was in the active set. Let us first show that if x is not in the first community, then T_x has sub-exponential tails.

Lemma 2.B.4 *Set $\rho' = 1/2000$. If $C_k \geq (\log 3)/\rho'$ then*

$$\forall i \in \mathbb{N}^* \qquad \mathbb{P}(T_x \geq i) \leq e^{-\rho' C_k i}. \tag{2.26}$$

We refer to Section 2.B.4 for a proof of this lemma.

Let us now prove point 8. Let $T^{(1)} = |\mathcal{O}_0|$ and $V_x = \mathbf{1}_{T_x \geq I}$. Conditionally on $\mathcal{F}_{T^{(1)}}$, the variables $(V_x)_{x \in \mathcal{A}_0 \cap \{Z \neq 1\}}$ are i.i.d. Bernoulli random variables with parameter $r \leq e^{-\rho' C_k I}$ by equation (2.26). Therefore, for all $i \in \mathbb{N}$,

$$\mathbb{P}\left(\sum_{x \in \mathcal{A}_0 \cap \{Z \neq 1\}} V_x = i\right) \leq \frac{|\mathcal{A}_0|^i}{i!} r^i$$

so that

$$\begin{aligned}\mathbb{P}(|\mathcal{A}_I \cap \{Z \neq 1\}| \geq i) &\leq \sum_{j \geq i} \frac{|\mathcal{A}_0|^j}{j!} r^j \\ &\leq \frac{(|\mathcal{A}_0|r)^i}{i!} \sum_{j \geq 0} \left(\frac{|\mathcal{A}_0|r}{i}\right)^j \\ &\leq 2\frac{(|\mathcal{A}_0|r)^i}{i!}\end{aligned}$$

as soon as $i \geq 2|\mathcal{A}_0|r$. For $i = \lceil 1/s \rceil$, this condition holds if $16\sqrt{2} \leq (s\sqrt{T})^{\rho' C_k C_I - 1}$ which holds when $C_I C_k \geq 4/\rho'$ and $s\sqrt{T} \geq c'_{th}$.

Taking $i = \lceil 1/s \rceil$ and using that $i! \geq (i/e)^i$ for all $i \geq 1$, it follows that

$$\mathbb{P}\left(|\mathcal{A}_I \cap \{Z \neq 1\}| \geq \frac{1}{s}\right) \leq 2\left(\frac{e|\mathcal{A}_0|r}{\lceil 1/s \rceil}\right)^{\lceil 1/s \rceil} \leq 2\left(se|\mathcal{A}_0|r\right)^{1/s}$$

as soon as $se|\mathcal{A}_0|r \leq 1$.

We want to take r small enough such that $2(se|\mathcal{A}_0|r)^{1/s} \leq 1/T$, that is

$$\log(se|\mathcal{A}_0|) + s\log(2T) \leq (-\log r),$$

which holds as soon as

$$\rho' C_k I \geq \log(s\sqrt{T}) + \log(32e\sqrt{2}) + s\log T.$$

using $|\mathcal{A}_0| \leq 16\sqrt{2T}$.

Note that $\frac{s\log T}{\log(s\sqrt{T})} = 2\frac{s\sqrt{T}}{\log(s\sqrt{T})}\frac{\log\sqrt{T}}{\sqrt{T}} \leq 2$, since $\log(x)/x$ is decreasing for $x > e$ and $s\sqrt{T} \leq \sqrt{T}$, so that there exists a numerical constant c'_{thresh} such that if $s\sqrt{T} \geq c'_{\text{thresh}}$, then point 8 is implied by

$$\rho' C_k I \geq 4\log(s\sqrt{T}),$$

which holds when $C_I C_k \geq 4/\rho'$.

Proof of point 9. Note that

$$k\sum_{i=0}^{I-1} |\mathcal{A}_i \cap \{Z \neq 1\}| = k \sum_{x \in \mathcal{A}_0 \cap \{Z \neq 1\}} T_x.$$

Conditionally on $\mathcal{A}_0$, the random variables $(T_x)_{x \in \mathcal{A}_0 \cap \{Z \neq 1\}}$ are i.i.d. random variables which are stochastically dominated by random variables $Y_x \sim \mathcal{E}(\rho' C_k)$ by Equation (2.26). These exponential random variables satisfy

$$\mathbb{E}\left(Y_x - \frac{1}{\rho' C_k}\right)^2 \leq \frac{1}{(\rho' C_k)^2}$$

and for all $a \in \mathbb{N}$ such that $a \geq 3$

$$\mathbb{E}\left(Y_x - \frac{1}{\rho' C_k}\right)_+^a \leq \frac{a!}{(\rho' C_k)^a},$$

so that Bernstein's inequality (see for instance Proposition 2.9 of [Massart, 2007]) entails that for all $t > 0$

$$\mathbb{P}\left(\sum_{x \in \mathcal{A}_0} Y_x - \frac{|\mathcal{A}_0|}{\rho' C_k} \geq \frac{2\sqrt{|\mathcal{A}_0| t}}{\rho' C_k} + \frac{t}{\rho' C_k}\right) \leq e^{-t}$$

and therefore by taking $t = \log T$, with probability at least $1 - 1/T$:

$$\begin{aligned}\sum_{x \in \mathcal{A}_0 \cap \{Z \neq 1\}} T_x &\leq \frac{16\sqrt{2T}}{\rho' C_k} + \frac{2\sqrt{16\sqrt{2T}\log T}}{\rho' C_k} + \frac{\log T}{\rho' C_k} \\ &\leq \frac{32\sqrt{T}}{\rho' C_k}\end{aligned}$$

as soon as $T \geq T_0'$ for some numerical constant T_0'. Hence, with probability at least $1 - 1/T$,

$$k \sum_{i=0}^{I-1} |\mathcal{A}_i \cap \{Z \neq 1\}| \leq \frac{64\sqrt{T}}{\rho' s}$$

using $k \leq 2C_K/s$.

Proof of point 10. The same proof as the one of Equation (2.28) shows that for all $x \in \mathcal{A}_0 \cap \{Z = 1\}$, for all $i \geq 1$ and for all $t > 0$,

$$\mathbb{P}\left(\hat{p}_{x,i} < p - \frac{|p-q|}{8} - |p-q|\sqrt{\frac{t}{2ki}} - 2\sqrt{2p\frac{t}{ki}} - 2\frac{t}{ki}\right) \leq 3e^{-t}, \tag{2.27}$$

so that by union bound and the inequality $k \geq C_k$,

$$\mathbb{P}\left(\exists i \geq 1, \quad \hat{p}_{x,i} < \frac{7p+q}{8} - |p-q|\left[\sqrt{\frac{\log(2\pi^2 i^2)}{C_k i}}\left(\frac{1}{\sqrt{2}} + 2\sqrt{2}\right) + 2\frac{\log(2\pi^2 i^2)}{C_k i}\right]\right) \leq \frac{1}{4}.$$

Therefore, if C_k is larger than a numerical constant,

$$\mathbb{P}\left(\exists i \geq 1, \quad \hat{p}_{x,i} < \frac{3p+q}{4}\right) \leq \frac{1}{4},$$

which, combined with point 5 of Theorem 2.B.2, implies

$$\mathbb{P}\left(\exists i \geq 1, \quad \hat{p}_{x,i} < \hat{\tau}\right) \leq \frac{1}{4}.$$

Let $V_x = \mathbf{1}_{x\in\mathcal{A}_I}$ for all $x \in \mathcal{A}_0 \cap \{Z = 1\}$. The above inequality ensures that conditionally on $\mathcal{A}_0$, the $(V_x)_{x\in\mathcal{A}_0\cap\{Z=1\}}$ are i.i.d. Bernoulli random variable with parameter $r \geq 3/4$. Therefore, Hoeffding's inequality entails

$$\mathbb{P}\left(|\mathcal{A}_I \cap \{Z = 1\}| \leq \frac{3|\mathcal{A}_0 \cap \{Z = 1\}|}{4} - \sqrt{|\mathcal{A}_0|\frac{\log T}{2}}\right) \leq \frac{1}{T}.$$

Let us assume for now that $|\mathcal{A}_0 \cap \{Z = 1\}| \geq 2\sqrt{2T}$ with probability $1 - 1/T$. Then this ensures that for T larger than some numerical constant,

$$\mathbb{P}\left(|\mathcal{A}_I \cap \{Z = 1\}| \leq \sqrt{2T}\right) \leq \frac{1}{T} + \frac{1}{T}.$$

To conclude, note that conditionally on $\mathcal{N}$, the random variable $|\mathcal{A}_0 \cap \{Z = 1\}|$ is an hypergeometric random variable with parameters $(\lceil 8\sqrt{2T}\rceil, r', n - N)$ where

$$r' = \frac{\frac{n}{2} - |\mathcal{N} \cap \{Z = 1\}|}{n - N} \geq \frac{\frac{n}{2} - \frac{3}{4}\frac{n}{4}}{n} \geq \frac{5}{16}$$

by points 1 and 2 of Theorem 2.B.2. Therefore, Equation (2.41) implies that

$$\mathbb{P}\left(|\mathcal{A}_0 \cap \{Z = 1\}| \leq \frac{5}{16}8\sqrt{2T} - \sqrt{\frac{16\sqrt{2T}\log T}{2}}\right) \leq \frac{1}{T},$$

so that for T larger than a numerical constant

$$\mathbb{P}\left(|\mathcal{A}_0 \cap \{Z = 1\}| \leq 2\sqrt{2T}\right) \leq \frac{1}{T}.$$

2.B.4 Proof of Lemma 2.B.4

Let $x \in \mathcal{A}_0 \cap \{Z \neq 1\}$ and assume that we are in the event of probability at least $1 - 9/T$ where Theorem 2.B.2 holds. For all $i \in \mathbb{N}^*$,

$$\begin{aligned}\mathbb{P}(T_x \geq i) &= \mathbb{P}\left(\forall j \in \{1, \dots, i\}, \hat{p}_{x,j} \geq \hat{\tau}\right)\\ &\leq \mathbb{P}\left(\hat{p}_{x,i} \geq \hat{\tau}\right)\\ &\leq \mathbb{P}\left(\hat{p}_{x,i} \geq \frac{p + 3q}{4}\right)\end{aligned}$$

using point 5 of Lemma 2.B.2.

Following the same proof as in point 5 of Theorem 2.B.2, one can show that for all $x \in \mathcal{A}_0 \cap \{Z \neq 1\}$, all $i \geq 1$ and all $t > 0$,

$$\mathbb{P}\left(\hat{p}_{x,i} \geq q + \frac{|\mathcal{V}^-_{x,i}|}{|\mathcal{V}_{x,i}|}|p - q| + 2\sqrt{2p\frac{t}{ki}} + 2\frac{t}{ki}\right) \leq 2e^{-t}$$

where $\mathcal{V}^-_{x,i} := \mathcal{V}_{x,i} \cap \{Z \neq 1\}$.

For $s\sqrt{T} \geq c'_{\text{thresh}}$, with c'_{thresh} such that $\frac{\log(s\sqrt{T})}{s\sqrt{T}} \leq 1/64$, one has $\frac{1}{s} \leq N/64$. Then, points 2 and 4 of Lemma 2.B.2 imply that $|\mathcal{N} \cap \{\hat{Z} = 1\} \cap \{Z \neq 1\}| \leq N/64$ and $\hat{N}_1 := |\mathcal{N} \cap \{\hat{Z} = 1\}| \geq N/8$. Therefore, the proportion of misclassified vertices in $\mathcal{N} \cap \{\hat{Z} = 1\}$ is at most $1/8$, so that conditionally on $\hat{N}_1$ and the event of Lemma 2.B.2 $|\mathcal{V}^-_{x,i}|$ is stochastically dominated by an hypergeometric distribution with parameters $(ki, 1/8, \hat{N}_1)$. Hence, Equation (2.41) entails

$$\mathbb{P}\left(\frac{|\mathcal{V}^-_{x,i}|}{|\mathcal{V}_{x,i}|} \geq \frac{1}{8} + \sqrt{\frac{t}{2ki}}\right) \leq e^{-t},$$

so that for all $i \geq 1$ and $t > 0$,

$$\mathbb{P}\left(\hat{p}_{x,i} \geq q + \frac{|p-q|}{8} + |p-q|\sqrt{\frac{t}{2ki}} + 2\sqrt{2p\frac{t}{ki}} + 2\frac{t}{ki}\right) \leq 3e^{-t}. \tag{2.28}$$

Note that

$$\begin{aligned}\frac{p+3q}{4} - \left(q + \frac{|p-q|}{8} + |p-q|\sqrt{\frac{t}{2ki}} + 2\sqrt{2p\frac{t}{ki}} + 2\frac{t}{ki}\right) \\ \geq \frac{|p-q|}{8} - \sqrt{\frac{t}{C_k i}}\left(\frac{|p-q|\sqrt{s}}{\sqrt{2}} + 2\sqrt{2ps}\right) - 2\frac{ts}{C_k i} \\ \geq |p-q|\left(\frac{1}{8} - \sqrt{\frac{t}{C_k i}}\left(\frac{1}{\sqrt{2}} + 2\sqrt{2}\right) - 2\frac{t}{C_k i}\right).\end{aligned}$$

since $s = (p-q)^2/p \leq 1$.

Thus, there exists a numerical constant $\rho = 10^{-3}$ such that by taking $t = \rho C_k i$,

$$\frac{p+3q}{4} \geq q + \frac{|p-q|}{8} + |p-q|\sqrt{\frac{t}{2ki}} + 2\sqrt{2p\frac{t}{ki}} + 2\frac{t}{ki},$$

so that

$$\mathbb{P}\left(\hat{p}_{x,i} \geq \frac{p+3q}{4}\right) \leq 3e^{-\rho C_k i}$$

and finally by letting $\rho' = \rho/2$ and if $C_k \geq (\log 3)/\rho'$:

$$\forall i \in \mathbb{N}^* \qquad \mathbb{P}(T_x \geq i) \leq e^{-\rho' C_k i}.$$

2.C Proof of the Constrained Upper Bound

This section proves the upper bound in Theorem 2.4.1. Recall that $B = (B_T \wedge \sqrt{T})/2$ in the Constrained Algorithm page 103.

It is enough to prove the upper bound in Theorem 2.4.1 in the case where $sB \geq c_{thresh}$ for some numerical constant $c_{thresh} \geq 1$. Indeed, if $sB \leq c_{thresh}$, Equation (2.8) automatically holds with $c_2 \geq c_{thresh}$. Hereafter, it is then assumed that $sB \geq c_{thresh}$.

The first step of the Constrained Algorithm page 103 is almost identical to that of the Unconstrained Algorithm after replacing $\sqrt{T}$ by $B = (B_T \wedge \sqrt{T})/2$ in the cardinality of the kernel. The following lemma is a slight variant of Lemma 2.B.2 in this setting. The proof is omitted.

Lemma 2.C.1 *There exist numerical constants $c_{thresh} \geq e$ and $B_0 \geq 1$, such that, if $B \geq B_0$ and $sB \geq c_{thresh}$ and $T/B \leq n/136$, then with probability at least $1 - 9/(sB)$:*

1. *the number N_{init} of sampled nodes satisfies $N_{init} \leq \frac{n}{8} - 4\sum_{t=1}^{t_f-1} N^{(t)}$,*
2. *at least $N_{init}/4$ nodes of each community have been sampled, that is $|\{Z = j\} \cap \mathcal{N}_{init}| \geq N_{init}/4$ for each $j \in \{1, 2\}$,*
3. *the proportion $\varepsilon_{N_{init}}$ of misclassified nodes satisfies*

$$\varepsilon_{N_{init}} = \inf_{\pi \text{ permutation on } \{1,2\}} \frac{1}{2N_{init}} \sum_{k=1}^{2} |\{Z = k\} \Delta \{\hat{Z} = \pi(k)\}| \leq \frac{4}{512^2} \frac{1}{sB}, \tag{2.29}$$

4. *$|\hat{\tau} - \frac{p+q}{2}| \leq \frac{p-q}{4}$.*

At the end of the first step, $|\mathcal{O}_0|$ pairs have been sampled and the sampling-regret therefore does not exceed $\mathbb{E}[|\mathcal{O}_0|] = c_{\mathcal{O}_0} B/s \leq c_{\mathcal{O}_0} T/(sB)$ since, by definition of B, $T \geq B^2$.

Let us proceed with the second step. To show that the sampling regret in the second step does not exceed $O(T/(sB))$, it is sufficient to prove that there exist two numerical constants c_{proba} and c_{regret} such that for any (T, B) satisfying $sB \geq c_{thresh}$ and $T/B \leq n/136$, the number of bad pairs sampled during the second step is bounded from above by $c_{regret} T/(sB)$ with probability at least $1 - c_{proba}/(sB)$. Indeed, since the number of bad pairs sampled in the second step $N^{bad}_{step2}(\psi, T)$ cannot be larger than T, it directly follows that the sampling-regret during the second step is upper bounded by

$$\mathbb{E}\left[N^{bad}_{step2}(\psi, T)\right] \leq c_{regret} \frac{T}{sB_T} + T \frac{c_{proba}}{sB_T} \leq c' \frac{T}{sB_T}.$$

The following lemma provides such a control of the number of bad pairs accumulated in step 2, as well as an upper bound on the number of misclassified nodes. It is a counterpart to Lemma 2.B.3 of the unconstrained case.

Lemma 2.C.2 *There exist two numerical constants $B_0' \geq 1$ and $c'_{thresh} \geq e$ such that if $B \geq B_0'$, $sB \geq c'_{thresh}$ and $T/B \leq n/136$, then with probability at least $1 - 63/(sB)$, Lemma 2.C.1 holds and for all iterations of `SCREENING` in point 6 of the constrained algorithm,*

5. *it is always possible to sample $|\mathcal{A}_0|$ new vertices: $|V^{(0)}| \geq \ldots \geq |V^{(t_f-1)}| \geq \frac{7n}{8}$;*
6. *No node from $\mathcal{A}_0$ has more than $2B$ adjacent pairs sampled during the whole execution of the constrained algorithm.*
7. *the algorithm does not run out of connections with the reference kernel: $kI \leq N^{(0)} \leq \ldots \leq N^{(t_f-1)}$;*

and there exists a numerical constant C_{fail} such that for all $t \in \{1, \dots, t_f\}$, during the call `SCREENING` $\left(\mathcal{N}^{(t-1)}, N^{(t)}, B, \hat{\tau}, V^{(t-1)}\right)$,

8. *the number of bad pairs sampled during the t^{th}-call to* `SCREENING` *from nodes in $\mathcal{A}_0$ is controlled:*
   $\sum_{x \in \mathcal{A}_0} \left|\left\{y_a^x : (x, y_a^x) \text{ sampled and } Z_{y_a^x} \neq Z_x\right\}\right| \leq C_{fail} \frac{N^{(t)}}{s}$;
9. *few vertices from the wrong community remain:* $|\mathcal{N}^{(t)} \cap \{Z \neq 1\}| \leq 8N^{(t)}/(sB)$;
10. *enough vertices from community 1 remain for the construction of the kernel $\mathcal{N}^{(t)}$ of $N^{(t)}$ nodes:* $\sum_{j=1}^{m} |\mathcal{A}_I^{(j)} \cap \{Z = 1\}| \geq N^{(t)}$;

As a consequence, the total number of bad pairs sampled during the second step is upper bounded by

$$C_{fail} \sum_{t=1}^{t_f} \frac{N^{(t)}}{s} \leq 2C_{fail} \frac{N_{t_f}}{s} \leq 4C_{fail} \frac{T}{sB},$$

with probability larger than $1 - 63/(sB)$.

We refer to Section 2.C.1 for a proof of Lemma 2.C.2.

Let us now conclude the proof of the upper bound of Theorem 2.4.1. In the third step, the kernel $\mathcal{N}^{(t_f)}$ has $\lceil T/B \rceil \leq 2T/B$ nodes and a proportion of misclassified nodes smaller than $8/(sB)$ with probability larger than $1 - 63/(sB)$ by point 9 of Lemma 2.C.2. Since each node of $\mathcal{N}^{(t_f)}$ is sampled at most B times, the number of bad pairs sampled during the third step is smaller than $16T/(sB)$ with probability at least $1 - 63/(sB)$, and smaller than T otherwise.

Hence, using again that we always have $N^{bad}(\psi, T) \leq T$, the total sampling-regret $\mathbb{E}\left[N^{bad}(\psi, T)\right]$ during the whole process is $O(T/(sB))$. The proof of the upper bound of Theorem 2.4.1 is complete.

2.C.1 Proof of Lemma 2.C.2

Lemma 2.C.2 simultaneously controls all the iterations of `SCREENING`. To prove it, we use the following lemma which controls each iteration.

Lemma 2.C.3 *There exists a numerical constant $c'_{thresh} \geq e$ such that the following holds.*

Let $\mathcal{N} \subset V_{init}$, $N' \in \mathbb{N}$, $B > 0$, $\nu \in [0, 1]$ and $V \subset V_{init}$, and

$$(\mathcal{N}', V') = \texttt{SCREENING}(\mathcal{N}, N', B, \nu, V). \tag{2.30}$$

Write $N = |\mathcal{N}|$.

Assume that $sB \geq c'_{thresh}$, that $B \leq 4N' \leq 4N \log(sB)$, that the proportion of misclassified nodes $|\mathcal{N} \cap \{Z \neq 1\}|/|\mathcal{N}|$ is upper bounded by $c_{misclas}/(sB)$ for some constant $c_{misclas} \in [8/512^2, 8]$, that $\nu \in [\frac{p+3q}{4}, \frac{3p+q}{4}]$, that $|V| \geq 7n/8$ and that no node in V is adjacent to a pair sampled before this call to `SCREENING`*. Then with probability at least $1 - 6/(sN')$,*

1. *the proportion* $|\mathcal{N}' \cap \{Z \neq 1\}|/|\mathcal{N}'|$ *of misclassified nodes after* `SCREENING` *is upper bounded by* $c^{after}_{misclas}/(sB)$ *where* $c^{after}_{misclas} = c_{misclas} \vee 8$ *if* $N' \geq B\log(sB)^{3/2}$ *and* $c^{after}_{misclas} = 512 c_{misclas}$ *otherwise.*

2. *the number of sampled bad pairs is controlled: there exists a numerical constant* C_{fail} *(for instance* $C_{fail} = 26C_k + 2 = 65002$*) such that*

$$\sum_{x \in \mathcal{A}_0} \left|\left\{y_a^x : \ (x, y_a^x) \text{ was sampled and } Z_{y_a^x} \neq Z_x\right\}\right| \leq C_{fail}\frac{N'}{s}. \tag{2.31}$$

3. *no node in* $\mathcal{N}'$ *or* V *has more than* B *adjacent pairs sampled during this call to* `SCREENING`.

4. $|V'| \geq |V| - 4N'$.

5. *it is possible to construct the kernel* $\mathcal{N}'$ *with* N' *nodes after Step 3:* $\sum_{j=1}^{m} |\mathcal{A}_I^{(j)} \cap \{Z = 1\}| \geq N'$.

6. *no node in* V' *is adjacent to a pair sampled before or during this call to* `SCREENING`.

Lemma 2.C.3 is proved in Section 2.C.2.

To prove Lemma 2.C.2, we control the t_f screening calls at the second step of the constrained algorithm page 103 as follows. For the first step, denote by E_0 the event of probability $1 - 9/(sB)$ where all the points of Lemma 2.C.1 are true. For each $t \in \{1, \dots, t_f\}$, denote by E_t the event where all the points of Lemma 2.C.3 are satisfied by the output of `SCREENING` at the t^{th}-call, which is $(\mathcal{N}^{(t)}, V^{(t)}) = \texttt{SCREENING}\left(\mathcal{N}^{(t-1)}, N^{(t)}, B, \hat{\tau}, V^{(t-1)}\right)$. On the event $\bigcap_{0 \leq t \leq t_f} E_t$, all the points of Lemma 2.C.2 can be easily derived, see Section 2.C.1.1 for a detailed proof.

Therefore, Lemma 2.C.2 holds with a probability at least $\mathbb{P}\left(\bigcap_{0 \leq t \leq t_f} E_t\right)$. To prove that $\bigcap_{0 \leq t \leq t_f} E_t$ holds with high probability, we proceed by induction. First, the event E_0 holds with probability at least $1 - 9/(sB)$ by Lemma 2.C.1. Next, for any $t \in \{1, \dots, t_f\}$, we check in Section 2.C.1.2, that, on the event $E_0 \cap \ldots \cap E_{t-1}$, the assumptions of Lemma 2.C.3 holds at the t^{th}-call of the `SCREENING` routine. Hence, according to Lemma 2.C.3, conditionally on the event $E_0 \cap \ldots \cap E_{t-1}$, the event E_t holds with probability at least $1 - 6/(sN^{(t)})$. By induction, we thus have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{0 \leq t \leq t_f} E_t\right) &= \mathbb{P}\left(E_0\right)\, \mathbb{P}\left(E_1 | E_0\right) \ldots \mathbb{P}\left(E_{t_f} | E_{t_f - 1}, \ldots, E_0\right) \\ &\geq \left(1 - \frac{9}{sB}\right) \prod_{t=0}^{t_f} \left(1 - \frac{6}{sN^{(t)}}\right),\end{aligned}$$

which is larger than

$$1 - \frac{9}{sB} - \sum_{t=1}^{t_f} \frac{6}{sN^{(t)}} = 1 - \frac{9}{sB} - \frac{6}{sN^{(0)}} \sum_{t=1}^{t_f - 1} \lfloor \log(sB) \rfloor^{-t} - \frac{6}{s\lceil T/B \rceil}$$
$$\geq 1 - \frac{9}{sB} - \frac{12\log(sB)}{sB} \times \frac{\lfloor \log(sB) \rfloor^{-1}}{1 - \lfloor \log(sB) \rfloor^{-1}} - \frac{6}{s(T/B)}$$
$$\geq 1 - \frac{9}{sB} - \frac{48}{sB} - \frac{6}{sB} = 1 - \frac{63}{sB},$$

using for the last inequality that $B \leq \sqrt{T}/2$ and $sB \geq c'_{thresh}$ for some numerical constant $c'_{thresh} > 0$.

To conclude, Lemma 2.C.2 holds with probability at least $1 - 63/(sB)$, provided that the conclusions of Lemma 2.C.2 hold on the event $\bigcap_{0 \leq t \leq t_f} E_t$, and that the assumptions of Lemma 2.C.3 are satisfied at each call of `SCREENING`. These two points are proved in the next two subsections.

2.C.1.1 The Conclusions of Lemma 2.C.2 holds on $\bigcap_{0 \leq t \leq t_f} E_t$

Assume that the event $\bigcap_{0 \leq t \leq t_f} E_t$ holds, and let us show that all the points of Lemma 2.C.2 are fulfilled.

Points 7, 8 and 10. Points 8 and 10 of Lemma 2.C.2 follow directly from Point 2 and Point 5 of Lemma 2.C.3. As for Point 7, it is satisfied when

$$4C_k C_I \frac{\log(sB)}{s} \leq \frac{B}{2\log(sB)},$$

which holds as soon as $sB \geq c'_{thresh}$ for some numerical constant c'_{thresh}.

Point 9. In the initial kernel, the proportion of misclassified nodes is upper bounded by Lemma 2.C.1 as follows

$$|\mathcal{N}^{(0)} \cap \{Z \neq 1\}| / |\mathcal{N}^{(0)}| \leq 2\varepsilon_N \leq \frac{8}{512^2} \times \frac{1}{sB}.$$

For the next kernel $\mathcal{N}^{(1)}$, it implies that

$$|\mathcal{N}^{(1)} \cap \{Z \neq 1\}| / |\mathcal{N}^{(1)}| \leq 512 \frac{8}{512^2} \times \frac{1}{sB} = \frac{8}{512} \times \frac{1}{sB}$$

using $c_{misclas} = 8/512^2$ in the point 1 of Lemma 2.C.3. For the subsequent kernels, the proportion of misclassified nodes is upper bounded as above, updating the value of $c_{misclas}$ at each step. We thus have

$$|\mathcal{N}^{(2)} \cap \{Z \neq 1\}| / |\mathcal{N}^{(2)}| \leq 512 \frac{8}{512} \times \frac{1}{sB} = \frac{8}{sB},$$

and for all $t \geq 3$,

$$|\mathcal{N}^{(t)} \cap \{Z \neq 1\}| / |\mathcal{N}^{(t)}| \leq \frac{8}{sB},$$

since $N^{(t)} \geq B\log(sB)^{3/2}$ as soon as $t \geq 3$ and $sB \geq c'_{thresh}$ for some numerical constant c'_{thresh}.

Point 5. At the t^{th}-call to `SCREENING`, the output of "new" nodes $V^{(t)}$ satisfies the recursive inequality $|V^{(t)}| \geq |V^{(t-1)}| - 4N^{(t)}$ by construction of the algorithm. The sequence of inequalities telescopes, leaving

$$|V^{(t)}| \geq |V^{(0)}| - \sum_{s=1}^{t} 4N^{(s)},$$

which is larger than $7n/8$ since $|V^{(0)}| = n - N_{init}$ and $N_{init} \leq n/8 - \sum_{s=1}^{t_f-1} 4N^{(s)}$ by the point 1 of Lemma 2.C.1.

Point 6. A node can fall into four categories:
1/ it is never used;
2/ it is used in Step 1 and possibly in the first iteration of `SCREENING`. Then the number of adjacent sampled pairs is at most $N_{init} + B$ by construction of Step 1 and by point 3 of Lemma 2.C.3, which is smaller than $2B$ as soon as $sB \geq c_{thresh}$ for some numerical constant c_{thresh};
3/ it is used in (at most) two consecutive iterations of `SCREENING` (and nowhere else). Then the number of adjacent sampled pairs is at most $2B$ by Lemma 2.C.3;
4/ it is used in the last iteration of `SCREENING` and (possibly) in Step 3. Then the number of adjacent sampled pairs is at most $B + B$ by Lemma 2.C.3 and by construction of Step 3.

2.C.1.2 Check of the Assumptions of Lemma 2.C.3

Assume that the events $E_0, \dots, E_{t-1}$ hold together, and let us check Lemma 2.C.3 assumptions. First, the condition $sB \geq c'_{thresh}$ comes from Lemma 2.C.2. Then, following the proof of Point 9, we can check that $|\mathcal{N}^{(t-1)} \cap \{Z \neq 1\}|/|\mathcal{N}^{(t-1)}| \leq c_{misclas}/(sB)$ for $c_{misclas} \in [8/512^2, 8]$. For the threshold $\hat{\tau}$ taking value in $[\frac{p+3q}{4}, \frac{3p+q}{4}]$, it is stated in Lemma 2.C.1. The input of "new" nodes $V^{(t-1)}$ satisfies $|V^{(t-1)}| \geq 7n/8$, as seen above in the proof of Point 5. Finally, the inequality $B \leq 4N^{(t)} \leq 4N^{(t-1)}\log(sB)$ is satisfied by construction of the algorithm, as soon as $sB \geq c'_{thresh}$ for some numerical constant c'_{thresh}.

2.C.2 Proof of Lemma 2.C.3: Control of `SCREENING`

In this section, we work conditionally to $\mathcal{F}_{T_{\text{start}}}$ where T_{start} is the number of pairs sampled before the current call to `SCREENING`.

Let us state the two main technical results that allow to prove Lemma 2.C.3. Write $\mathcal{V}(x) := \mathcal{V}_j$ for each $j \in \{1, \dots, m\}$ and $x \in \mathcal{A}_0^{(j)}$. The first one controls the properties of the sets $(\mathcal{V}(x))_{x\in\mathcal{A}_0}$. Given a subset of nodes S, denote by misclas(S) the set of misclassified nodes in S, that is the set of all $x \in S$ such that $Z_x \neq 1$ in `SCREENING`.

Lemma 2.C.4 *The sets $(\mathcal{V}(x))_{x\in\mathcal{A}_0}$ satisfy*

1. *For all $y \in \mathcal{N}$, $|\{x \in \mathcal{A}_0 : y \in \mathcal{V}(x)\}| \leq B$,*

2. $\mathbb{P}\left(\left|\left\{x \in \mathcal{A}_0 : |\operatorname{misclas}(\mathcal{V}(x))| \geq \frac{kI}{16}\right\}\right| \geq \frac{c_{misclas}^{after}}{2}\frac{N'}{sB}\right) \leq \frac{2}{sN'}$

where $c_{misclas}^{after}$ *is defined as in Lemma 2.C.3,*

3. $\sum_{x\in\mathcal{A}_0} |\operatorname{misclas}(\mathcal{V}(x))| \leq \frac{N'}{s}$.

The proof of the above lemma is postponed to Section 2.C.3. The next lemma allows to control the effectiveness of Step 3 of `SCREENING`. Its proof follows the same lines as the proof of Lemma 2.B.4 (for proving (2.32)) and Point 10 of Lemma 2.B.3 (for proving (2.33)), it is therefore omitted.

Lemma 2.C.5 *Conditionally to the choice of the set* $\mathcal{A}_0$ *and* $(\mathcal{V}(x))_{x\in\mathcal{A}_0}$*, the variables* $(T_x)_{x\in\mathcal{A}_0}$ *are independent and for all* $x \in \mathcal{A}_0$ *and all* $i \in \{1,\dots,I\}$,

$$\mathbb{P}\left(T_x \geq i \,\middle|\, Z_x \neq 1 \text{ and } |\operatorname{misclas}(\mathcal{V}(x))| \leq \frac{kI}{16}\right) \leq e^{-i} \tag{2.32}$$

$$\mathbb{P}\left(T_x \geq I \,\middle|\, Z_x = 1 \text{ and } |\operatorname{misclas}(\mathcal{V}(x))| \leq \frac{kI}{16}\right) \geq \frac{3}{4}. \tag{2.33}$$

Let us now prove Lemma 2.C.3. Note that Points 4 and 6 follow from the construction of the algorithm and that Point 3 follows straighforwardly from point 1 of Lemma 2.C.4 (for the nodes from $\mathcal{N}$) and from the construction of the algorithm (for the nodes from $\mathcal{A}_0$).

2.C.2.1 Proof of Point 1

$$\begin{aligned}
|\mathcal{N}' \cap \{Z \neq 1\}| &\leq \sum_{j=1}^{m} |\operatorname{misclas}(\mathcal{A}_I^{(j)})| \\
&= \sum_{j=1}^{m} \sum_{\substack{x\in\mathcal{A}_0 \text{ s.t.} \\ |\operatorname{misclas}(\mathcal{V}(x))|>\frac{kI}{16}}} \mathbf{1}_{x\in\mathcal{A}_I^{(j)} \text{ and } Z_x\neq 1} + \sum_{j=1}^{m} \sum_{\substack{x\in\mathcal{A}_0 \text{ s.t.} \\ |\operatorname{misclas}(\mathcal{V}(x))|\leq\frac{kI}{16}}} \mathbf{1}_{x\in\mathcal{A}_I^{(j)} \text{ and } Z_x\neq 1} \\
&\leq \sum_{x\in\mathcal{A}_0} \mathbf{1}_{|\operatorname{misclas}(\mathcal{V}(x))|\geq\frac{kI}{16}} + \sum_{\substack{x\in\mathcal{A}_0 \text{ s.t.} \\ |\operatorname{misclas}(\mathcal{V}(x))|\leq\frac{kI}{16}}} \mathbf{1}_{T_x\geq I \text{ and } Z_x\neq 1} \\
&\leq \frac{c_{\text{misclas}}^{\text{after}}}{2}\frac{N'}{sB} + \sum_{\substack{x\in\mathcal{A}_0 \text{ s.t.} \\ |\operatorname{misclas}(\mathcal{V}(x))|\leq\frac{kI}{16}}} \mathbf{1}_{T_x\geq I \text{ and } Z_x\neq 1}
\end{aligned}$$

with probability at least $1 - 2/(sN')$ by point 2 of Lemma 2.C.4.

The second term is dominated by a binomial random variable with parameters $(|\mathcal{A}_0|, e^{-I})$ by Lemma 2.C.5, so it is dominated by a binomial random variable X with parameters

$(4N', 1/(sB)^{1026})$ since $I \geq 1026 \log(sB)$. Equation (2.45) implies that for $sB \geq 512$ (which is implied by $c'_{thresh} \geq 512$),

$$\mathbb{P}\left(\frac{1}{4N'}X \geq \frac{1}{512sB}\right) \leq \exp\left(-\frac{1}{2}4N'\frac{1}{512sB}\log\left(\frac{(sB)^{1025}}{512}\right)\right) \leq \exp\left(-4N'\frac{\log(sB)}{sB}\right).$$

We want this probability to be smaller than $1/(sN')$, that is

$$\frac{\log(sB)}{sB} \geq s\frac{\log(sN')}{4sN'},$$

which is true since $s \leq 1$, and the function $x \longmapsto \frac{\log x}{x}$ is nonincreasing for $x \geq e$, and $4sN' \geq sB \geq e$ by assumption.

Therefore,

$$\mathbb{P}\left(|\mathcal{N}' \cap \{Z \neq 1\}| \geq \frac{c^{\text{after}}_{\text{misclas}} + (8/512)}{2}\frac{N'}{sB}\right) \leq \frac{3}{sN'}$$

which implies Point 1 (since $c^{\text{after}}_{\text{misclas}} \geq 8/512$ by definition).

2.C.2.2 Proof of Point 2

Given a subset S of $\mathcal{A}_0$, denote by $\text{bad}(S)$ the number of sampled bad pairs coming from nodes in S during `SCREENING`, that is

$$\text{bad}(S) = \sum_{x \in S} |\{y_i^x : i \leq k((T_x + 1) \wedge I) \text{ and } Z_{y_i^x} \neq Z_x\}|.$$

The total number of bad pairs sampled during `SCREENING` can be decomposed into

$$\begin{aligned}
\text{bad}\left(\mathcal{A}_0\right) = &\sum_{x \in \mathcal{A}_0} \text{bad}(\{x\})\mathbf{1}_{|\,\text{misclas}(\mathcal{V}(x))| > \frac{kI}{16}} \\
&+ \sum_{x \in \mathcal{A}_0 \cap \{Z=1\}} \text{bad}(\{x\})\mathbf{1}_{|\,\text{misclas}(\mathcal{V}(x))| \leq \frac{kI}{16}} \\
&+ \sum_{x \in \mathcal{A}_0 \cap \{Z\neq 1\}} \text{bad}(\{x\})\mathbf{1}_{|\,\text{misclas}(\mathcal{V}(x))| \leq \frac{kI}{16}} \\
&\leq kI \sum_{x \in \mathcal{A}_0} \mathbf{1}_{|\,\text{misclas}(\mathcal{V}(x))| \geq \frac{kI}{16}} \qquad (2.34) \\
&+ \sum_{x \in \mathcal{A}_0 \cap \{Z=1\}} |\,\text{misclas}(\mathcal{V}(x))| \qquad (2.35) \\
&+ \sum_{x \in \mathcal{A}_0 \cap \{Z\neq 1\}} k(T_x + 1)\mathbf{1}_{|\,\text{misclas}(\mathcal{V}(x))| \leq \frac{kI}{16}} \qquad (2.36)
\end{aligned}$$

The first sum is controlled by Point 2 of Lemma 2.C.4:

$$\mathbb{P}\left((2.34) \geq \frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{N'kI}{sB}\right) \leq \frac{2}{sN'}.$$

Thus, since $kI/B \leq 4C_kC_I \log(sB)/(sB)$ by definition, there exists a constant c'_{thresh} such that $kI/B \leq 2/c^{\text{after}}_{\text{misclas}}$ as soon as $sB \geq c'_{thresh}$, so that

$$\mathbb{P}\left((2.34) \geq \frac{N'}{s}\right) \leq \frac{2}{sN'}.$$

Likewise, by Point 3 of Lemma 2.C.4,

$$(2.35) \leq \frac{N'}{s}.$$

By Lemma 2.C.5, the variables T_x in the third sum are stochastically dominated by i.i.d. exponential random variables with parameter 1. Therefore, using the inequality $k \leq 2C_k/s$, the term (2.36) is stochastically dominated by

$$8C_k\frac{N'}{s} + 2\frac{C_k}{s}\sum_{i=1}^{4N'} Y_i$$

where $(Y_i)_{i\in\mathbb{N}^*}$ are i.i.d. exponential random variables with parameter 1. These exponential random variables satisfy

$$\mathbb{E}\left(Y_i - 1\right)^2 \leq 1$$

and for all $a \in \mathbb{N}$ such that $a \geq 3$

$$\mathbb{E}\left(Y_i - 1\right)_+^a \leq a!,$$

so that Bernstein's inequality (see for instance Proposition 2.9 of [Massart, 2007]) entails for all $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^{4N'} Y_i - 4N' \geq 4\sqrt{N't} + t\right) \leq e^{-t}$$

and therefore by taking $t = N'$, with probability at least $1 - e^{-N'} \geq 1 - 1/N' \geq 1 - 1/(sN')$

$$\sum_{i=1}^{4N'} Y_i \leq 9N'.$$

Hence, with probability at least $1 - 3/(sN')$,

$$\text{bad}\left(\mathcal{A}_0\right) \leq (26C_k + 2)\frac{N'}{s}.$$

2.C.2.3 Proof of Point 5.

Write $\mathcal{A}_I = \bigcup_{j=1}^m \mathcal{A}_I^{(j)}$ and for each $x \in \mathcal{A}_0$, let $V_x = \mathbf{1}_{x\in\mathcal{A}_I}$ indicate whether x has been kept until the end of Step 3 of `SCREENING`. Lemma 2.C.5 ensures that the random variables $(V_x)_{x\in\mathcal{A}_0\cap\{Z=1\} \text{ s.t. } |\text{misclas}(\mathcal{V}(x))|\leq\frac{kI}{16}}$ dominate i.i.d. Bernoulli random variables with parameter 3/4. Therefore, Hoeffding's inequality (2.41) entails

$$\mathbb{P}\left(\left|\mathcal{A}_I \cap \{Z=1\}\right| \leq \frac{3\left|\mathcal{A}_0 \cap \{Z=1\} \cap \{x : |\,\text{misclas}(\mathcal{V}(x))| \leq \frac{kI}{16}\}\right|}{4} - \sqrt{|\mathcal{A}_0|\frac{\log(sN')}{2}}\right) \leq \frac{1}{sN'}.$$

Note that $|\{x \in \mathcal{A}_0 \text{ s.t. } |\operatorname{misclas}(\mathcal{V}(x))| > \frac{kI}{16}\}| \leq \frac{c^{\text{after}}_{\text{misclas}}}{2} N'/(sB)$ with probability at least $1 - 2/(sN')$ by Lemma 2.C.4. Since $|\mathcal{A}_0| = 4N'$, the previous equation entails

$$\mathbb{P}\left(|\mathcal{A}_I \cap \{Z = 1\}| \leq \frac{3|\mathcal{A}_0 \cap \{Z = 1\}|}{4} - \frac{3\, c^{\text{after}}_{\text{misclas}} N'}{8sB} - \sqrt{4N' \frac{\log(sN')}{2}}\right) \leq \frac{3}{sN'}.$$

Let us assume for now that $|\mathcal{A}_0 \cap \{Z = 1\}| \geq \frac{11}{7} N'$ with probability at least $1 - 1/(sN')$. Then this ensures that for N' and sB larger than some numerical constants (which is guaranteed by $B \geq B_0$ and $sB \geq c'_{thresh}$),

$$\mathbb{P}\left(|\mathcal{A}_I \cap \{Z = 1\}| \leq N'\right) \leq \frac{4}{sN'},$$

which gives point 5, provided that $|\mathcal{A}_0 \cap \{Z = 1\}| \geq \frac{11}{7} N'$.

The random variable $|\mathcal{A}_0 \cap \{Z = 1\}|$ is an hypergeometric random variable with number of draws $4N'$ and initial probability of a winning draw $r' \in [\frac{3}{7}, \frac{4}{7}]$ because the number of nodes that have not been sampled at the start of `SCREENING` is bigger than $7n/8$ by assumption and because the true communities are balanced.

Therefore, Hoeffding's inequality (2.41) implies

$$\mathbb{P}\left(|\mathcal{A}_0 \cap \{Z = 1\}| \leq \frac{3}{7} 4N' - \sqrt{\frac{4N' \log(sN')}{2}}\right) \leq \frac{1}{sN'},$$

so that for N' large enough (which is implied by $B \geq B_0$ for some numerical constant B_0).

$$\mathbb{P}\left(|\mathcal{A}_0 \cap \{Z = 1\}| \leq \frac{11}{7} N'\right) \leq \frac{1}{sN'}.$$

2.C.3 Proof of Lemma 2.C.4

2.C.3.1 Points 1 and 3

To check Point 1, it suffices to check that $\lceil 4N'/m \rceil \leq B$. For $sB \geq c'_{thresh}$ with a numerical constant c'_{thresh} large enough, one has $m = \lfloor N/(kI) \rfloor \geq N/(2kI)$ and

$$\left\lceil \frac{4N'}{m} \right\rceil \leq \frac{16N'kI}{N} \leq 64 C_k C_I \frac{(\log(sB))^2}{s} \leq B. \tag{2.37}$$

For Point 3, note that

$$\begin{aligned}\sum_{x\in\mathcal{A}_0} |\operatorname{misclas}(\mathcal{V}(x))| &\leq \left\lceil \frac{4N'}{m} \right\rceil \sum_{j=1}^{m} |\operatorname{misclas}(\mathcal{V}_j)| \\ &\leq 16kI\frac{N'}{N}|\operatorname{misclas}(\mathcal{N})| \\ &\leq 64C_kC_I\frac{\log(sB)}{s}c_{\text{misclas}}\frac{N'}{sB} \\ &\leq 64C_kC_I\frac{\log(sB)}{sB}c^{\text{after}}_{\text{misclas}}\frac{N'}{s} \\ &\leq \frac{N'}{s}\end{aligned}$$

by assumption on the the number of misclassified nodes in $\mathcal{N}$, and as soon as $sB \geq c'_{thresh}$ for some numerical constant c'_{thresh}.

2.C.3.2 Point 2, Small Kernels

In this section, we assume $N' < B\log(sB)^{3/2}$. By Equation (2.37).

$$\begin{aligned}\left|\left\{x\in\mathcal{A}_0 : |\operatorname{misclas}(\mathcal{V}(x))| \geq \frac{kI}{16}\right\}\right| &\leq \left\lceil \frac{4N'}{m} \right\rceil \sum_{j=1}^{m} \mathbf{1}_{|\operatorname{misclas}(\mathcal{V}_j)|\geq\frac{kI}{16}} \\ &\leq 16\frac{N'}{N}kI\sum_{j=1}^{m}\mathbf{1}_{|\operatorname{misclas}(\mathcal{V}_j)|\geq\frac{kI}{16}}.\end{aligned}$$

Note that

$$\begin{aligned}\frac{kI}{16}\sum_{j=1}^{m}\mathbf{1}_{|\operatorname{misclas}(\mathcal{V}_j)|\geq\frac{kI}{16}} &\leq \sum_{j=1}^{m}|\operatorname{misclas}(\mathcal{V}_j)| \\ &= |\operatorname{misclas}(\mathcal{N})| \leq \frac{c_{\text{misclas}}N}{sB}\end{aligned}$$

by assumption, so that

$$\begin{aligned}\left|\left\{x\in\mathcal{A}_0 : |\operatorname{misclas}(\mathcal{V}(x))| \geq \frac{kI}{16}\right\}\right| &\leq 16\frac{N'}{N}kI\times\frac{16}{kI}\frac{c_{\text{misclas}}N}{sB} \\ &= 256c_{\text{misclas}}\frac{N'}{sB} \\ &= \frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{N'}{sB}.\end{aligned}$$

This bound is not random, it holds with probability 1.

2.C.3.3 Point 2, Large Kernels

In this section, we assume $N' \geq B\log(sB)^{3/2}$.

The number of misclassified nodes in each $\mathcal{V}_j$ can be controlled more easily by introducing a coupling with i.i.d. Bernoulli random variables. Note that this coupling is a theoretical tool and does not appear in the algorithm.

Lemma 2.C.6 *Let K be a random variable taking values in $\{0,\dots,N\}$. Let $(X_x)_{x\in\mathcal{N}}$ be a vector of random variables taking values in $\{0,1\}$ such that*

- $\sum_{x\in\mathcal{N}} X_x = K$
- *the distribution of $(X_x)_{x\in\mathcal{N}}$ is invariant under permutation of $\mathcal{N}$*

Note that these two points together with the distribution of K characterize the distribution of $(X_x)_{x\in\mathcal{N}}$. Then for all $u>0$, there exists a coupling with i.i.d. Bernoulli random variables $(Y_x)_{x\in\mathcal{N}}$ with parameter u such that by writing $M=\sum_{x\in\mathcal{N}} Y_x$, M is independent of $(X_x)_{x\in\mathcal{N}}$ and

$$M \geq K \implies (\forall x\in\mathcal{N},\ X_x \leq Y_x)\,. \tag{2.38}$$

Proof of Lemma 2.C.6. Let M be a binomial random variable with parameters (N,u) such that M and K are independent. Let $(\widetilde{X}_i)_{1\leq i\leq N}$ and $(\widetilde{Y}_i)_{1\leq i\leq N}$ be random variables such that conditionally to M and K and for all $1\leq i\leq N$,

$$\widetilde{X}_i = \begin{cases} 1 & \text{if } i\leq K \\ 0 & \text{otherwise} \end{cases}$$

$$\widetilde{Y}_i = \begin{cases} 1 & \text{if } i\leq M \\ 0 & \text{otherwise} \end{cases}.$$

Let σ be a uniform random variable in the set of bijections from $\{1,\dots,N\}$ to $\mathcal{N}$ that is independent of K, M, $(\widetilde{X}_i)_i$ and $(\widetilde{Y}_i)_i$, and define $X'_x=\widetilde{X}_{\sigma^{-1}(x)}$ and $Y_x=\widetilde{Y}_{\sigma^{-1}(x)}$ for all $x\in\mathcal{N}$.

Then the random vector $(X'_x)_{x\in\mathcal{N}}$ has the same distribution as the random vector $(X_x)_{x\in\mathcal{N}}$, the random variables $(Y_x)_{x\in\mathcal{N}}$ are i.i.d. Bernoulli random variables with parameter u, and Equation (2.38) holds for these two vectors. □

Let M and $(Y_x)_{x\in\mathcal{N}}$ be the random variables given by Lemma 2.C.6 applied to $(X_x)_{x\in\mathcal{N}} = (\mathbf{1}_{\widehat{Z}_x\neq Z_x})_{x\in\mathcal{N}}$, $K=|\operatorname{misclas}(\mathcal{N})|$ and $u = 2c^{\text{after}}_{\text{misclas}}/(sB) + 4\log(sB)^2/B$. Note that the algorithm is invariant by permutation of the nodes of $\mathcal{N}$, so that we may assume without loss of generality that the distribution of these $(X_x)_{x\in\mathcal{N}}$ is invariant by permutation of $\mathcal{N}$.

By Assumption of Lemma 2.C.3, we have $K\leq c_{\text{misclas}}/(sB)$. Let us show that $M\geq c_{\text{misclas}}/(sB)$ with probability at least $1-1/(sN')$, which implies $M\geq K$ with probability at least $1-1/(sN')$. Since M is a binomial random variable with parameters (N,u), Bernstein's inequality (2.42) entails

$$\mathbb{P}\left(M \leq Nu - \sqrt{2Nut} - t\right) \leq e^{-t}.$$

Since $\sqrt{2ab} \leq \frac{a}{2} + b$ for all $a, b > 0$, it holds with probability at least $1 - 1/(sN')$

$$\begin{aligned} M &\geq Nu - \frac{Nu}{2} - \log(sN') - \log(sN') \\ &\geq \frac{c_{\text{misclas}} N}{sB} + 2N\frac{\log(sB)^2}{B} - 2\log(sN'). \end{aligned}$$

Note that

$$\frac{2\log(sN')}{2N\frac{\log(sB)^2}{B}} \leq \frac{\frac{\log(sN\log(sB))}{N}}{\frac{\log(sB)^2}{B}} = \frac{\frac{\log(sN\log(sB))}{sN\log(sB)}}{\frac{\log(sB)}{sB}} \leq 1,$$

as soon as $sB \geq e$ since the application $x \longmapsto (\log x)/x$ is nonincreasing for $x \geq e$ and $sN\log(sB) \geq sB \geq e$ (the second last inequality comes from the assumption $N' = N\lfloor \log(sB) \rfloor \geq B\log(sB)^{3/2}$ made at the beginning of the current subsection). Therefore,

$$\mathbb{P}\left(M \leq \frac{c_{\text{misclas}} N}{sB} \right) \leq \frac{1}{sN'},$$

and finally, according to Lemma 2.C.6,

$$\mathbb{P}\left(\forall x \in \mathcal{N}, \quad \mathbf{1}_{\widehat{Z}_x \neq Z_x} \leq Y_x \right) \geq 1 - \frac{1}{sN'}. \tag{2.39}$$

We can now proceed to the conclusion of the proof of Point 2 when $N' \geq B\log(sB)^{3/2}$. We have

$$\begin{aligned} \left| \left\{ x \in \mathcal{A}_0 : |\operatorname{misclas}(\mathcal{V}(x))| \geq \frac{kI}{16} \right\} \right| &\leq \left\lceil \frac{4N'}{m} \right\rceil \sum_{j=1}^{m} \mathbf{1}_{|\operatorname{misclas}(\mathcal{V}_j)| \geq \frac{kI}{16}} \\ &\leq 16\frac{N'}{N} kI \sum_{j=1}^{m} \mathbf{1}_{\sum_{x\in\mathcal{V}_j} Y_x \geq \frac{kI}{16}} \end{aligned}$$

with probability at least $1 - 1/(sN')$ by Equations (2.37) and (2.39).

Note that $\sum_{j=1}^{m} \mathbf{1}_{\sum_{x\in\mathcal{V}_j} Y_x \geq \frac{kI}{16}}$ is a binomial random variable with parameters $(m, \mathbb{P}(\sum_{x\in\mathcal{V}_j} Y_x \geq \frac{kI}{16}))$, and that $\sum_{x\in\mathcal{V}_j} Y_x$ is a binomial random variable with parameters (kI, u) with $u = 2c^{\text{after}}_{\text{misclas}}/(sB) + 4\log(sB)^2/B$. Since $5u \leq 1/16$ for $sB \geq c'_{thresh}$, we can apply Equation (2.45) to obtain

$$\mathbb{P}\left(\sum_{x\in\mathcal{V}_j} Y_x \geq \frac{kI}{16} \right) \leq \exp\left(-\frac{kI}{32} \log\frac{1}{16u} \right). \tag{2.40}$$

Note that

$$\begin{aligned} \log\frac{1}{16u} &\geq \log\frac{sB}{256(1 \vee (s\log(sB)^2))} \\ &= \log(sB) - \log 256 - 0 \vee \log(s\log(sB)^2) \\ &\geq \frac{2}{3}\log(sB) - 0 \vee \log((sB)^{1/3}) \\ &\geq \frac{2}{3}\log(sB) - \log(sB)/3 = \frac{1}{3}\log(sB), \end{aligned}$$

when $sB \geq c'_{thresh}$ for c'_{thresh} large enough. Therefore, Equation (2.40) implies

$$\mathbb{P}\left(\sum_{x\in\mathcal{V}_j} Y_x \geq \frac{kI}{16}\right) \leq \exp\left(-\frac{kI}{96}\log(sB)\right).$$

It remains to control the probability that a binomial random variable with parameters $\left(m, \exp\left(-\frac{kI}{96}\log(sB)\right)\right)$ exceeds $\frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{N'/(sB)}{16kIN'/N}$. To apply Equation (2.40), check that

$$\begin{aligned}
\frac{\frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{N'/(sB)}{16kIN'/N}}{m\exp\left(-\frac{kI}{96}\log(sB)\right)} &= \frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{\frac{N'}{m}}{16kI\frac{N'}{N}}\frac{1}{sB}\exp\left(\frac{kI}{96}\log(sB)\right) \\
&\geq \frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{1}{16sB}\exp\left(\frac{kI}{96}\log(sB)\right) \quad \text{since } m \leq \frac{N}{kI} \\
&\geq \exp\left(\frac{kI}{200}\log(sB)\right) \geq 5,
\end{aligned}$$

for $sB \geq c'_{thresh}$. Thus, Equation (2.45) and $c^{\text{after}}_{\text{misclas}} = c_{\text{misclas}} \vee 8$ imply

$$\begin{aligned}
\mathbb{P}\left(16\frac{N'}{N}kI\sum_{j=1}^{m}\mathbf{1}_{\sum_{x\in\mathcal{V}_j}Y_x\geq\frac{kI}{16}} \geq \frac{c_{\text{misclas}}\vee 8}{2}\frac{N'}{sB}\right) &\leq \exp\left(-\frac{c_{\text{misclas}}\vee 8}{4}\frac{N'/(sB)}{16kIN'/N}\frac{kI}{200}\log(sB)\right) \\
&\leq \exp\left(-\frac{N\log(sB)}{1600sB}\right) \\
&\leq \exp\left(-\frac{N'}{1600sB}\right) \quad \text{since } N\log(sB) \geq N'.
\end{aligned}$$

We want this probability to be smaller than $1/(sN')$, that is

$$\frac{N'}{1600sB} \geq \log(sN')$$

which holds as soon as $N' \geq \left\lceil 2 \times 1600sB\log(1600s^2B)\right\rceil$, which is implied by the assumption $N' \geq B\log(sB)^{3/2}$ for $sB \geq c'_{\text{thresh}}$. Thus,

$$\mathbb{P}\left(\left|\left\{x \in \mathcal{A}_0 : |\,\text{misclas}(\mathcal{V}(x))| \geq \frac{kI}{16}\right\}\right| \geq \frac{c^{\text{after}}_{\text{misclas}}}{2}\frac{N'}{sB}\right) \leq \frac{2}{sN'}.$$

The proof is complete.

2.D Probabilistic Inequalities

We recall Bernstein and Hoeffding inequalities for binomial and hypergeometric distributions.

Lemma 2.D.1 *For $n \geq 1$, $p \in [0,1]$ and $N \geq n$, let X be either a binomial random variable with parameters (n,p) or a sum of m i.i.d. hypergeometric random variables with parameters $(\frac{n}{m}, p, N)$. Then, for all $t > 0$,*

$$\mathbb{P}\left(X - np \geq \sqrt{\frac{nt}{2}}\right) \leq e^{-t} \qquad \text{and} \qquad \mathbb{P}\left(|X - np| \geq \sqrt{\frac{nt}{2}}\right) \leq 2e^{-t} \tag{2.41}$$

and

$$\mathbb{P}\left(X - np \geq \sqrt{2npt} + t\right) \leq e^{-t}, \tag{2.42}$$

$$\mathbb{P}\left(|X - np| \geq \sqrt{2npt} + t\right) \leq 2e^{-t}. \tag{2.43}$$

The following lemma allows to control large deviations of binomial and hypergeometric random variables.

Lemma 2.D.2 *Let X be either a binomial random variable with parameters (n,p) or a sum of m i.i.d. hypergeometric random variables with parameters $(\frac{n}{m}, p, N)$. Then for all $c \in [p,1]$,*

$$\mathbb{P}\left(X \geq nc\right) \leq e^{-n \cdot kl(c,p)} \tag{2.44}$$

where $kl(c,p) = c\log(c/p) + (1-c)\log((1-c)/(1-p))$.

In particular, if $c \geq 5p$,

$$\mathbb{P}\left(X \geq nc\right) \leq e^{-\frac{1}{2}nc\log\frac{c}{p}}. \tag{2.45}$$

Proof of Lemma 2.D.2. The large deviation Inequality (2.44) is derived by the classical Cramèr-Chernoff's method (see for instance [Massart, 2007], Chapter 2).

For Inequality (2.45), note that for all $0 < \alpha < 1/p$,

$$\begin{aligned}
kl(\alpha p, p) &= \frac{\alpha p}{2}\log\alpha + \left[\frac{\alpha p}{2}\log\alpha - (1-\alpha p)\log\frac{1-p}{1-\alpha p}\right] \\
&= \frac{\alpha p}{2}\log\alpha + \left[\frac{\alpha p}{2}\log\alpha - (1-\alpha p)\log\left(1 + p\frac{\alpha - 1}{1-\alpha p}\right)\right] \\
&\geq \frac{\alpha p}{2}\log\alpha + \left[\frac{\alpha p}{2}\log\alpha - p(\alpha - 1)\right] \\
&\geq \frac{\alpha p}{2}\log\alpha + p\left[\frac{\alpha\log\alpha}{2} + 1 - \alpha\right],
\end{aligned}$$

and the term inside the square brackets is positive as soon as $\alpha \geq 5$. □

We also recall some classical controls on the Kullback-Leibler divergence between two Bernoulli distribution.

Lemma 2.D.3 *For any $p_1, p_2 \in [0,1]$,*

$$\frac{(p_1 - p_2)^2}{p_1 \vee p_2} \leq kl(p_1, p_2) \leq \frac{(p_1 - p_2)^2}{p_1(1-p_1) \wedge p_2(1-p_2)}.$$

In particular, for any $q \leq p \leq 1/2$,

$$s = \frac{(p-q)^2}{p+q} \leq \frac{(p-q)^2}{p} \leq kl(p,q) \vee kl(q,p) \leq \frac{2(p-q)^2}{q} = 2(1 + p/q)s.$$

Miscellaneous inequalities. The following inequality is used repeatedly in the proofs.

Lemma 2.D.4 *For all $x > 0$ and $y \geq 0$,*

$$x \geq (2y \log y) \vee e \implies \frac{x}{\log x} \geq y. \tag{2.46}$$

Chapter 3

Lower Bounds and Conjectures in Pair-Matching

In this chapter, the pair-matching problem is investigated in more complex models: in a stochastic block model with K communities and a random geometric graph model. Some constraints are also imposed on the explorative behaviour of strategies, enforcing the exploration of the latent space (balanced sampling constraint). Some lower-bounds are derived by adapting the proofs of the lower-bounds of the previous chapter, and some detailed arguments are presented to explain how it should be possible to derive matching upper-bounds. A special attention is paid on the impact of the balanced sampling constraint on the optimal regret. In the considered stochastic block model, the balanced sampling constraint induces an additional factor $\sqrt{K}$ in the regret of the optimal strategy.

Contents

3.1 Introduction

In the previous chapter, the pair-matching problem has been investigated in the simple case where the graph has the structure of a two communities SBM. However, analysis and visualization of real-life data often require more complex models. This motivates us to study the pair-matching problem in more general settings such as K-classes SBM and its continuous limit, the so-called graphon model. Specifically, we present conjectures on the optimal regrets in such settings, including proofs of matching lower bounds. A special attention is paid to strategies constrained to explore many communities in SBM or a large portion of the latent space in graphons. While it is a desirable feature in many applications (such as matching players in sport tournaments or reconstructing protein-protein interaction networks), it is not satisfied by optimal strategies in [Giraud et al., 2019].

We will discuss the conjecture of the previous chapter and give a partial proof for it, before looking at some generalizations, with constrained strategies and geometric graphs. In particular, we highlight the fact that the price to pay for exploring equally a constant fraction of the communities in SBM is an additional factor $\sqrt{K}$ in the regret.

The chapter is organized as follows. In Section 2, we formulate the problem in SBM and geometric graphs. Section 3 collects our conjectures on the regret, together with partial results on the lower bounds. Appendices are devoted to the proofs.

NOTATIONS. c and c' denote absolute positive constants, whose values may change along the chapter. One write $x \gtrsim y$ (respectively $x \lesssim y$) if there exists a constant c such that $x \geq cy$ (resp. $x \leq cy$). One denote by $x \asymp y$ if $x \lesssim y$ and $x \gtrsim y$.

3.2 Setup

In this section, we first recall the framework of the previous chapter for a general model of random graph, before specifying the set-up for the case of community structure and geometric structure. Then, we present the situation where we impose a constraint on the explorative behaviour of strategies.

3.2.1 The pair-matching problem

A random graph model. The n vertices are indexed by the set $V = \{1, \dots, n\}$. Successful matches are represented by a set of edges E between nodes in V: there is a successful match between a and b in V if and only if the pair $\{a, b\}$ belongs to E. Hereafter, a set of two distinct elements in V is called a *pair*, and $\mathcal{E}$ denote the set of all pairs of nodes. The graph $(V = \{1, \dots, n\}, E)$ is conveniently represented by its adjacency matrix $A \in \mathbb{R}^{n \times n}$, whose entries satisfy $A_{ab} = 1$ if $\{a, b\} \in E$ and $A_{ab} = 0$ otherwise. For any pair $e = \{a, b\}$, the notations A_e and A_{ab} are used indifferently. The graph considered here is undirected and without loop, thus A is symmetric with diagonal entries equal to zero.

The graph (V, E) is modeled as a stochastic process. Let $\{P_{ij}\}_{1 \leq j < i \leq n}$ be parameters in $[0, 1]$. Then, the entries of A below the diagonal are independently sampled as Bernoulli

random variables with parameters $\{P_{ij}\}$.

$$A_{ij} \sim \mathcal{B}(P_{ij}), \text{ for } 1 \leq i < j \leq n.$$

A strategy. The graph (V, E) is unobserved at first. The decision maker (or pair-matcher) uncovers it step by step by adaptive queries. At step $t = 1$, it selects a pair $\widehat{e}_1 \in \mathcal{E}$ and observes the interaction $A_{\widehat{e}_1}$. At next step $t > 1$, it picks a new pair $\widehat{e}_t \in \mathcal{E} \setminus \{\widehat{e}_1, \ldots, \widehat{e}_{t-1}\}$, using the information available at time t, that is $(\widehat{e}_1, A_{\widehat{e}_1}, \ldots, \widehat{e}_{t-1}, A_{\widehat{e}_{t-1}})$, and collects the new observation $A_{\widehat{e}_t}$. The decision maker may randomize its choice, but cannot pick a pair twice, or use any information other than the past observations.

Formally, let $U_1, U_2, \ldots$ be i.i.d uniform random variables in $[0, 1]$, independent of A and representing the sequence of internal randomness for the algorithm. A strategy of the pair-matcher is a sequence $\psi = (\psi_t;\, 1 \leq t \leq \binom{n}{2})$ of measurable functions ψ_t which maps the space of past observations $(\mathcal{E} \times \{0, 1\})^{t-1}$ and internal randomness $[0, 1]^t$ to the set of available pairs at time t, that is, $\mathcal{E} \setminus \widehat{\mathcal{E}}_{t-1}$ with $\widehat{\mathcal{E}}_{t-1} = \{\widehat{e}_1, \ldots, \widehat{e}_{t-1}\}$. Thus, the function ψ_t takes as input the observed graph $(A_e)_{e \in \widehat{\mathcal{E}}_{t-1}}$, and possibly an internal "new" randomness U_t, to output a new pair $\widehat{e}_t$:

$$\widehat{e}_t = \psi_t(\widehat{\mathcal{E}}_{t-1}, (A_e)_{e \in \widehat{\mathcal{E}}_{t-1}}, U_1, \ldots, U_t),$$

except for the first pair that is sampled as $\widehat{e}_1 = \psi_1(U_1)$.

The objective. The objective of the decision maker is to discover as many edges as possible in average. In other words, its strategy ψ should maximize the expected number of discoveries:

$$\mathbb{E}_\mu \left[\sum_{t=1}^{T} A_{\widehat{e}_t} \right] = \sum_{t=1}^{T} \mathbb{E}_\mu \left[P_{\widehat{e}_t} \right].$$

It is equivalent to minimize the regret which is the difference between the best possible performance and the performance of ψ. Formally, the regret is defined as

$$\mathbb{E}_\mu \left[R_T(\psi) \right] = \sum_{t=1}^{T} P_{e_t^*} - \sum_{t=1}^{T} \mathbb{E}_\mu \left[P_{\widehat{e}_t} \right]$$

where $P_{e_1^*} \geq \ldots \geq P_{e_T^*}$ are the T largest probabilities of connection in the graph (i.e. largest entries of $\{P_e\}_{e \in \mathcal{E}}$).

Without assumptions on the graph structure, past observations are useless for making decisions, implying that linear regret is inevitable. In order to study interesting strategies, we thus consider the following structures.

3.2.2 The case of community structure

Let $K \geq 2$ be an integer and a divisor of n, and $0 < q < p < 1$ be two parameters. In a stochastic block model with K classes, there exists an unknown partition of the nodes $\{1, \ldots, n\}$ into K groups of same size, which is denoted by $G = \{G_1, G_2, \ldots, G_K\}$ with $|G_j| = n/K$. Then, the probability of connection P_{ij} is equal to p if i and j are in the

same group, and equal to q otherwise. The collections of such graph distributions is written $\mathrm{SBM}(n, K, p, q)$.

As in the previous chapter, we consider henceforth the setting where $p/q \leq \rho^*$ for some constant $\rho^* > 1$. Similarly as in the case $K = 2$ investigated in the previous chapter, the signal to noise ratio

$$s = \frac{(p-q)^2}{q + (p-q)/K} \tag{3.1}$$

drives the difficulty of community recovery when $p/q \leq \rho^*$.

First, [Banks et al., 2016] shows that when the number of nodes n goes to infinity and p, q scale as $1/n$ with n, non-trivial community recovery can be obtained if and only if $ns \gtrsim K \log(K)$. Polynomial-time algorithms are proposed in [Bordenave et al., 2018, Abbe and Sandon, 2015, Stephan and Massoulié, 2018] and achieve non-trivial community recovery when $ns > K^2$. For smaller s, it is conjectured in [Decelle et al., 2011] that non-trivial community recovery is impossible in polynomial-time.

In addition, for some constants $c > 1$, $c' > 0$ and when $ns > cK^2$, polynomial-time algorithms in [Chin et al., 2015, Gao et al., 2017, Fei and Chen, 2019, Giraud and Verzelen, 2019] enjoy a misclassification rate bounded from above by $\exp(-c'ns/K)$. The exponential decay with ns/K was shown to be optimal in [Zhang et al., 2016].

Let $\mu \in \mathrm{SBM}(n, K, p, q)$ be a distribution of a K-classes SBM. The set $\mathcal{E}^{good}(\mu)$ of good pairs is defined as the set of pairs $\{a, b\}$ with a and b from the same community. The set $\mathcal{E}^{bad}(\mu)$ of bad pairs is the set of pairs $\{a, b\}$ with a and b from two different communities. Since $p > q$, optimal strategies should sample as many pairs in $\mathcal{E}^{\mathrm{good}}$ as possible.

Formally, consider a time horizon T smaller than $|\mathcal{E}^{\mathrm{good}}| = K\binom{n/K}{2} \sim n^2/(2K)$. For any pair $e = \{a, b\}$ and any strategy ψ, let $N_e(\psi, T) \in \{0, 1\}$ indicate if the pair e has been sampled during the T queries of the strategy ψ. Then, the expected number of discoveries for ψ is equal to

$$\mathbb{E}_\mu \left[\sum_{t=1}^{T} A_{\widehat{e}_t} \right] = pT - (p-q)\, \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right]$$

where $N^{bad}(\psi, T) = \sum_{e \in \mathcal{E}^{bad}} N_e(\psi, T)$ is the number of sampled bad pairs up to time T. The maximal expected value of discoveries is achieved by any oracle strategy ψ^* sampling only edges in $\mathcal{E}^{\mathrm{good}}$. In that case, $N^{bad}(\psi^*, T) = 0$ and the expected number of discoveries is equal to pT. Therefore, the regret of the strategy ψ is

$$\mathbb{E}_\mu \left[R_T(\psi) \right] = pT - \mathbb{E}_\mu \left[\sum_{t=1}^{T} A_{\widehat{e}_t} \right] = (p-q)\, \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right].$$

Thus, as long as $T \leqslant |\mathcal{E}_{\mathrm{good}}|$, the regret is proportional to the expected number of sampled bad pairs $\mathbb{E}_\mu \left[N^{bad}(\psi, T) \right]$. Accordingly, we will analyse this last quantity rather than the regret, and will refer to it as the *sampling-regret*.

3.2.3 The case of geometric structure

A geometric graph corresponds to a modeling where a latent point x_i in a metric space is associated to each node, and the probability of connection P_{ij} is a function of the distance

between x_i and x_j.

In this chapter, we consider a specific distribution of geometric graph. Let $[0,1)$ be the tore endowed with the shortest path distance d, that is, $d(x,y) = |x-y| \wedge (1-|x-y|)$. We assume that there exist latent points $x_1, \dots, x_n \in [0,1)$ associated with the graph nodes $V = \{1, \dots, n\}$, and the probabilities of connection are equal to

$$P_{ij} = \frac{3}{4} - \frac{d(x_i, x_j)}{4}. \tag{3.2}$$

Note that $P_{ij} \in [1/2, 3/4]$. We also assume that the latent points $(x_1, \dots, x_n)$ are deterministic and take the special form $(1/n, 2/n, \dots, n/n)$ up to some unknown permutation of the coordinates, i.e. there exists an unknown permutation σ such that $x_i = \sigma(i)/n$ for all $i \in [n]$. Hereafter, the collection of such distributions is denoted by $\mathrm{GG}(n)$.

Let $\mu \in GG(n)$ be the distribution of a geometric graph with latent points $x_1, \dots, x_n$. Any strategy has an expected number of discoveries equal to

$$\mathbb{E}_\mu \left[\sum_{t=1}^{T} A_{\widehat{e}_t} \right] = \frac{3T}{4} - \frac{1}{4} \sum_{t=1}^{T} \sum_{a,b=1}^{n} d(x_a, x_b)\, \mathbb{E}_\mu \left[\mathbf{1}_{\widehat{e}_t = \{a,b\}} \right]. \tag{3.3}$$

Thus, our initial objective of maximizing (3.3) is equivalent to minimizing the following sum of distances

$$D(\psi, T) = \sum_{t=1}^{T} \sum_{a,b=1}^{n} d(x_a, x_b)\, \mathbb{E}_\mu \left[\mathbf{1}_{\widehat{e}_t = \{a,b\}} \right], \tag{3.4}$$

which we call *dispersion*. Let $D(\psi^*, T)$ denote the minimal value of dispersion achieved by an oracle strategy ψ^* sampling T pairs corresponding to T closest pairs of latent points in the tore $[0,1)$. Then, for any strategy ψ, the regret is equal to

$$\mathbb{E}_\mu \left[R_T(\psi) \right] = \frac{1}{4} \left(D(\psi, T) - D(\psi^*, T) \right),$$

which is proportional to the difference between the dispersion of ψ and the best possible dispersion in time horizon T.

3.2.4 Constraint for a balanced-sampling

Natural constraints are imposed on strategies in various applications. For example, in online video games, a bad algorithm matches a few players many times while keeping the others waiting. In order to avoid such undesirable features, we force strategies to explore the graph. It is actually necessary since optimal strategies do not do it naturally when no constraints are imposed on them. Indeed, the unconstrained optimal strategy of [Giraud et al., 2019] only focuses on a few nodes from a same community, by making a local exploration of the graph.

Accordingly, we will (also) study the situation where the following constraints are imposed on strategies. In SBM, the constraint (3.5) forces strategies to sample many different communities. Similarly in geometric graphs, the constraint (3.8) entails an exploration of a large portion of the latent space.

The definitions of the constraints require some notations. Let $N_B(\psi, T)$ denote the number of sampled pairs in $B \times B$ until time T, i.e. $N_B(\psi, t) = \sum_{\{a,b\} \in B \times B} N_{\{a,b\}}(\psi, t)$. For any node a in V, and any interval I in the tore $[0, 1]$, simply write $a \in I$ for $x_a \in I$.

3.2.4.1 Balanced Sampling Constraint in SBM

Denote by $N^{good}(\psi, T) = \sum_{e \in \mathcal{E}^{good}} N_e(\psi, T)$ the number of sampled good pairs up to time T. Then, for a strategy ψ interacting with a SBM of partition $(G_1, \dots, G_K)$, the constraint of balanced sampling is satisfied if the following property holds.

Balanced Sampling Constraint in SBM.
Let h_{BS}, c_G, c_P be in $(0, 1]$ and T_0 be a positive integer. With a probability larger than c_P and for $T \geq T_0$, the inequality

$$N_{G_j}(\psi, T) \geq h_{BS} \frac{T}{K} \tag{3.5}$$

holds for at least $c_G K$ different groups G_j.

The parameter h_{BS} represents the strength of the balanced sampling constraint. We will investigate how the optimal sampling-regret depends on this parameter.

We ask that the balanced sampling constraint (3.5) holds for T larger than some T_0 because, for some choices of c_G, c_P and h_{BS}, no algorithm can fulfill (3.5) for small values of T. Let us explain informally this point.

In the early stage where T is too small to get enough information on the latent partition $(G_1, \dots, G_K)$, it is likely that we cannot do better than random sampling of the pairs. Actually, as explained in Section 3.3.2.2, sub-linear regret can only be achieved for $T \gtrsim (1 \vee (h_{BS}K))(K/s)^2$, suggesting that there is nothing substantially better than random sampling up to this time. Let us check if (3.5) can be satisfied by a random sampling. For a random sampling, the probability to sample a pair within G_j is $1/K^2$, so the mean value of $N_{G_j}(\psi_{random}, T)$ is T/K^2. This last quantity is larger than $h_{BS}T/K$ if and only if $h_{BS}K \leq 1$, which means that a random sampling cannot satisfy the constraint (3.5) if $h_{BS}K > 1$. Hence, for $h_{BS}K > 1$, it is likely that no algorithm can fulfill (3.5) in the early stage $T \lesssim (1 \vee (h_{BS}K))(K/s)^2$. This informal reasoning can be made rigorous, as stated in the next lemma, proved in Appendix 3.A.3.1.

Lemma 3.2.1 *No algorithm ψ can fulfill the property (3.5) when*

$$h_{BS}K > \frac{128}{c_P c_G} \quad \text{and} \quad T < \left(\frac{c_p c_G}{2048(1+\rho^*)}\right)^2 \left(\frac{K}{s}\right)^2 h_{BS}K. \tag{3.6}$$

In order to avoid meaningless statements, we will assume in our results for balanced sampling strategies that at least one of the two conditions holds

$$h_{BS}K \leq \frac{128}{c_P c_G} \quad \text{or} \quad T \geq T_0 = \left(\frac{c_p c_G \rho^*}{2048}\right)^2 \left(\frac{K}{s}\right)^2 h_{BS}K. \tag{3.7}$$

3.2.5 Balanced Sampling Constraint in GG.

In SBM, the balanced sampling constraint (3.5) requires that a least a fixed proportion of the groups are sampled linearly in time. We require a similar constraint for the geometric graph model: in a fixed proportion of the unit interval $[0,1]$, any interval I with length at least $T^{-1/5}$ should be sampled proportionally to $|I|T$. More formally, we impose the following constraint.

Balanced Sampling Constraint in GG. *Let $c_G, c_P, c_l, h_{BS} \in]0,1]$ be positive constants. Assume that there exists an interval $U \subset [0,1]$ of length $|U| \geq c_G$ such that the following property holds with probability at least c_P. For any interval $I \subset U$ of length $|I| \geq c_l T^{-1/5}$, the number of sampled pairs in I is large enough to satisfy*

$$\sum_{t=1}^{T} \mathbf{1}_{\widehat{e}_t = \{a,b\} \in I \times I} \geq h_{BS}|I|T \ , \tag{3.8}$$

where $\{a,b\} \in I \times I$ means that x_a and x_b are in I.

As above, in the early stage of the exploration, it is likely that nothing smarter than some random sampling can be performed. In such a case, the probability to sample a pair within I is $|I|^2$ and the average number of sampled pairs within I is $|I|^2 T$. This last quantity is not larger than the lower bound in (3.8), since $|I|$ can be as small as $T^{-1/5}$. Hence, the constraint (3.8) cannot hold in the early stage of the exploration.

We also notice that an interval I contains $\asymp (|I|n)^2$ within pairs, so the constraint (3.8) can only hold if $|I|T \lesssim (|I|n)^2$ for all I with $|I| \gtrsim T^{-1/5}$. This is possible only if $T^{3/5} \lesssim n$. Such a condition appears in Corollary 3.3.5 and Conjecture 3.

3.3 Results

3.3.1 Assumption

Although the labeling of the nodes in $\{1,\dots,n\}$ is arbitrary, the performance of optimal strategies should not depend on this arbitrariness. Accordingly, as in the previous chapter, we restrict our analysis to the set of strategies that are invariant to labelling of nodes. For convenience of the reader, we recall how this assumption is formalized.

For any pair $e \in \mathcal{E}$ and any strategy ψ, recall that $N_e(\psi,t) := \mathbf{1}_{e \in \widehat{\mathcal{E}}_t}$ indicates if the pair e has been sampled up to time t. Let μ be a distribution on graph with nodes $V = \{1,\dots,n\}$ and σ be a permutation of V. For any pair $\{a,b\} \in \mathcal{E}$, let $\sigma(\{a,b\}) := \{\sigma(a), \sigma(b)\}$ and μ^σ denote the distribution of $(A_{\sigma(e)})_{e \in \mathcal{E}}$, where $(A_e)_{e \in \mathcal{E}}$ is distributed according to μ.

Invariance to labelling (IL). *The distribution of the outcomes of the strategy ψ is invariant by permutations of the nodes labels: For any graph distribution μ and any permutation σ on V, the distribution of $\left(N_e(\psi,t) : e \in \mathcal{E}, 1 \leq t \leq \binom{n}{2}\right)$ under μ^σ is the same as the distribution of $\left(N_{\sigma(e)}(\psi,t) : e \in \mathcal{E}, 1 \leq t \leq \binom{n}{2}\right)$ under μ.*

$$N_e(\psi,t)\big|_{\mu^\sigma} \overset{distrib}{=} N_{\sigma(e)}(\psi,t)\big|_{\mu} \tag{3.9}$$

3.3.2 Lower bounds and conjectures in SBM

Let us discuss the case where the number of groups K is larger than 2. Contrary to the two communities case studied in [Giraud et al., 2019], we expect a statistical-computational gap for $K \geq 5$. Below, we give a partial lower bound for the optimal rates conjectured in [Giraud et al., 2019], and then we extend this result to the constrained case.

Let Ψ_T be the set of strategies fulfilling the assumption (3.9), and $\Psi_{T,h_{BS}}$ the subset of constrained strategies, that is, those satisfying the constraint (3.5) and the assumption (3.9).

3.3.2.1 Unconstrained case Ψ_T

We first recall the conjecture from [Giraud et al., 2019, section 5] on the pair-matching rates, which involve the term s defined in (3.1). Let Ψ_T^{poly} denote the intersection of Ψ_T with polynomial-time algorithms.

Conjecture 1 *Assume that $0 \leq q < p \leq 1/2$ and $p/q \leq \rho^*$ for some $\rho^* > 1$. Assume also that $s \leq (c(1+\rho^*))^{-1}$ for some $c > 1$. Then, without computational constraint, we conjecture that*

$$\inf_{\psi \in \Psi_T} \mathbb{E}\left[N^{bad}(\psi, T)\right] \asymp \left(\left(\frac{K\log(K)}{s}\right)^2 \vee \frac{K\sqrt{T}}{s}\right) \wedge T, \tag{3.10}$$

where the constants involved in $\asymp$ only depend on ρ^*.

Moreover, with polynomial-time constraint,

$$\inf_{\psi \in \Psi_T^{poly}} \mathbb{E}\left[N^{bad}(\psi, T)\right] \asymp \left(\left(\frac{K^2}{s}\right)^2 \vee \frac{K\sqrt{T}}{s}\right) \wedge T, \tag{3.11}$$

where the constants in $\asymp$ only depend on ρ^*.

The above conjecture gives both upper bounds and matching lower bounds on the sampling-regret, thus characterizing the optimality for pair-matching. For a detailed discussion about this conjecture, we refer the reader to the previous chapter, or equivalently to [Giraud et al., 2019, section 5].

Lower bounds. As we can see in (3.10) and (3.11), there is a common term $(K\sqrt{T}/s) \wedge T$. We actually prove a lower bound matching this term in Theorem 3.3.1, by generalizing the approach developed in the previous chapter. The proof of this lower bound can be found in Appendix 3.A.2.

Theorem 3.3.1 *Under the assumptions of Conjecture 1, we have*

$$\inf_{\psi \in \Psi_T} \mathbb{E}_\mu\left[N^{bad}(\psi, T)\right] \geq \frac{1}{130}\left(\frac{K\sqrt{T}}{16(1+\rho^*)s} \wedge T\right).$$

for all $\mu \in SBM(n, K, p, q)$.

Hence, for the lower bounds of (3.10) and (3.11), the term $(K\sqrt{T}/s) \wedge T$ is proved, while the others, $(K\log(K)/s)^2$ and $(K^2/s)^2$, remain to be proven. Unfortunately, we do not have a proof for them. These conjectural terms are based on informal arguments presented in [Giraud et al., 2019, section 5] and reminded below.

The term $(K^2/s)^2$ is conjectured to be the barrier below which non-trivial recovery of the communities becomes impossible in polynomial time. Indeed, for a SBM of N nodes, it was conjectured in [Decelle et al., 2011] that no polynomial-time algorithm achieves non-trivial classification when $Ns \lesssim K^2$, which is commonly referred to as the Kesten-Stigum threshold. If the conjecture of [Decelle et al., 2011] holds, then, after spending T queries to observe all pairs of a graph of $N \asymp \sqrt{T}$ nodes, we only have a trivial classification when $T \lesssim (K^2/s)^2$. This makes us to believe that the sampling-regret would grow linearly with T as long as $T \lesssim (K^2/s)^2$. This results in the conjectural bound below

$$\inf_{\psi \in \Psi_T^{poly}} \mathbb{E}\left[N^{bad}(\psi, T)\right] \gtrsim \left(\frac{K^2}{s}\right)^2 \wedge T.$$

Similarly, the term $(K\log(K)/s)^2$ is based on some impossibility result in clustering. Specifically, non-trivial classification was proved in [Banks et al., 2016] to be impossible below the information theoretic threshold $Ns \lesssim K\log(K)$. Then, following the same lines as above, this suggest a linear regret as long as $T \lesssim (K\log(K)/s)^2$, thus implying the following lower bound

$$\inf_{\psi \in \Psi_T^{poly}} \mathbb{E}\left[N^{bad}(\psi, T)\right] \gtrsim \left(\frac{K\log(K)}{s}\right)^2 \wedge T.$$

Upper bounds. Beyond these impossible regimes, there remains the question of finding algorithms that match the performances (3.10) and (3.11).

For the case with polynomial-time constraint, when $T \gtrsim (K^2/s)^2$, we think that (3.11) can be achieved by a generalization of the algorithm of [Giraud et al., 2019], modulo some minor changes to handle the transition from 2 to K classes. Roughly, this algorithm seeks to identify $\asymp \sqrt{T}$ nodes in a single group and then pairs them together. The cost of identifying $\sqrt{T}$ nodes within a single group is expected to be $O((K/s)^2 + K\sqrt{T}/s)$, and the last step has no cost. So the regret of the algorithm should match the conjectured rate (3.11).

Proving that (3.10) holds when $T \gtrsim (K\log(K)/s)^2$ is a more delicate question. If there existed a turnkey algorithm with vanishing classification error for $Ns \gtrsim K\log(K)$, the method in [Giraud et al., 2019] could be used. However, such theoretical guarantees for clustering algorithms are unknown. The closest result is in [Zhang et al., 2016] where the authors prove a vanishing error when $Ns/(K\log(K)) \to \infty$. Hence, to prove (3.10), future work could investigate [Zhang et al., 2016] and try to get sharper results on clustering.

3.3.2.2 The constrained case $\Psi_{T,h_{BS}}$

We now study the case where strategies are constrained to sample a positive fraction of the communities. Recall that $\Psi_{T,h_{BS}}$ denotes the set of constrained strategies, i.e. those

fulfilling the sampling constraint (3.5) and the invariance assumption (3.9). Let $\Psi^{poly}_{T,h_{BS}}$ be the intersection of $\Psi_{T,h_{BS}}$ with polynomial-time algorithms.

As in the above sub-section, we start with a conjecture on the optimal sampling-regrets.

Conjecture 2 *Assume that (3.7) and the hypotheses of Conjecture 1 hold. Then, for any $\mu \in SBM(n,K,p,q)$, we conjecture that when $\Psi_{T,h_{BS}} \neq \emptyset$:*

$$\inf_{\psi \in \Psi_{T,h_{BS}}} \mathbb{E}_\mu \left[N^{bad}(\psi,T) \right] \asymp \left(\left(\frac{K \log(K)}{s} \right)^2 \vee \frac{K\sqrt{(h_{BS}K \vee 1)T}}{s} \right) \wedge T, \tag{3.12}$$

where the constants involved in $\asymp$ only depend on c_P, c_G and ρ^*.

Similarly, we conjecture that when $\Psi^{poly}_{T,h_{BS}} \neq \emptyset$:

$$\inf_{\psi \in \Psi^{poly}_{T,h_{BS}}} \mathbb{E} \left[N^{bad}(\psi,T) \right] \asymp \left(\left(\frac{K^2}{s} \right)^2 \vee \frac{K\sqrt{(h_{BS}K \vee 1)T}}{s} \right) \wedge T, \tag{3.13}$$

where, again, the constants involved in $\asymp$ only depend on c_P, c_G and ρ^*.

Before discussing the rates appearing in Conjecture 2, let us emphasize that the assumption (3.7) removes degenerate cases where $\Psi_{T,h_{BS}}$ is empty, but, even when (3.7) holds, $\Psi_{T,h_{BS}}$may still be empty for some choices of h_{BS}, c_P, c_G and T_0. In particular, as discussed above the Lemma 3.2.1, page 150, it is likely that, when no conditions on T are enforced, the condition $h_{BS}K < 1$ is needed for having $\Psi_{T,h_{BS}} \neq \emptyset$.

The term $Ks^{-1}\sqrt{T}$ in Conjecture 1 is replaced by $Ks^{-1}\sqrt{(h_{BS}K \vee 1)T}$ in Conjecture 2, due to the additional constraint of balanced sampling (3.5). If Conjecture 2 is true, this means that the constraint affects the optimal rates only if $h_{BS} \gtrsim K^{-1}$.

Let us give some intuition on the rate $T \wedge (Ks^{-1}\sqrt{(h_{BS}K \vee 1)T})$. Following the same lines of reasoning as in the previous chapter, a natural strategy in this constrained scenario is the following. First, for $c_G K$ groups, we identify $\asymp \sqrt{h_{BS}T/K}$ nodes per group, and then we match these nodes within their group. As the last step has no cost in terms of regret, the regret should be proportional to the number of pairs needed to identify the groups of these $c_G K \times \sqrt{h_{BS}T/K}$ nodes. Since we need $\asymp Ks^{-1}$ queries to identify the group of one node, the total regret is of order

$$c_G K \times \sqrt{h_{BS}T/K} \times Ks^{-1} = c_G K s^{-1}\sqrt{h_{BS}KT} \ ,$$

in order to identify the groups of $c_G K\sqrt{h_{BS}T/K}$ nodes. Hence, any strategy suffers a linear regret as long as T is smaller than $Ks^{-1}\sqrt{h_{BS}KT}$, and then a regret proportional to $Ks^{-1}\sqrt{h_{BS}KT}$. Finally, sampling the within group pairs, we have queried at most

$$c_G K \times h_{BS}T/K = c_G h_{BS}T < T$$

pairs. Therefore, after spending $c_G h_{BS}T$ queries as explained above, the constraint of balanced sampling (3.5) is satisfied, and we can spend the remaining queries with the smallest possible sampling regret. This last part is done using the strategy of the previous chapter: we identify

$\asymp \sqrt{T}$ nodes in one of the group (inducing a $O(Ks^{-1}\sqrt{T})$ sampling regret) and then we spend the remaining budget (at no cost) by matching them together. The resulting regret is thus

$$\asymp T \wedge \left(Ks^{-1}\sqrt{h_{BS}KT} + Ks^{-1}\sqrt{T} \right),$$

which is the rate conjectured in (3.12), up to the additional term $\left(s^{-1}K\log(K)\right)^2$ required for non trivial community recovery.

Lower-bounds. We can see that the term $K\sqrt{(h_{BS}K \vee 1)T}/s$ appears in both sampling-regrets (3.12) and (3.13). The next theorem gives a lower bound matching this term. Its proof, given in Appendix 3.3.2, is an extension of the proof of Theorem 3.3.1.

For simplicity, we assume below that $c_G K$ is an even integer. This assumption can be readily removed at the price of somewhat worst constants. By convention, when $\Psi_{T,h_{BS}} = \emptyset$, the infimum over $\Psi_{T,h_{BS}}$ is set to $+\infty$.

Theorem 3.3.2 *Under the assumptions of Conjecture 2, and for $c_G K$ an even integer, we have*

$$\inf_{\psi \in \Psi_{T,h_{BS}}} \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \geq \frac{c_G c_P}{130} \left(\frac{K\sqrt{((h_{BS}K) \vee 1)T}}{16(1+\rho^*)s} \wedge T \right)$$

for any $\mu \in SBM(n, K, p, q)$.

We thus obtain one part of the lower bounds for (3.12) and (3.13). The other parts, $(K\log(K)/s)^2$ and $(K^2/s)^2$, are purely conjectural and have already been discussed in details in the previous sub-section.

In the special case where $h_{BS} = 1$, we actually have a complete lower bound for (3.12). Indeed, for $h_{BS} = 1$, the optimal regret (3.12) gets simplified as follows

$$\inf_{\psi \in \Psi_{T,1}} \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \asymp \frac{K\sqrt{KT}}{s} \wedge T, \tag{3.14}$$

since the term $(K\log(K)/s)^2$ is negligible compared to the terms of (3.14). Therefore, we can see that Theorem 3.3.2 gives a complete lower bound for (3.14) in that specific case.

Upper-bounds. It remains to prove the upper bounds for Conjecture 2. We think that it can be done by adapting the proofs of those in Conjecture 1. The rationale for the corresponding algorithms would be the one depicted after Conjecture 2.

Finally, the tight link between both cases (unconstrained-constrained) suggests that the latter will follow once the former is solved.

3.3.3 Lower bound and conjecture in geometric graphs

Let us now investigate the optimal rate for pair-matching in the random geometric graph model (3.2). The regret $R_T(\psi)$ of a strategy ψ is proportional to the difference $D(\psi, T) -$

$D(\psi^*, T)$ between the dispersion $D(\psi, T)$ defined by (3.4) and the dispersion $D(\psi^*, T)$ of an oracle strategy. Accordingly, we evaluate the scaling of the oracle dispersion $D(\psi^*, T)$ and give a lower bound on the dispersion $D(\psi, T)$, from which we deduce a lower bound on the regret $R_T(\psi)$.

The following lemma gives the scaling of the dispersion of an oracle strategy, which is the best strategy among all the ideal strategies which benefit from a perfect knowledge of the hidden structure. Let Ψ_T^{id} denote the set of ideal strategies.

Lemma 3.3.3 *For ideal strategies, the smallest possible dispersion is of the order of*

$$\inf_{\psi^* \in \Psi_T^{id}} \mathbb{E}_\mu \left[D(\psi^*, T) \right] \asymp \frac{T}{n} \vee \left(\frac{T}{n} \right)^2 , \tag{3.15}$$

for any $\mu \in GG(n)$. Besides, there exist oracles satisfying both (3.15) and the sampling constraint (3.8).

The proof of Lemma 3.3.3 is given in the Appendix 3.B.1.

Moving on to procedures based on data, let Ψ_T^{cons} denote the set of strategies fulfilling the invariance to labeling (3.9) and sampling constraint (3.8).

By convention, when $\Psi_T^{cons} = \emptyset$, the infimum over Ψ_T^{cons} is set to $+\infty$ in the next two results.

Theorem 3.3.4 *There exist a numerical constant T_0 and a constant $C_B > 0$ depending only on the parameters of the assumption (3.8), such that, for any $T \geq T_0$ and $\mu \in GG(n)$, we have*

$$\inf_{\psi \in \Psi_T^{cons}} \mathbb{E}_\mu \left[D(\psi, T) \right] \geq C_B \left(T^{4/5} \vee \frac{T}{n} \vee \left(\frac{T}{n} \right)^2 \right) . \tag{3.16}$$

The term $(T/n) \vee (T/n)^2$ follows directly from Lemma 3.3.3 and the fact that a strategy based on data cannot outperform the oracle performance. The lower bound with respect to $T^{4/5}$ is proven in Appendix 3.B.2. The proof is adapted from the proof of Theorem 3.3.2.

As explained page 151, below the constraint (3.8), the condition (3.8) can only hold in the regime $n \gtrsim T^{3/5}$, where the above lower bound is of the order of $T^{4/5}$. We then have the immediate corollary.

Corollary 3.3.5 *There exist a numerical constant $c > 0$ and a constant $C'_B > 0$ depending only on the parameters of the assumption (3.8), such that, for any $n \geq cT^{3/5}$ and $\mu \in GG(n)$, we have*

$$\inf_{\psi \in \Psi_T^{cons}} \mathbb{E}_\mu \left[R_T(\psi) \right] \geq C'_B T^{4/5} .$$

Let us try to get some intuition on the $T^{4/5}$ rate by mimicking the analysis done for constrained SBM. Similarly as in Kleinberg's analysis of continuous armed bandits [Kleinberg, 2004], we can split the tore $[0, 1)$ into K intervals of size $1/K$. Then, we can seek to identify

$\asymp \sqrt{T/K}$ nodes per interval and finally to match them together in each interval. Thus, the total budget spent in the last matching step is $\asymp K \times (T/K) = T$ as desired. Assume that we have already identified the latent positions of a large number N of nodes uniformly spread over the K intervals. Then, the problem of locating a new node i, by querying pairs between i and the N reference nodes, can be approximated by a bandit problem, where each interval represents an arm. The separation between typical points of adjacent intervals is $1/K$. Classical arguments in bandit problems (with unimodal reward) suggest that the regret for identifying one node in one interval should be of the order of K (up to log factors). Hence, the total regret for identifying $\asymp \sqrt{T/K}$ nodes in each interval is expected to be of the order of (dropping all log factors)

$$K \times \sqrt{T/K} \times K = K^{3/2}\sqrt{T} \ .$$

When matching together pairs within each interval, we sample $\asymp T/K$ pairs within each interval, as required by the balanced sampling constraint (3.8). The cost for querying a pair within an interval is of order of $1/K$ (on average). So, the regret of the within interval matching is of order

$$K \times \frac{T}{K} \times \frac{1}{K} = \frac{T}{K}.$$

Hence, the total regret that we get is

$$\asymp K^{3/2}\sqrt{T} + T/K \quad \text{(up to log factors)}.$$

This regret is minimal for $K \asymp T^{1/5}$, leading to the $T^{4/5}$ regret rate.

The main flaw to turn the above reasoning into a valid proof, is that it is unclear how we can identify the locations of the N initial nodes, with a regret not exceeding $T^{4/5}$. Actually, the above approximation by the K-armed bandit problem is accurate if, the time horizon $\asymp K^2$ of each bandit, is smaller than the number $\asymp N/K$ of nodes per intervals. This implies that N should be larger than $K^3 = T^{3/5}$. So, a basic localization algorithm querying all the pairs within the N nodes would generate a regret larger than $T^{4/5}$.

Making the above argument rigorous is not immediate, and designing an algorithm matching the $T^{4/5}$ rate (up to possible log terms) is a delicate task. In Appendix 3.C, we describe an algorithm relying on two main ingredients: iteratively increasing the set of localized points (similarly in spirit as for constrained SBM in Chapter 2) and a divide-and-localize policy in order to keep the regret under control. The conjectural performance of this algorithm is as follows.

Conjecture 3 *The strategy ψ, described in Appendix 3.C, satisfies the sampling constraint (3.8) and the following inequality, for any $\mu \in GG(n)$,*

$$\mathbb{E}_\mu\left[D(\psi, T)\right] \lesssim T^{4/5}\log(T)^2 \log\log(T), \tag{3.17}$$

as soon as $n \geq cT^{3/5}\log(T)\log\log(T)$ for a large enough constant $c > 0$.

In the regime $n \gtrsim T^{3/5}\log(T)\log\log(T)$, the upper bound (3.17) matches the lower bound (3.16), up to some logarithmic factors. The combination of the two results characterizes the optimal dispersion in geometric graphs.

We do not have a complete proof for (3.17), but we have strong arguments to support it. Our arguments rely on the conjectural existence of a latent points localization algorithm satisfying the next property.

Conjectural latent points localization algorithm.
We conjecture the existence of an algorithm fulfilling the following localization property.
For any latent positions $x_1, \dots, x_N$ in the tore $[0,1)$, and for any adjacency matrix $\{A_{ij}\}_{1\leq i,j\leq N}$ generated by the random geometric graph model (3.2), there exist estimators $\hat{x}_1, \dots, \hat{x}_N$ such that for some distance preserving transformation $\widehat{Q}$ in the tore (translation or reflection)

$$\max_{i\in[N]} d(\widehat{Q}x_i, \hat{x}_i) \lesssim \sqrt{\frac{\log(N)}{N}}, \tag{3.18}$$

with large probability.

We do not prove this conjecture, but we present some supporting arguments below.

The $\sqrt{\log(n)/n}$ rate can be intuited as follows. In a thought experiment, assume that the locations of the latent points $x_1, \dots, x_{n-1}$ are revealed. Our goal is then to estimate x_n on the basis of the observations $(A_{in})_{i=1,\dots,n-1}$. We have $A_{in} = f(x_i, x_n) + E_{in}$, where $f(x,y) = 3/4 - d(x,y)/4$ and $E_{1n}, \dots, E_{(n-1)n}$ are $n-1$ independent sub-Gaussian random variables with variance proxy 1. For a candidate location $x \in [0,1]$, we can compute the residual sum of squares

$$\begin{aligned} RSS_n(x) &= \frac{1}{n-1}\sum_{i=1}^{n-1}(A_{in} - f(x_i,x))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1}(f(x_i,x_n) - f(x_i,x))^2 + \frac{2}{n-1}\sum_{i=1}^{n-1}(f(x_i,x_n) - f(x_i,x))E_{in} \\ &\quad + \frac{1}{n-1}\sum_{i=1}^{n-1} E_{in}^2. \end{aligned}$$

The difference $f(x_i,x_n) - f(x_i,x)$ behaves like $d(x,x_n)$. So, the first term in the right-hand side typically behaves like $d(x,x_n)^2$, while the second term behaves like a sub-Gaussian random variable with variance proxy $d(x,x_n)^2/n$. The third term does not depend on x. So, minimizing $RSS_n(x)$ over x on a regular grid with step size $1/n$, we find $\hat{x}_n$ such that $d(\hat{x}_n, x_n) = O(\sqrt{\log(n)/n})$ with large probability.

In the next chapter, the above argument is made rigorous to prove the existence of algorithms fulfilling (3.18) when $x_1, \dots, x_n$ are regularly spread on the tore $[0,1)$ and $P_{ij} = f(x_i, x_j)$ with f an (unknown) bi-Lipschitz function. Due to this extra assumption on the distribution of $x_1, \dots, x_n$, the result of the next chapter does not imply the above conjecture on latent points localization.

In appendix 3.C, we present a sketch of a possible proof for Conjecture 3, based on the above conjectural latent points localization algorithm.

3.4 Summary and perspectives

This chapter collects lower bounds for the pair-matching problem, by generalizing the techniques from Chapter 2. In terms of results, we prove partial lower bounds for K-classes SBM and complete lower bounds for geometric graphs.

We also introduce a constraint to make strategies explore the latent space. In that case, our conjectures show how the regret may depend on this constraint, which sheds light on the trade-off between exploration and exploitation.

However, the conjectures for optimal regrets are not proved. Specifically, in K-classes SBM, complete lower bounds remain to be proven, and the question is left open. As discussed earlier, our belief is that some impossibility results (or well accepted conjectures) in the clustering literature should be linked to the solution. On the other hand, the upper bounds will be the subject of future work. Indeed, we think that the strategy of [Giraud et al., 2019] could be generalized, at least to some extent. For instance, it should be easy in the regimes where the time horizon T is relatively large compared to the number of communities.

For geometric graphs, we think that our conjecture on optimal regret should be proved soon. Indeed, this chapter already presents matching lower bounds, together with a sketch of a possible proof for upper bounds (which match up to some logarithmic factors). But, this sketch for upper bounds relies on the conjecture of a latent points localization algorithm, which requires more work to be solved.

The pressing question of latent points localization is the subject of the next chapter. Although we obtain some results with the desired rate (3.18) (as we will see), these theorems are only proven under strict assumptions, which are not appropriate for the pair-matching problem. Therefore, latent points localization is still a work in progress.

Appendices

3.A Proof of the lower bound in SBM

In this section, we follow similar lines as in the proof of the lower bound of the previous chapter. However, for the sake of clarity, we write a complete proof.

The proof requires some notations. For any node $a \in [n]$, let $G(a) \subset [n]$ be the set of nodes from the same group as a. For any strategy ψ, define $N_a(\psi, t) = \sum_{b \in V : b \neq a} N_{\{a,b\}}(\psi, t)$ the number of times the node a has been sampled after t queries of the strategy ψ. For any set of nodes B, write $N_{aB}(\psi, t) = \sum_{b \in B} N_{\{a,b\}}(\psi, t)$ the number of times a pair between a and a node of B has been sampled up to time t. Recall that $\mathcal{E}$ denotes the set of all pairs in $\{1, \dots, n\}$.

3.A.1 Distributional properties under the assumption of invariance to labeling

The invariance to labelling property enforces some invariances on the distributions of $(N_e(\psi, T) : e \in \mathcal{E})$ and $(N_a(\psi, T) : a = 1, \dots, n)$ and $(N_{aG(a)}(\psi, T) : a = 1, \dots, n)$.

Let μ be a distribution in SBM(n, K, p, q) associated to a partition $G = \{G_1, G_2, \dots, G_K\}$ of $\{1, \dots, n\}$. Consider a permutation σ of $[n]$ which leaves the partition G invariant, that is, for all $i \in [K]$, $\sigma(G_i) = G_{\tau(i)}$ for some permutation τ of $[K]$. Then, the distribution μ^σ is equal to the distribution μ. Hence the invariance to labelling property ensures that for any permutation σ leaving G invariant, the vectors $(N_e(\psi, t) : e \in \mathcal{E}; t = 1, \dots, \binom{n}{2})$ and $(N_{\sigma(e)}(\psi, T) : e \in \mathcal{E}; t = 1, \dots, \binom{n}{2})$ have the same distribution. As a consequence, the following properties holds.

Claim 3.A.1 *When the strategy ψ fulfills the invariance to labelling property, then the random variables $(N_e(\psi, T) : e \in \mathcal{E}^{good})$ are pair-wise exchangeable. The same property holds for $(N_e(\psi, T) : e \in \mathcal{E}^{bad})$ and $(N_a(\psi, T) : a = 1, \dots, n)$ and $(N_{aG(a)}(\psi, T) : a = 1, \dots, n)$.*

Proof. Let $\{a, b\}, \{a', b'\}$ denote two pairs in $\mathcal{E}^{good}$ and let σ be a G-invariant permutation such that $\sigma(\{a, b\}) = \{a', b'\}$, and $\sigma(\{a', b'\}) = \{a, b\}$. Since $\mu = \mu^\sigma$ and ψ is invariant to labelling, the random variables $(N_{\{a,b\}}, N_{\{a',b'\}})$ and $(N_{\{a',b'\}}, N_{\{a,b\}})$ have the same distribution. The same reasoning applies for pairs in $\mathcal{E}^{bad}$.

Consider now two nodes $a, b \in \{1, \dots, n\}$. Let σ be a G-invariant permutation on $\{1, \dots, n\}$ such that $\sigma(a) = b$ and $\sigma(b) = a$. Since $\mu = \mu^\sigma$ and ψ is invariant to labelling, the random variables $(N_a(\psi, T), N_b(\psi, T))$ and $(N_b(\psi, T), N_a(\psi, T))$ have the same distribution. Likewise, the random variables $(N_{aG(a)}(\psi, T), N_{bG(b)}(\psi, T))$ and $(N_{bG(b)}(\psi, T), N_{aG(a)}(\psi, T))$ have the same distribution. □

Without assumption on ψ, the distribution of $N^{bad}(\psi, T)$ may depend on the distribution μ of the SBM. On the other hand, when the strategy ψ is assumed to be invariant to labeling, the distribution of $N^{bad}(\psi, T)$ does not depend on the distribution μ in SBM(n, K, p, q); See next claim.

Claim 3.A.2 *For any $\mu, \mu' \in$SBM(n, K, p, q), the distribution of $N^{bad}(\psi, T)$ under μ is the same as under μ'.*

Proof. Let μ, μ' be two distributions in SBM(n, K, p, q). By definition, there exists a permutation σ on $\{1, \dots, n\}$ such that $\mu' = \mu^\sigma$. Since $\mathcal{E}^{bad}(\mu^\sigma) = \sigma^{-1}(\mathcal{E}^{bad}(\mu))$, it follows from the invariance assumption (3.9) that the distribution under μ^σ of $\sum_{e \in \mathcal{E}^{bad}(\mu^\sigma)} N_e(\psi, T)$ is the same as the distribution under μ of $\sum_{e \in \mathcal{E}^{bad}(\mu)} N_e(\psi, T)$. □

3.A.2 Proof of Theorem 3.3.1

Let $kl(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$ be the Kullback-Leibler divergence between two Bernoulli distributions with means p and q. In the following, we set $\tilde{s} := kl(p, q) \vee kl(q, p)$. We will prove the following lower bound

$$\inf_{\psi \in \Psi_T} \mathbb{E}_\mu \left[N^{bad}(\psi, T) \right] \geq \frac{1}{130} \left[\frac{K\sqrt{T}}{16\tilde{s}} \wedge T \right], \tag{3.19}$$

for any $\mu \in$ SBM(n, K, p, q) and $\tilde{s} \leq 1/16$. Claim 3.D.1 ensures that $\tilde{s} \leq (1 + \rho^*)s$ under the assumptions of Theorem 3.3.1, so the Theorem 3.3.1 follows.

Let $N_a^{bad}(\psi, T)$ be the number of sampled pairs $\{a, b\}$ with b not in the community of a. Hereafter in the proof, the strategy ψ is fixed and, to simplify notations, the dependency of N_a and N_a^{bad} on ψ is dropped out: $N_a^{bad}(\psi, T)$ is denoted $N_a(T)$ and $N_a^{bad}(\psi, T)$ is denoted $N_a^{bad}(T)$. Let also $N_a^{good}(T) = N_a(T) - N_a^{bad}(T)$. The number of sampled pairs between-group is

$$N^{bad}(T) = \frac{1}{2} \sum_{a=1}^{n} N_a^{bad}(T).$$

For $t \geq 0$, let $\mathcal{F}_t$ be the σ-algebra gathering information available up to time t: $\mathcal{F}_t$ is the σ-algebra generated by $(\widehat{\mathcal{E}}_t, (A_e)_{e \in \widehat{\mathcal{E}}_t}, U_1, \dots, U_{t+1})$.

As in the previous chapter, the main tool for proving equation (3.19) is the next lemma, which is directly adapted from the Bandit literature.

Lemma 3.A.3 *Let $\tilde{T}$ be a stopping time with respect to the filtration $(\mathcal{F}_t)_{t \geq 0}$. Let $\mu, \mu' \in SBM(n, K, p, q)$ and let $P = (P_{ab})_{a<b}$ and $P' = (P'_{ab})_{a<b}$ denote their connection probabilities, that is $P_{ab} = \mu(\{a, b\} \in E)$ and $P'_{ab} = \mu'(\{a, b\} \in E)$ for all $a, b \in V$. If $\tilde{T} \leq T$ a.s., then for any $\mathcal{F}_{\tilde{T}}$-measurable random variable $\mathcal{Z}$ taking values in $[0, 1]$,*

$$\sum_{a<b} \mathbb{E}_\mu[N_{\{a,b\}}(\tilde{T})] kl(P_{ab}, P'_{ab}) \geq kl(\mathbb{E}_\mu[\mathcal{Z}], \mathbb{E}_{\mu'}[\mathcal{Z}]). \tag{3.20}$$

With this lemma, the core inequality of the proof can be established. This inequality shows that if $N_{aG(a) \cup G_k}(t) = O(1/\tilde{s})$, then $N^{bad}_{aG(a) \cup G_k}(t)$ is of the same order of magnitude than $N_{aG(a) \cup G_k}(t)$.

We remind the reader that n/K is an integer. Let $G = (G_1, G_2, \dots, G_K)$ be a partition of $\{1, \dots, n\}$ into K groups of same size, where $G_1 = \{1, \dots, n/K\}$, $G_2 = \{n/K + 1, \dots, 2n/K\}$, $\dots$, $G_K = \{((K-1)n/K) + 1, \dots, n\}$. Let $\mu \in SBM(n, K, p, q)$ be the distribution of a conditional SBM with classes $G_1, G_2, \dots, G_K$, within-group connection probability p and between-group connection probability q. Unless specified, $\mathbb{E} = \mathbb{E}_\mu$ in the following. The next lemma, which is a variant of Lemma 2.A.4 in Chapter 2, is proved in section 3.A.4.1.

Lemma 3.A.4 *Let M be a positive integer such that $16M\tilde{s} \leq 1$ and define the stopping time $\tilde{T} = T \wedge \inf\{t : \max(N_{1G_1 \cup G_K}(t), N_{nG_1 \cup G_K}(t)) \geq M\}$. Setting $N_{1+n}(T) = N_{1G_1 \cup G_K}(T) + N_{nG_1 \cup G_K}(T)$ and $N^{bad}_{1+n}(T) = N_{1G_K}(T) + N_{nG_1}(T)$, we have*

$$\mathbb{E}\left[N^{bad}_{1+n}(\tilde{T})\right] \geq \frac{1}{4}\mathbb{E}\left[N_{1+n}(\tilde{T})\right] \geq \frac{1}{4}\mathbb{E}\left[N_{1G_1 \cup G_K}(T) \wedge M\right]. \tag{3.21}$$

For proving the lower bound (3.19), we also need the next combinatorial lemma, whose proof is given in Section 3.A.4.2. Recall that $G(a) \subset [n]$ denotes the set of nodes in the same community as a.

Lemma 3.A.5 *Let M and T be two positive integers. If*

$$N^{good}(T) := \frac{1}{2}\sum_{a=1}^{n} N_{aG(a)}(T) \geq \frac{T}{2},$$

then

$$\sum_{a=1}^{n}(N_{aG(a)}(T) \wedge M) \geq \frac{1}{2}\left((M\sqrt{T}) \wedge \frac{T}{4}\right).$$

The lower bound (3.19) can now be proved. Recall that for any strategy $\psi \in \Psi_T$, the assumption of invariance to labeling implies that the sampling-regret $\mathbb{E}_\mu\left[N^{bad}(\psi, T)\right]$ does not depend on $\mu \in \mathrm{SBM}(n, K, p, q)$, see Claim 3.A.2. Therefore, it is sufficient to prove (3.19) for any strategy ψ invariant by labelling, with the distribution μ defined above Lemma 3.A.4.

Let M be a positive integer such that

$$1 \leq M \leq \frac{1}{16\tilde{s}}.$$

Claim 3.A.1 ensures that $\mathbb{E}\left[N_{\{a,b\}}(T)\right] = \mathbb{E}\left[N_{\{1,n\}}(T)\right]$ for any pair $\{a, b\} \in \mathcal{E}^{bad}$. Hence

$$\mathbb{E}\left[N^{bad}(T)\right] = \frac{n^2}{2}\frac{(K-1)}{K}\mathbb{E}\left[N_{\{1,n\}}(T)\right] \geq \frac{n(K-1)}{4}\mathbb{E}\left[N^{bad}_{1+n}(T)\right].$$

Then, using Lemma 3.A.4, we obtain for $K \geq 2$,

$$\begin{aligned} 32\,\mathbb{E}\left[N^{bad}(T)\right] \geq 8n(K-1)\,\mathbb{E}\left[N^{bad}_{1+n}(T)\right] &\geq 4nK\,\mathbb{E}\left[N^{bad}_{1+n}(\tilde{T})\right] \\ &\geq nK\,\mathbb{E}\left[N_{1G_1 \cup G_K}(T) \wedge M\right] \\ &\geq nK\,\mathbb{E}\left[N_{1G_1}(T) \wedge M\right]. \end{aligned}$$

Claim 3.A.1 ensures that $\mathbb{E}\left[N_{aG(a)}(T) \wedge M\right] = \mathbb{E}\left[N_{1G_1}(T) \wedge M\right]$ for all $a \in \{1, \ldots, n\}$, therefore the last inequalities give

$$32\,\mathbb{E}\left[N^{bad}(T)\right] \geq nK\,\mathbb{E}\left[N_{1G_1}(T) \wedge M\right] \geq K\sum_{a=1}^{n}\mathbb{E}\left[N_{aG(a)}(T) \wedge M\right]. \tag{3.22}$$

Combining with Lemma 3.A.5, it follows that

$$\begin{aligned}64\,\mathbb{E}\left[N^{bad}(T)\right] &\geq 2K\sum_{a=1}^{n}\mathbb{E}\left[N_{aG(a)}(T)\wedge M\right]\\ &\geq 2K\,\mathbb{E}\left[\mathbf{1}_{N^{good}(T)\geq T/2}\sum_{a=1}^{n}(N_{aG(a)}(T)\wedge M)\right]\\ &\geq K\left((M\sqrt{T})\wedge\frac{T}{4}\right)\mathbb{E}\left[\mathbf{1}_{N^{good}(T)\geq T/2}\right].\end{aligned}$$

Finally, since $T = N^{good}(T) + N^{bad}(T)$, we get

$$64\,\mathbb{E}\left[N^{bad}(T)\right] \geq K\left((M\sqrt{T})\wedge\frac{T}{4}\right)\mathbb{E}\left[\mathbf{1}_{N^{bad}(T)<T/2}\right].$$

On the other hand, we have

$$\mathbb{E}\left[N^{bad}(T)\right] \geq \mathbb{E}\left[\mathbf{1}_{N^{bad}(T)\geq T/2}N^{bad}(T)\right] \geq \frac{T}{2}\,\mathbb{E}\left[\mathbf{1}_{N^{bad}(T)\geq T/2}\right].$$

Hence, summing the last two inequalities, we obtain

$$\begin{aligned}65\,\mathbb{E}\left[N^{bad}(T)\right] &\geq \frac{T}{2}\,\mathbb{E}\left[\mathbf{1}_{N^{bad}(T)\geq T/2}\right] + K\left((M\sqrt{T})\wedge\frac{T}{4}\right)\mathbb{E}\left[\mathbf{1}_{N^{bad}(T)<T/2}\right]\\ &\geq \frac{T}{2}\wedge(KM\sqrt{T})\wedge\frac{KT}{4}\\ &\geq \frac{T}{2}\wedge(KM\sqrt{T}).\end{aligned}$$

For $\tilde{s}\leq 1/16$, taking M equal to the integer part of $1/(16\tilde{s})$ gives

$$65\,\mathbb{E}\left[N^{bad}(T)\right] \geq \frac{K\sqrt{T}}{32\tilde{s}}\wedge\frac{T}{2}.$$

Since the sampling-regret does not depend on the choice of μ, the proof of Theorem 3.3.1 is complete. □

3.A.3 Proof of Lemma 3.2.1 and Theorem 3.3.2

To prove Lemma 3.2.1 and Theorem 3.3.2, we start from Lemma 3.A.6 below, which is a variant of Lemma 3.A.5. Its proof is given in Section 3.A.4.3.

Lemma 3.A.6 *Let M and T be two positive integers. Let us write*

$$\mathcal{F}_{good} := \left\{\mathrm{Card}\left\{i\in[K]:\quad N_{G_i}(\psi,T)\geq h_{BS}\frac{T}{K}\right\}\;\geq\; c_G K\right\},$$

for the event where the balanced-sampling constraint (3.5) holds. Then, on the event $\mathcal{F}_{good}$, we have

$$\sum_{a=1}^{n}(N_{aG(a)}(T)\wedge M)\geq\frac{c_G}{2}\left((M\sqrt{h_{BS}KT})\wedge\frac{h_{BS}T}{2}\right).$$

Let M be a positive integer such that

$$1 \leq M \leq \frac{1}{16\tilde{s}}, \quad \text{for} \quad \tilde{s} = kl(p,q) \vee kl(q,p).$$

Following the same lines as in the proof of Theorem 3.3.1, we can start from the Inequality (3.22), which we recall for convenience:

$$32\,\mathbb{E}\left[N^{bad}(T)\right] \geq K \sum_{a=1}^{n} \mathbb{E}\left[N_{aG(a)}(T) \wedge M\right]. \tag{3.23}$$

Combining (3.23) with Lemma 3.A.6, and taking M equal to the integer part of $1/(16\tilde{s})$, it follows that

$$32\,\mathbb{E}\left[N^{bad}(T)\right] \geq \frac{c_G}{2} K \left(\frac{\sqrt{h_{BS}KT}}{32\tilde{s}} \wedge \frac{h_{BS}T}{2}\right) \mathbb{P}\left[\mathcal{F}_{good}\right]. \tag{3.24}$$

Both the Lemma 3.2.1 and the Theorem 3.3.2 follow from this lower bound. Let us start with the proof of Lemma 3.2.1.

3.A.3.1 Proof of Lemma 3.2.1

From (3.24) and $\mathbb{E}\left[N^{bad}(T)\right] \leq T$, we get that

$$\mathbb{P}\left[\mathcal{F}_{good}\right] \leq \frac{128T}{c_G K \left(\frac{\sqrt{h_{BS}KT}}{16\tilde{s}} \wedge (h_{BS}T)\right)} = \frac{128}{c_G}\left(\frac{1}{h_{BS}K} \vee \frac{16\tilde{s}\sqrt{T}}{K\sqrt{h_{BS}K}}\right).$$

Hence, $\mathbb{P}\left[\mathcal{F}_{good}\right] \geq c_P$ is not possible when

$$h_{BS}K > \frac{128}{c_P c_G} \quad \text{and} \quad T < \left(\frac{c_p c_G}{2048}\right)^2 (h_{BS}K)\left(\frac{K}{\tilde{s}}\right)^2.$$

The proof of Lemma 3.2.1 then follows since $\tilde{s} \leq (1+\rho^*)s$.

3.A.3.2 Proof of Theorem 3.3.2

When $h_{BS}K \leq 128/(c_P c_G)$, then (3.19) already ensures that

$$\mathbb{E}\left[N^{bad}(T)\right] \geq \frac{1}{130}\left[\frac{K\sqrt{T}}{16\tilde{s}} \wedge T\right].$$

For $h_{BS}K > 128/(c_P c_G)$ and $T \geq T_0$, we have from (3.24) and (3.5)

$$\mathbb{E}\left[N^{bad}(T)\right] \geq \frac{c_G c_P}{128} K \left(\frac{\sqrt{h_{BS}KT}}{16\tilde{s}} \wedge (h_{BS}T)\right).$$

(we notice that as $T \geq T_0$, with T_0 defined in (3.7), the right-hand side is always smaller than T.) Since $h_{BS}K > 128/(c_P c_G)$, we then obtain

$$\mathbb{E}\left[N^{bad}(T)\right] \geq \frac{c_G c_P}{128}\left(\frac{K\sqrt{h_{BS}KT}}{16\tilde{s}} \wedge T\right).$$

We have proved that

$$\inf_{\psi\in\Psi_{T,h_{BS}}} \mathbb{E}_\mu\left[N^{bad}(\psi,T)\right] \geq \frac{c_G c_P}{130}\left(\frac{K\sqrt{((h_{BS}K)\vee 1)T}}{16\tilde{s}} \wedge T\right),$$

and Theorem 3.3.2 follows from $\tilde{s} \leq (1+\rho^*)s$.

3.A.4 Proofs of lemmas

3.A.4.1 Proof of Lemma 3.A.4

The last inequality in (3.21) follows directly from

$$N_{1+n}(\tilde{T}) \geq N_{1G_1\cup G_K}(T)\mathbf{1}_{T<\tilde{T}} + M\mathbf{1}_{T\geq\tilde{T}} \geq N_{1G_1\cup G_K}(T)\wedge M.$$

It remains to show the first inequality. Consider the transposition $\sigma = (1,n)$ of 1 and n which switches the labels 1 and n while keeping other nodes unchanged. Let μ^σ be the distribution of $(A_{\sigma(a),\sigma(b)})_{ab}$. The partition $G^\sigma = \{G_1^\sigma, G_2^\sigma \ldots, G_K^\sigma\}$ associated to μ^σ, corresponds to G with 1 and n switched, that is $G_1^\sigma = \{n, 2, \ldots, n/K\}$, $G_K^\sigma = \{((K-1)n/K)+1, \ldots, n-1, 1\}$ and $G_j^\sigma = G_j$ for all $j \in [K]\setminus\{1,K\}$.

Let M be a positive integer and set

$$\mathcal{Z} = \frac{N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})}{2M} \in [0,1].$$

By invariance to labelling,

$$\mathbb{E}_{\mu^\sigma}\left[N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})\right] = \mathbb{E}_\mu\left[N_{1G_1}(\tilde{T}) + N_{nG_K}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right].$$

Hence, Lemma 3.A.3 ensures that,

$$\begin{aligned}
&(kl(p,q)\vee kl(q,p))\,\mathbb{E}_\mu\left[N_{1G_1\cup G_K}(\tilde{T}) + N_{nG_1\cup G_K}(\tilde{T})\right] \\
&\geq kl\left(\mathbb{E}_\mu\left[N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})\right]/(2M), \mathbb{E}_{\mu^\sigma}\left[N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})\right]/(2M)\right) \\
&= \frac{1}{2M} kl\left(\mathbb{E}_\mu\left[N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})\right], \mathbb{E}_\mu\left[N_{1G_1}(\tilde{T}) + N_{nG_K}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right]\right) \\
&\geq \frac{1}{2M}\frac{\left(\mathbb{E}_\mu\left[N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})\right] - \mathbb{E}_\mu\left[N_{1G_1}(\tilde{T}) + N_{nG_K}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right]\right)^2}{\mathbb{E}_\mu\left[N_{1G_K}(\tilde{T}) + N_{nG_1}(\tilde{T})\right] \vee \mathbb{E}_\mu\left[N_{1G_1}(\tilde{T}) + N_{nG_K}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right]},
\end{aligned}$$

where the last line follows from Claim 3.D.1. Setting $N_{1+n}^{good}(T) = N_{1G_1}(T) + N_{nG_K}(T)$, the last inequality can be written as

$$\begin{aligned}
2M\tilde{s}\,\mathbb{E}\left[N_{1+n}(\tilde{T})\right]&\left(\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right] \vee \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right]\right) \\
&\geq \left(\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right] - \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right]\right)^2. \quad (3.25)
\end{aligned}$$

If $\mathbb{E}\left[N_{1+n}^{good}(\tilde{T})\right] \leq \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right]$, then

$$2\,\mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right] \geq \mathbb{E}\left[N_{1+n}^{good}(\tilde{T})\right] + \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right] = \mathbb{E}\left[N_{1+n}(\tilde{T})\right]$$

and Lemma 3.A.4 follows.

Assume therefore that $\mathbb{E}\left[N_{1+n}^{good}(\tilde{T})\right] \geq \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right]$. It follows that

$$2\,\mathbb{E}\left[N_{1+n}^{good}(\tilde{T})\right] \geq \mathbb{E}\left[N_{1+n}^{good}(\tilde{T})\right] + \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right] = \mathbb{E}\left[N_{1+n}(\tilde{T})\right],$$

so inequality (3.25) implies

$$4M\tilde{s}\,\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right]^2 \geq \left(\mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right] - \mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right]\right)^2.$$

Rearranging the expression gives

$$\begin{aligned}\mathbb{E}\left[N_{1+n}^{bad}(\tilde{T})\right] &\geq \mathbb{E}\left[N_{1+n}^{good}(\tilde{T}) + 2N_{\{1,n\}}(\tilde{T})\right]\left(1 - \sqrt{4M\tilde{s}}\right) \\ &\geq \frac{1}{2}\,\mathbb{E}\left[N_{1+n}^{good}(\tilde{T})\right] \geq \frac{1}{4}\,\mathbb{E}\left[N_{1+n}(\tilde{T})\right],\end{aligned}$$

since $M \leq 1/(16\tilde{s})$ by assumption. The proof of Lemma 3.A.4 is complete. □

3.A.4.2 Proof of Lemma 3.A.5

Decomposing the sum, we have

$$\sum_{a=1}^{n}(N_{aG(a)}(T) \wedge M) = \sum_{j\in[K]}\sum_{a\in G_j}(N_{aG_j}(T) \wedge M).$$

Recall that $N_{G_j}(T)$ is the number of sampled pairs between two nodes of G_j.

Claim 3.A.7 *The inequality*

$$\forall j \in [K], \quad \sum_{a\in G_j}(N_{aG_j}(T) \wedge M) \geq \left(M\sqrt{N_{G_j}(T)}\right) \wedge \frac{N_{G_j}(T)}{2}$$

holds true.

From the above claim we deduce that

$$\sum_{a=1}^{n}(N_{aG(a)}(T) \wedge M) \geq \sum_{j\in[K]}\left(M\sqrt{N_{G_j}(T)}\right) \wedge \frac{N_{G_j}(T)}{2}.$$

Hence, to prove Lemma 3.A.5 it suffices to show the next inequality

$$\sum_{j\in[K]}\left(M\sqrt{N_{G_j}(T)}\right) \wedge \frac{N_{G_j}(T)}{2} \geq \frac{1}{2}\left((M\sqrt{T}) \wedge \frac{T}{4}\right). \tag{3.26}$$

Define C_1 the set of indices $j \in [K]$ fulfilling

$$(M\sqrt{N_{G_j}(T)}) \wedge \frac{N_{G_j}(T)}{2} = \frac{N_{G_j}(T)}{2},$$

and C_2 all j such that

$$(M\sqrt{N_{G_j}(T)}) \wedge \frac{N_{G_j}(T)}{2} = M\sqrt{N_{G_j}(T)}.$$

Depending on the value of $\sum_{j\in C_1} N_{G_j}(T)$, we distinguish the two following cases.

- If $\sum_{j\in C_1} N_{G_j}(T) \geq T/4$, then

$$\sum_{j\in[K]} \left((M\sqrt{N_{G_j}(T)}) \wedge \frac{N_{G_j}(T)}{2} \right) \geq \sum_{j\in C_1} \frac{N_{G_j}(T)}{2} \geq \frac{T}{8}.$$

- If $\sum_{j\in C_1} N_{G_j}(T) < T/4$, we have

$$\sum_{j\in C_2} N_{G_j}(T) = N^{good}(T) - \sum_{j\in C_1} N_{G_j}(T) \geq \frac{T}{4} \tag{3.27}$$

since $N^{good}(T) \geq T/2$ by assumption. Then, using the basic inequality $\sum_i \sqrt{y_i} \geq \sqrt{\sum_i y_i}$ for any sequence of positive reals $y_1, y_2, \ldots$, we get

$$\sum_{j\in[K]} \left(M\sqrt{N_{G_j}(T)} \right) \wedge \frac{N_{G_j}(T)}{2} \geq M \sum_{j\in C_2} \sqrt{N_{G_j}(T)} \geq M\sqrt{\sum_{j\in C_2} N_{G_j}(T)},$$

so that, by (3.27),

$$\sum_{j\in[K]} \left(M\sqrt{N_{G_j}(T)} \right) \wedge \frac{N_{G_j}(T)}{2} \geq \frac{M\sqrt{T}}{2}.$$

The combination of the two cases gives inequality (3.26), and Lemma 3.A.5 follows. □

Proof of Claim 3.A.7. It is enough to prove the inequality of the claim for the case $j = 1$, hence we only consider the group G_1. Define the sets $S_1 = \{a \in G_1 : N_{aG_1}(T) \leq M\}$ and $S_2 = \{a \in G_1 : N_{aG_1}(T) > M\}$. If $\sum_{a\in S_1} N_{aG_1}(T) \geq N_{G_1}(T)/2$ then $\sum_{a\in G_1}(N_{aG_1}(T) \wedge M) \geq \sum_{a\in S_1} N_{aG_1}(T) \geq N_{G_1}(T)/2$, and the inequality of the claim follows for this case.

Assume now that $\sum_{a\in S_1} N_{aG_1}(T) < N_{G_1}(T)/2$. Since $2N_{G_1}(T) = \sum_{a\in G_1} N_{aG_1}(T)$, we have

$$\begin{aligned} 2N_{G_1}(T) \leq N_{G_1}(T)/2 + \sum_{a\in S_2} N_{aG_1}(T) &= N_{G_1}(T)/2 + \sum_{a\in S_2} N_{aS_1}(T) + \sum_{a\in S_2} N_{aS_2}(T) \\ &= N_{G_1}(T)/2 + \sum_{a\in S_1} N_{aS_2}(T) + \sum_{a\in S_2} N_{aS_2}(T) \\ &\leq N_{G_1}(T) + |S_2|^2. \end{aligned}$$

Hence, $|S_2| \geq \sqrt{N_{G_1}(T)}$ and

$$\sum_{a\in G_1} (N_{aG_1}(T) \wedge M) \geq |S_2|M \geq (M\sqrt{N_{G_1}(T)}).$$

The inequality of the claim follows from the above. □

3.A.4.3 Proof of Lemma 3.A.6

Following the first paragraph in the proof of Lemma 3.A.5, we get from Claim 3.A.7

$$\sum_{a=1}^{n}(N_{aG(a)}(T)\wedge M)\geq\sum_{j\in[K]}\left(M\sqrt{N_{G_j}(T)}\right)\wedge\frac{N_{G_j}(T)}{2}.$$

Hence, to prove Lemma 3.A.6, we only need to prove that on the event $\mathcal{F}_{good}$, the inequality

$$\sum_{j\in[K]}\left(M\sqrt{N_{G_j}(T)}\right)\wedge\frac{N_{G_j}(T)}{2}\geq\frac{c_G}{2}\left((M\sqrt{h_{BS}KT})\wedge\frac{h_{BS}T}{2}\right),$$

holds.

We use the decomposition on the sets C_1 and C_2 as in Lemma 3.A.5 :

$$\begin{aligned}\sum_{j\in[K]}\left(M\sqrt{N_{G_j}(T)}\right)\wedge\frac{N_{G_j}(T)}{2}&\\ \geq\sum_{j\in C_1}\frac{N_{G_j}(T)}{2}+\sum_{j\in C_2}\left(M\sqrt{N_{G_j}(T)}\right)&\end{aligned}\tag{3.28}$$

and lower bound separately the two terms.

Let us work on the event $\mathcal{F}_{good}$. Then, at least a constant fraction c_G of the K classes, say $G_{j_1},\dots,G_{j_{c_GK}}$, fulfills

$$N_{G_{j_r}}(T)\geq h_{BS}\frac{T}{K},\tag{3.29}$$

for $r\in[c_GK]$. Depending on the number of classes $G_{j_1},\dots,G_{j_{c_GK}}$ that are in C_1, we distinguish two cases:

- If at least $c_GK/2$ classes among the $G_{j_1},\dots,G_{j_{c_GK}}$ belong to C_1, then

$$\sum_{j\in C_1}\frac{N_{G_j}(T)}{2}\geq\frac{c_GK}{2}\frac{h_{BS}T}{2K}=c_Gh_{BS}\frac{T}{4},$$

  using (3.29) in the first inequality.

- Otherwise, at least $c_GK/2$ classes among the $G_{j_1},\dots,G_{j_{c_GK}}$ belong to C_2 and

$$\sum_{j\in C_2}\left(M\sqrt{N_{G_j}(T)}\right)\geq\frac{c_GK}{2}M\sqrt{\frac{h_{BS}T}{K}}=c_GM\frac{\sqrt{h_{BS}KT}}{2},$$

  by inequality (3.29).

Plugging the last two displays in (3.28), we get on the event $\mathcal{F}_{good}$

$$\sum_{j\in[K]}\left(M\sqrt{N_{G_j}(T)}\right)\wedge\frac{N_{G_j}(T)}{2}\geq\frac{c_G}{2}\left(M\sqrt{h_{BS}KT}\wedge\frac{h_{BS}T}{2}\right).\tag{3.30}$$

3.B Proof of the lower bound in geometric graphs

3.B.1 Proof of Lemma 3.3.3

Without loss of generality, assume that the latent positions $x_1, \dots, x_n$ are already in order, that is, $x_i = i/n$ for all $i = 1, \dots, n$ (note that the equality $x_i = i/n$ is seen between two elements of the tore $[0,1]/\{0 = 1\}$). Given an integer l and a node a, define $\mathcal{L}_a(l)$ as the set of l-left nearest neighbors of a, i.e.:

$$\mathcal{L}_a(l) := \big\{b \in \{a-1, a-2, \dots, a-l\}\big\}$$

which is a set seen in the tore $[n]/\{0 = n\}$. Since an ideal strategy has perfect knowledge of the ground truth, it can minimize the dispersion by querying only pairs of nodes that are close neighbors in $[0,1]$. Below, we compute the dispersion of such a strategy ψ^*. Let $\mu \in \mathrm{GG}(n)$.

If $T \leq n-1$, then

$$\mathbb{E}_\mu\left[D(\psi^*, T)\right] = \sum_{a=1}^{T} \sum_{b \in \mathcal{L}_a(1)} d(x_a, x_b) = \frac{T}{n}.$$

Assume now that $T \geq n$, and denote by $m = \lceil T/n \rceil$ the ceiling of T/n. Then,

$$\inf_{\psi \in \Psi_T^{id}} \mathbb{E}_\mu\left[D(\psi, T)\right] \asymp \sum_{a=1}^{n} \sum_{b \in \mathcal{L}_a(m)} d(x_a, x_b),$$

which is of the order of

$$n \sum_{b=1}^{m} \frac{b}{n} \asymp m^2 \asymp \left(\frac{T}{n}\right)^2.$$

The equation (3.15) of the lemma derives from the above displays. Besides, ψ^* can be chosen so that the sampling constraint (3.8) is satisfied. Hence, Lemma 3.3.3 is proved. □

3.B.2 Proof of Theorem 3.3.4

No strategy can have a smaller dispersion than the oracle value given in Lemma 3.3.3, hence

$$\inf_{\psi \in \Psi_T^{cons}} \mathbb{E}_\mu\left[D(\psi, T)\right] \gtrsim \frac{T}{n} \vee \left(\frac{T}{n}\right)^2.$$

This gives two terms in the lower bound of Theorem 3.3.4.

If $T \geq 2^{-5} n^{5/3}$, then Theorem 3.3.4 follows directly from the above display. Therefore, we focus in the remaining of the proof on the case $T \leq 2^{-5} n^{5/3}$.

The proof for the remaining term (in the lower bound of Theorem 3.3.4) requires some notations. Let $\delta \in]0,1[$ and $\mu \in GG(n)$ the distribution of a geometric graph with latent points $x_1, \dots, x_n$. Then, for any node $a \in V$, define its neighborhood $V_{(a)}$ as

$$V_{(a)} = \{x \in [0,1] : \ d(x, x_a) \leq \delta\}.$$

Let $\mathcal{E}^{bad}(\mu)$ be the set of all pairs $\{a,b\}$ such that $x_b \notin V_{(a)}$, or equivalently $d(x_a,x_b) > \delta$. Throughout the proof, the pairs in $\mathcal{E}^{bad}(\mu)$ are called bad pairs. Let $N_a^{bad}(\psi,T)$ be the number of sampled bad pairs involving the node a, up to time T. The total number of sampled bad pairs is

$$N^{bad}(\psi,T) = \frac{1}{2}\sum_{a=1}^{n} N_a^{bad}(\psi,T).$$

In the dispersion $D(\psi,T)$, a sampled bad pair entails a term larger than δ, so that

$$\mathbb{E}_\mu\left[D(\psi,T)\right] \geq \delta\,\mathbb{E}_\mu\left[N^{bad}(\psi,T)\right].$$

We will show the lemma below by adapting the proofs seen in SBM.

Lemma 3.B.1 *Let δ be a real such that $2n^{-1} \leq \delta \leq 2^{-11}$. Then, for any $\mu \in GG(n)$, there exists a constant $C_{3.B.1}$, depending only on the parameters of the assumption (3.8), such that*

$$\inf_{\psi\in\Psi_T^{cons}} \mathbb{E}_\mu\left[N^{bad}(\psi,T)\right] \geq C_{3.B.1}\left(\frac{\sqrt{T}}{\delta^{5/2}} \wedge T\right). \tag{3.31}$$

Combining the two last inequalities, we get

$$\inf_{\psi\in\Psi_T^{cons}} \mathbb{E}_\mu\left[D(\psi,T)\right] \geq C_{3.B.1}\left(\frac{\sqrt{T}}{\delta^{3/2}} \wedge \delta T\right).$$

We can choose $\delta := T^{-1/5}$ since it satisfies the condition of Lemma 3.B.1, that is, $2n^{-1} \leq \delta \leq 2^{-11}$. Indeed, we have $2n^{-1} \leq T^{-1/5}$ because $T \leq 2^{-5}n^{5/3}$ by assumption. And we also have $T^{-1/5} \leq 2^{-11}$ for $T \geq T_0 = 2^{55}$. Therefore, we can take $\delta := T^{-1/5}$ in the last display to obtain

$$\inf_{\psi\in\Psi_T^{cons}} \mathbb{E}_\mu\left[D(\psi,T)\right] \geq C_{3.B.1}\, T^{4/5},$$

and the proof of Theorem 3.3.4 is complete.

3.B.3 Proof of Lemma 3.B.1

In this section, we follow similar lines as in the proofs for SBM.

3.B.3.1 Distributional properties under the assumption of invariance to labeling

Let μ be a distribution in GG(n). Consider a permutation σ that preserves the distances, that is, $d(x_{\sigma(i)},x_{\sigma(j)}) = d(x_i,x_j)$ for all $i,j \in [n]$. The distribution μ^σ is therefore equal to the distribution μ. It follows from the invariance to labelling assumption that the vectors $(N_e(\psi,t) : e \in \mathcal{E}; t = 1,\dots,\binom{n}{2})$ and $(N_{\sigma(e)}(\psi,T) : e \in \mathcal{E}; t = 1,\dots,\binom{n}{2})$ have the same distribution. As a consequence, the following properties hold.

Claim 3.B.2 *When the strategy ψ fulfills the invariance to labelling property (3.9), the random variables $(N_a(\psi,T) : a = 1,\dots,n)$ are pair-wise exchangeable. The same property holds for $(N_a^{bad}(\psi,T) : a = 1,\dots,n)$.*

Proof. Consider two nodes $a, b \in \{1, \dots, n\}$. Let σ be a permutation preserving the latent distances such that $\sigma(a) = b$ and $\sigma(b) = a$. Since $\mu = \mu^\sigma$ and ψ is invariant to labelling, the random variables $(N_a(\psi, T), N_b(\psi, T))$ and $(N_b(\psi, T), N_a(\psi, T))$ have the same distribution. Likewise, the random variables $(N_a^{bad}(\psi, T), N_b^{bad}(\psi, T))$ and $(N_b^{bad}(\psi, T), N_a^{bad}(\psi, T))$ have the same distribution. □

The assumption (3.9) implies that $N^{bad}(\psi, T)$ does not depend on $\mu \in \mathrm{GG}(n)$.

Claim 3.B.3 *For any $\mu, \mu' \in GG(n)$, the distribution of $N^{bad}(\psi, T)$ under μ is the same as under μ'.*

Proof. Let μ, μ' be two distributions in $\mathrm{GG}(n)$. By definition, there exists a permutation σ on $\{1, \dots, n\}$ such that $\mu' = \mu^\sigma$. Since $\mathcal{E}^{bad}(\mu^\sigma) = \sigma^{-1}(\mathcal{E}^{bad}(\mu))$, it follows from assumption (3.9) that the distribution under μ^σ of $\sum_{e \in \mathcal{E}^{bad}(\mu^\sigma)} N_e(\psi, T)$ is the same as the distribution under μ of $\sum_{e \in \mathcal{E}^{bad}(\mu)} N_e(\psi, T)$. □

3.B.3.2 Proof of of Lemma 3.B.1

Hereafter in the proof, the strategy ψ is fixed and to simplify notations, $N_a^{bad}(\psi, T)$ is denoted by $N_a(T)$, and $N_a^{bad}(\psi, T)$ by $N_a^{bad}(T)$. Let also $N_a^{good}(T) = N_a(T) - N_a^{bad}(T)$.

Lemma 3.A.3 is again the main tool to derive the core inequality of the proof, which is stated in Lemma 3.B.4. This inequality shows that if $N_a(t) = O(\delta^{-2})$, then $N_a^{bad}(t)$ is of the same order of magnitude as $N_a(t)$.

Let $x_1, \dots, x_n$ be latent points satisfying $x_i = i/n$ for all $i \in [n]$, and let δ be a real such that $2/n \le \delta \le 2^{-11}$. Denote $V = (V_1, V_2, \dots, V_n)$ their respective neighborhoods of diameter 2δ, that is, $V_i = \{x \in [0, 1] : d(x, x_i) \le \delta\}$. Let $\mu \in GG(n)$ be the distribution of a geometric graph with the latent points $x_1, \dots, x_n$. In the rest of the proof, we write $\mathbb{E} = \mathbb{E}_\mu$.

Among the $x_1, \dots, x_n$, let x_k denote a point in between 3δ and 4δ-away from x_1, that is, satisfying the inequalities $3\delta \le d(x_1, x_k) \le 4\delta$. The next lemma corresponds to Lemma 3.A.4 for SBM, and is proved in section 3.B.4.

Lemma 3.B.4 *Let M be a positive integer such that $M\delta^2 \le 2^{-11}$, and define the stopping time $\tilde{T} = T \wedge \inf\{t : \max(N_1(t), N_k(t)) \ge M\}$. Setting $N_{1+k}(T) = N_1(T) + N_k(T)$ and $N_{1+k}^{good}(T) = N_1^{good}(T) + N_k^{good}(T)$ and $N_{1+k}^{bad}(T) = N_1^{bad}(T) + N_k^{bad}(T)$, we have*

$$\mathbb{E}\left[N_{1+k}^{bad}(\tilde{T})\right] \ge \frac{1}{4}\,\mathbb{E}\left[N_{1+k}(\tilde{T})\right] \ge \frac{1}{4}\,\mathbb{E}\left[N_1(T) \wedge M\right]. \tag{3.32}$$

For proving the lower bound, we also need the next combinatorial result, which is based on the constraint of balanced sampling. The proof derives from Lemma 3.A.6 for SBM. Indeed, by analogy with SBM, let K be an integer of the order of δ^{-1}, and let $I_1, \dots, I_K$ be a partition of the latent space $[0, 1]$ into K segments of equal length, except for the last segment I_K at the end of $[0, 1]$ which may be smaller. Accordingly, let $G_1, \dots, G_K$ be the K groups of corresponding nodes, that is, whose latent points are respectively in $I_1, \dots, I_K$. Note that it is not exactly as in SBM, since the last group G_K may have a smaller cardinality than the

other groups, and also the $K-1$ first groups $G_1, \dots, G_{K-1}$ may not have exactly the same cardinality (up to a ± 1 difference). However, we can still follow the same lines as in the proof of Lemma 3.A.6 and obtain the following lemma. The proof is sketched in section 3.B.4.

We denote by $G(a)$ the group of nodes in which a belongs to.

Lemma 3.B.5 *Let M and T be two positive integers, and δ be a real such that $2n^{-1} \leq \delta \leq 2^{-11}$. Then, there exists a constant $C_{3.B.5}$, depending only on the parameters of the assumption (3.8), such that*

$$\mathbb{E}\left[\sum_{a=1}^{n}(N_{aG(a)}(T) \wedge M)\right] \geq C_{3.B.5}\left((M\sqrt{\frac{T}{\delta}}) \wedge T\right).$$

The lower bound (3.31) can now be proved. Notice that for any strategy $\psi \in \Psi_T$, Claim 3.B.3 implies that $\mathbb{E}_\mu\left[N^{bad}(\psi, T)\right]$ does not depend on $\mu \in \mathrm{GG}(n)$. Therefore, it is sufficient to prove (3.31) with the distribution μ defined above Lemma 3.B.4.

Let M be a positive integer such that

$$1 \leq M \leq 2^{-11}\delta^{-2}.$$

Claim 3.B.2 ensures that $\mathbb{E}\left[N_a^{bad}(T)\right] = \mathbb{E}\left[N_b^{bad}(T)\right]$ for any nodes $a, b \in [n]$, hence

$$\mathbb{E}\left[N^{bad}(T)\right] \geq \frac{n}{2}\mathbb{E}\left[N_1^{bad}(T)\right] \geq \frac{n}{4}\mathbb{E}\left[N_{1+k}^{bad}(T)\right].$$

By Lemma 3.B.4, it follows that

$$16\,\mathbb{E}\left[N^{bad}(T)\right] \geq 4n\,\mathbb{E}\left[N_{1+k}^{bad}(T)\right] \geq n\,\mathbb{E}\left[N_1(T) \wedge M\right].$$

Claim 3.B.2 ensures that $\mathbb{E}\left[N_a(T) \wedge M\right] = \mathbb{E}\left[N_1(T) \wedge M\right]$ for all $a \in [n]$, so that

$$16\,\mathbb{E}\left[N^{bad}(T)\right] \geq n\,\mathbb{E}\left[N_1(T) \wedge M\right] = \sum_{a=1}^{n}\mathbb{E}\left[N_a(T) \wedge M\right]. \tag{3.33}$$

Hence

$$16\,\mathbb{E}\left[N^{bad}(T)\right] \geq \sum_{a=1}^{n}\mathbb{E}\left[N_{aG_{(a)}}(T) \wedge M\right] \geq C_{3.B.5}\left((M\sqrt{\frac{T}{\delta}}) \wedge T\right),$$

using Lemma 3.B.5 in the last inequality. For $\delta^2 \leq 2^{-11}$, we can take M equal to the integer part of $2^{-11}\delta^{-2}$ to get

$$\mathbb{E}\left[N^{bad}(T)\right] \geq 2^{-16}C_{3.B.5}\left(\frac{\sqrt{T}}{\delta^{5/2}} \wedge T\right).$$

Since $N^{bad}(T)$ does not depend on the choice of distribution μ, Lemma 3.B.1 is proved. □

3.B.4 Technical lemmas

3.B.4.1 Proof of Lemma 3.B.5

We use the analogy with SBM. Since the strategy ψ satisfies the property of balanced-sampling (3.8) for geometric graphs, it also fulfills the constraint (3.5) for SBM, with respect to the notations introduced above Lemma 3.B.5. In fact, the constraint (3.5) for SBM may not hold for the last group G_K because the corresponding interval I_K may be smaller than the others I_i, $i \neq K$. Also, the $K-1$ other groups may not have exactly the same cardinal numbers, with a ± 1 difference in their respective cardinal numbers. However, for $K = \lceil \delta^{-1} \rceil$ with $\delta \leq 2^{-11}$, the number of communities is large enough so that these minor differences from the SBM scenario are without consequences on the proof of Lemma 3.A.6. Hence, we can still follow the same lines as in Lemma 3.A.6 for SBM (taking $h_{BS} \asymp 1$) to get the desired bound of Lemma 3.B.5. □

3.B.4.2 Proof of Lemma 3.B.4

The second inequality in (3.32) follows directly from

$$N_{1+k}(\tilde{T}) \geq N_1(T)\mathbf{1}_{T<\tilde{T}} + M\mathbf{1}_{T\geq\tilde{T}} \geq N_1(T) \wedge M.$$

It remains to show the first inequality. Consider the transposition $\tau = (1, k)$ which switches the labels 1 and k while keeping the other nodes unchanged. Let μ^τ be the distribution of $(A_{\tau(a),\tau(b)})_{ab}$.

Let M be a positive integer and set

$$\mathcal{Z} = \frac{N_{1V_k}(\tilde{T}) + N_{kV_1}(\tilde{T})}{2M} \in [0, 1].$$

Note that V_1 and V_k are disjoint sets. By invariance to labelling, we have

$$\mathbb{E}_{\mu^\tau}\left[N_{1V_k}(\tilde{T}) + N_{kV_1}(\tilde{T})\right] = \mathbb{E}_{\mu}\left[N_{1V_1}(\tilde{T}) + N_{kV_k}(\tilde{T}) + 2N_{\{1,k\}}(\tilde{T})\right].$$

Using Claim 3.D.1 and the fact that $P_{ij} \in [1/2, 3/4]$ for all $i, j \in [n]$, we can show the following claim on the Kullback-Leibler divergence between two Bernoulli parameters. The proof is at the end of the section.

Claim 3.B.6 *$kl(P_{1b}, P_{\tau(1)\tau(b)}) \vee kl(P_{kb}, P_{\tau(k)\tau(b)}) \leq 2^7\delta^2$ for all $b \in [n] \setminus \{1, k\}$.*

Hence, Lemma 3.A.3 gives

$$\begin{aligned}
&2^7\delta^2\,\mathbb{E}_\mu\left[N_1(\tilde{T})+N_k(\tilde{T})\right]\\
&\geq kl\left(\mathbb{E}_\mu\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]/(2M),\mathbb{E}_{\mu^\tau}\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]/(2M)\right)\\
&=kl\left(\frac{\mathbb{E}_\mu\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]}{2M},\frac{\mathbb{E}_\mu\left[N_{1V_1}(\tilde{T})+N_{kV_k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]}{2M}\right)\\
&\geq\frac{1}{2M}\frac{\left(\mathbb{E}_\mu\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]-\mathbb{E}_\mu\left[N_{1V_1}(\tilde{T})+N_{kV_k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]\right)^2}{\mathbb{E}_\mu\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]\vee\mathbb{E}_\mu\left[N_{1V_1}(\tilde{T})+N_{kV_k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]},
\end{aligned}$$

where the last line follows from Claim 3.D.1. With the definition $N^{good}_{1+k}(T)=N_{1V_1}(T)+N_{kV_k}(T)$, the last inequality can be written as

$$\begin{aligned}
2^8M\delta^2\,\mathbb{E}\left[N_{1+k}(\tilde{T})\right]&\left(\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]\vee\mathbb{E}\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]\right)\\
&\geq\left(\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]-\mathbb{E}\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]\right)^2. \quad (3.34)
\end{aligned}$$

If $\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]\leq\mathbb{E}\left[N^{bad}_{1+k}(\tilde{T})\right]$, then

$$2\,\mathbb{E}\left[N^{bad}_{1+k}(\tilde{T})\right]\geq\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]+\mathbb{E}\left[N^{bad}_{1+k}(\tilde{T})\right]=\mathbb{E}\left[N_{1+k}(\tilde{T})\right]$$

and Lemma 3.B.4 follows.

Assume therefore that $\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]\geq\mathbb{E}\left[N^{bad}_{1+k}(\tilde{T})\right]$. It follows that

$$2\,\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]\geq\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]+\mathbb{E}\left[N^{bad}_{1+k}(\tilde{T})\right]=\mathbb{E}\left[N_{1+k}(\tilde{T})\right].$$

Besides, $\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]\geq\mathbb{E}\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]$, so inequality (3.34) entails

$$\begin{aligned}
&2^9M\delta^2\,\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]^2\\
&\qquad\geq\left(\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]-\mathbb{E}\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]\right)^2.
\end{aligned}$$

Rearranging the expression gives

$$\begin{aligned}
\mathbb{E}\left[N^{bad}_{1+k}(\tilde{T})\right]&\geq\mathbb{E}\left[N_{1V_k}(\tilde{T})+N_{kV_1}(\tilde{T})\right]\\
&\geq\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})+2N_{\{1,k\}}(\tilde{T})\right]\left(1-\sqrt{2^9M\delta^2}\right)\\
&\geq\frac{1}{2}\,\mathbb{E}\left[N^{good}_{1+k}(\tilde{T})\right]\geq\frac{1}{4}\,\mathbb{E}\left[N_{1+k}(\tilde{T})\right],
\end{aligned}$$

where we use $2^9M\delta^2\leq 1/4$, which holds because $M\delta^2\leq 2^{-11}$ by assumption. The proof of Lemma 3.B.4 is complete. □

Proof of Claim 3.B.6. Among the $x_1, \dots, x_n$, let x_k denote a point in between 3δ and 4δ-away from x_1, that is, satisfying the inequalities $3\delta \leq d(x_1, x_k) \leq 4\delta$. The permutation $\tau = (1, k)$ exchanges 1 and k and leaves the other indices invariant, which implies that $kl(P_{1b}, P_{\tau(1)\tau(b)}) \vee kl(P_{kb}, P_{\tau(k)\tau(b)})$ is equal to $kl(P_{1b}, P_{kb}) \vee kl(P_{kb}, P_{1b})$ for all $b \in [n] \setminus \{1, k\}$.

Claim 3.D.1 ensures that, for any $p, q \in]0, 1[$,

$$kl(p, q) \vee kl(q, p) \leq \frac{(p-q)^2}{p(1-p) \wedge q(1-q)}. \tag{3.35}$$

Hence

$$kl(P_{1b}, P_{kb}) \vee kl(P_{kb}, P_{1b}) \leq \frac{(P_{1b} - P_{kb})^2}{P_{1b}(1-P_{1b}) \wedge P_{kb}(1-P_{kb})},$$

for all $b \in [n] \setminus \{1, k\}$.

By definition of the geometric model, we have $|P_{1b} - P_{kb}| = |d(x_1, x_b) - d(x_k, x_b)| \leq d(x_1, x_k) \leq 4\delta$. Besides, for all $i, j \in [n]$, we know that $P_{ij} \in [1/2, 3/4]$, which implies that $P_{ij}(1 - P_{ij}) \geq 3/16$. Therefore,

$$kl(P_{1b}, P_{kb}) \vee kl(P_{kb}, P_{1b}) \leq \frac{(4\delta)^2}{3/16} \leq 2^7 \delta^2.$$

□

3.C Conjectural upper bound in geometric graph

This section collects some arguments to support Conjecture 3, which may take the form of detailed discussions or computations. The goal is to present a sketch of a pair-matching algorithm, together with an informal study of its dispersion.

We think that this work could lead to a rigorous proof of Conjecture 3, if the latent points localization property (3.18) were proven. At this point, we do not have a proof for (3.18), and accordingly, we will assume in this section the existence of an algorithm fulfilling (3.18). We will use it as a black-box and call it Localization-Algorithm. Note that the question of latent points localization is actually investigated in the next chapter, but our first results only holds under strict assumptions, which do not fit in the pair-matching algorithm presented here. Without going into much details, the problem boils down to the following fact: our latent points localization theorems in the next chapter are useful for latent points that are uniformly spread on the latent space, whereas our pair-matching algorithm below, uses Localization-Algorithm on subsets of latent points that are a bit different from the uniform distribution.

The pair-matching algorithm introduced in this section is an iterative procedure, which alternates between a task of latent points localization and a task of expansion many times. The task of latent points localization consists in finding the localisation of a group of latent points, using only the data A. Then, this set of estimated positions, say $\hat{X}$, is used as a reference set in the task of expansion, which recovers the positions of a new set of latent points, using S and the data A. Each iteration of the pair-matching algorithm is based on this process. As the number of iteration increases, there are more and more recovered latent positions, with a smaller and smaller error of reconstruction.

The rest of the section is organized as follows. In Section C.1 we present the function Posit(S, K) as a way of estimating latent positions from data, while controlling the dispersion. Section C.2 briefly describes the function UnimodalBandit(S, X, S', δ) which estimates the positions of a new set S', based on known positions X of a reference set S. Finally, these two functions are used for the construction of a pair-matching algorithm in Section C.3, where we also analyze its dispersion.

Warning: often we only focus on the orders of magnitude, and quantities may be written up to numerical constants.

C.1. latent points localization: recovering positions from data

Let S be a sub-set of the nodes $V = \{1, \dots, n\}$, with cardinal number $|S| = N$, and latent points $x_1, \dots, x_N$. Assume that the $x_1, \dots, x_N$ belong to an interval I of length $|I| \leq \delta$ in the tore $[0, 1]$. If we sample all pairs in $S \times S$, then the conjecture (3.18) ensures that Localization-Algorithm returns a vector of estimates $\hat{X} = (\hat{x}_1, \dots, \hat{x}_N)$ satisfying the following error of positioning

$$\max_{i \in [N]} d(\widehat{Q} x_i, \hat{x}_i) \lesssim \sqrt{\frac{\log(N)}{N}} \tag{3.36}$$

with high probability, for some orthogonal transformation $\widehat{Q}$ in the tore (i.e. a transformation preserving the distances). The cost of this operation (in terms of dispersion) is bounded by

$N^2\delta$, since the total number of sampled pairs is smaller than N^2, and the cost of sampling one pair is smaller than δ (because all latent points are at most δ-away from each other in $[0,1]$).

In order to reduce this cost, we suggest the following algorithm.

Function POSIT(S, K)

Inputs: a subset S of V, and an integer K.

1. Partition at random S into K subsets $S_1, \dots, S_K$ of same size.
2. For $i = 1, \dots, K$: run LOCALIZATION-ALGORITHM on S_i, and output a vector of positions $\hat{X}_i$.
3. [a]Glue together the K vectors $\hat{X}_1 \dots, \hat{X}_K$ into a single vector $\hat{X}$.

Output: a vector $\hat{X}$ of position estimates for the nodes of S.

[a]when transforming the K vectors of size $|S|/K$ into a vector of size $|S|$, we need to carefully handle the fact that the positions are only identifiable up to some orthogonal transformation Q.

Note that for each call to LOCALIZATION-ALGORITHM on S_i, the property (3.36) ensures that the error of positioning the nodes of S_i is smaller than $\sqrt{\log(N_i)/N_i}$, for $|S_i| = N_i$. Then, the point 3 in the function POSIT(S, K) glues the position estimates together, so that the error of positioning all nodes of $S = \cup_i S_i$ is smaller than $\sqrt{\log(N_1)/N_1}$. The "error of positioning the nodes of S" means the sup-error made by the vector of estimates $\hat{X}$.

Each call to LOCALIZATION-ALGORITHM on S_i yields a dispersion smaller than $N_i^2\delta$. Hence, over the K calls to LOCALIZATION-ALGORITHM, the function POSIT(S,K) makes a total dispersion smaller than

$$D(\text{POSIT}(S, K)) \lesssim KN_1^2\delta = \frac{N^2\delta}{K}.$$

Comparing the costs between POSIT(S, K) and the initial approach, we can see that POSIT(S, K) has a smaller cost since $N^2\delta/K \leq N^2\delta$. Consequently, we will use the function POSIT(S, K) when we need to estimate the latent positions from data, while controlling the dispersion.

C.2. Bandit: recovering latent positions using reference positions

Given a set $S \subset V$ where the positions are known (or well estimated), our aim is to recover the latent positions of a new set $S' \subset V \setminus \{S\}$, using the positions of S as reference. More precisely, if the error of positioning in S is bounded by δ, then the error for S' will also be smaller than δ. The interest of this step is to recover a bigger set S' from a smaller set S, at a very small cost (in terms of dispersion). In comparison, the above step of latent points localization yields a larger cost, but allows to estimate latent positions from scratch (without reference positions).

We assume the existence of such an algorithm, hereafter called UNIMODALBANDIT(S, X, S', δ), which takes as inputs a set S of nodes with given positions X, and a upper bound δ on the positioning error of X, and a new set $S' \subset V \setminus \{S\}$.

We now give a short description of UnimodalBandit(S, X, S', δ). In order to estimate the position of a node a in S', with an error at most δ, and a probability larger than $1 - T^{-2}$, the function UnimodalBandit(S, X, S', δ) samples $\delta^{-2}\log(T)$ pairs between a and the set S. This task can be done in such a way that the dispersion is smaller than $\delta^{-1}\log(T)$. UnimodalBandit(S, X, S', δ) repeats this task on each node of S' and then returns a vector of positions X' which satisfies a sup-error smaller than δ. During this process, we can see that the total number of sampled pairs is $|S'|\delta^{-2}\log(T)$, and the total dispersion is smaller than

$$D(\text{UnimodalBandit}(S, X, S', \delta)) \lesssim |S'|\delta^{-1}\log(T) \ .$$

We emphasize that a crucial condition for using UnimodalBandit(S, X, S', δ) with the above theoretical guarantees, is that the reference set S has at least $\delta^{-2}\log(T)$ nodes per interval of length δ in $[0, 1]$.

We do not prove these theoretical guarantees. However, we think that it can be done by adapting proofs from the bandit literature which deal with unimodal structure of reward, for instance, see [Yu and Mannor, 2011]. It also seems possible to adapt the strategy from the pair-matching problem [Giraud et al., 2019, Unconstrained Algorithm "Step 2: expanding the communities"].

C.3. Pair-matching

We are now ready to give a more formal description of our pair-matching algorithm. The procedure is written below. It calls the functions Posit(S, K) and UnimodalBandit(S, X, S', δ) together at each iteration (except at the initialization). The function Posit(S, K) refines previous estimations by giving better estimates of positions, whereas UnimodalBandit(S, X, S', δ) estimates the positions of new points for a small dispersion.

Pair-Matching algorithm

Inputs: T time horizon, $V = \{1, \dots, n\}$ set of nodes.

Internal constants: $s_i = 1 - (1/2)^i$ and $\delta_i \asymp T^{-(4+s_i)/25}$ for $i \geq 0$. Let $J_0 \asymp \delta_0^{-1}$, $J_i \asymp \delta_{i-1}^{-1}$ and $K_i \asymp T^{2(1-s_{i-1})/25}$ for $i \geq 1$.

Step 0: initial positioning

1. Pick uniformly at random a set S_0 of $N_0 = \delta_0^{-3} \log(T)$ nodes in V.
2. Set $\hat{X}_0 = \textsc{Posit}(S_0, J_0)$.

Steps i ≥ 1: iteratively bigger and better sets of positions

Set $I \asymp \log\log(T)$. For $i = 1, \dots, I$, repeat:

Bigger

3. Pick uniformly at random a set S_i of $N_i = T^{3/5} \log(T)$ nodes in $V \setminus \{\cup_{k=0}^{i-1} S_k\}$.
4. Run $\textsc{BanditUnimodal}(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})$, and output positions $\hat{X}_i$ of S_i.

Better

5. According to the positions $\hat{X}_i$, decompose S_i into J_i subsets $S_{i,1}, \dots, S_{i,J_i}$.
6. For $j = 1, \dots, J_i$, set $\hat{X}'_{i,j} = \textsc{Posit}(S_{i,j}, K_i)$.
7. Glue together the J_i vectors $\hat{X}'_{i,1}, \dots, \hat{X}'_{i,J_i}$ into a single vector $\hat{X}'_i$ of positions for S_i.
8. Set $\hat{X}_i = \hat{X}'_i$.

We analyse the dispersion of the above pair-matching algorithm.

Step 0: initial positioning

Pick uniformly at random a set S_0 of $N_0 = \delta_0^{-3} \log(T)$ nodes. Then, $\textsc{Posit}(S_0, J_0)$ partition S_0 into J_0 subsets $S_{0,j}$ of same size, that is, their cardinal numbers satisfy $N_0 = N_{0,1} + \ldots + N_{0,J_0}$ with

$$N_{0,j} = \frac{N_0}{J_0} = \delta_0^{-2} \log(T).$$

Then, it calls Localization-Algorithm on each set $S_{0,j}$. The latent points localization assumption (3.36) ensures that we recover up to an error $\sqrt{\log(N_{0,j})/N_{0,j}}$ the latent positions in each of these sets. This error is smaller than δ_0 since

$$\delta_0^2 \geq \frac{\log(T)}{\delta_0^{-2} \log(T)} = \frac{\log(N_{0,j})}{N_{0,j}}.$$

$\textsc{Posit}(S_0, J_0)$ finally glues together these J_0 sets, and returns position estimates $\hat{X}_0$ for

the whole set S_0. The total dispersion for the Step 0 is equal to the number of sampled pairs:

$$D(\text{Step } 0) = \sum_{j=1}^{J_0} N_{0,j}^2 \leq J_0 N_{0,1}^2 = \delta_0^{-5} \log(T)^2 = T^{4/5} \log(T)^2. \tag{3.37}$$

Step i with $i \geq 1$: recovering a bigger and better $(i+1)^{\text{th}}$-vectors of positions

a/*Bigger: positioning a set S_i at resolution δ_{i-1}, using position estimates of same resolution of the smaller set S_{i-1}*

Pick uniformly at random a set S_i of $N_i = T^{3/5} \log(T)$ nodes in $V \setminus \{\cup_{k=0}^{i-1} S_k\}$. There are enough nodes in the graph for this operation since we have assumed to be in the regime where $n \gtrsim T^{3/5} \log(T) \log\log(T)$.

Then, BanditUnimodal$(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})$ returns a vector of positions $\hat{X}_i$ for the nodes of S_i, with a positioning error smaller than δ_{i-1}.

Recall that for each node a of S_i, the number of sampled pairs between a and the set S_{i-1} is $\delta_{i-1}^{-2} \log(T)$. Also recall that the condition for using BanditUnimodal$(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})$ is that the reference set S_{i-1} has enough nodes for this task, i.e. it has at least $\delta_{i-1}^{-2} \log(T)$ nodes per interval of length δ_{i-1} in $[0,1]$. Here, this is translated into the following condition $N_{i-1}\, \delta_{i-1} \geq \delta_{i-1}^{-2} \log(T)$ which is easy to check.

During BanditUnimodal$(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})$, the total number of sampled pairs is $N_i \delta_{i-1}^{-2} \log(T)$, and the total dispersion is bounded by

$$D(\text{BanditUnimodal}(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})) \leq N_i \delta_{i-1}^{-1} \log(T) \leq T^{4/5} \log(T). \tag{3.38}$$

b/*Better: improving from δ_{i-1} to δ_i the localization of the nodes of S_i :*

Decompose the set S_i into J_i subsets according to their localisation in the latent space $[0,1]$, that is, their cardinal numbers satisfy $N_i = N_{i,1} + \ldots + N_{i,J_i}$ where each $N_{i,j}$ is the number of nodes in the j^{th}-interval of length $J_i^{-1} = \delta_{i-1}$ in $[0,1]$. We have

$$N_{i,j} = \frac{N_i}{J_i} = \delta_{i-1} N_i = T^{(11-s_{i-1})/25} \log(T).$$

For the positioning of each set $S_{i,j}$, the function Posit$(S_{i,j}, K_i)$ decomposes the set $S_{i,j}$ into K_i subsets $S_{i,j,k}$, $k \in [K_i]$. Their cardinal numbers satisfy $N_{i,j} = N_{i,j,1} + \ldots + N_{i,j,K_i}$, with

$$N_{i,j,k} = \frac{N_{i,j}}{K_i} = T^{(9+s_{i-1})/25} \log(T). \tag{3.39}$$

Then, Posit$(S_{i,j}, K_i)$ calls Localization-Algorithm on each $N_{i,j,k}$, so that we recover the latent positions in each of these sets, up to an error δ_i. Indeed, this guarantee follows from assumption (3.36) and the inequality

$$\delta_i^2 \geq \frac{\log(T)}{\delta_i^{-2} \log(T)} \geq \frac{\log(N_{i,j,k})}{N_{i,j,k}},$$

where we use $N_{i,j,k} \geq \delta_i^{-2} \log(T)$.

For a call LOCALIZATION-ALGORITHM on $N_{i,j,k}$, we have a dispersion smaller than $\delta_{i-1} N_{i,j,k}^2$ since all nodes of $S_{i,j,k}$ have their latent points in a same interval of length smaller than δ_{i-1}. Hence, for the function POSIT$(S_{i,j}, K_i)$, which calls LOCALIZATION-ALGORITHM K_i times, the total dispersion is smaller than

$$D(\text{POSIT}(S_{i,j}, K_i)) \leq \sum_{j=1}^{K_i} \delta_{i-1} N_{i,j,k}^2 = K_i \delta_{i-1} N_{i,j,1}^2 = \delta_{i-1} T^{4/5} \log(T)^2,$$

using (3.39) in the last equality.

Finally, the J_i calls to POSIT$(S_{i,j}, K_i)$, $j \in [J_i]$, allows to position the set S_i, with a dispersion bounded by

$$\sum_{j=1}^{J_i} D(\text{POSIT}(S_{i,j}, K_i)) \leq J_i \delta_{i-1} T^{4/5} \log(T)^2 = T^{4/5} \log(T)^2.$$

The last display and (3.38) allows to conclude that the dispersion of the i^{th}-step is smaller than

$$\begin{aligned} D(\text{Step i}) =& D(\text{BANDITUNIMODAL}(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})) + \sum_{j=1}^{J_i} D(\text{POSIT}(S_{i,j}, K_i)) \\ &\leq T^{4/5} \log(T)^2. \end{aligned} \tag{3.40}$$

Total dispersion

The total dispersion of the pair-matching algorithm is equal to the sum of the dispersions at each iteration i, that is

$$D(\text{Pair-matching algorithm}) = \sum_{i=0}^{I} D(\text{Step i}) \leq I\, T^{4/5} \log(T)^2, \tag{3.41}$$

invoking the bounds (3.37) and (3.40).

It remains to compute the number of iterations I until the T pairs have been sampled overall. At each step $i \geq 1$, BANDITUNIMODAL$(S_{i-1}, \hat{X}_{i-1}, S_i, \delta_{i-1})$ samples $N_i \delta_{i-1}^{-2}$ pairs, and the J_i calls to POSIT$(S_{i,j}, K_i)$ samples together $J_i K_i N_{i,j,1}^2$ pairs. Using (3.39), this gives a total number of sampled pairs during the i^{th}-step equal to

$$T^{3/5} \log(T) T^{2(4+s_{i-1})/25} + T^{(4+s_{i-1})/25} T^{2(1-s_{i-1})/25} T^{2(9+s_{i-1})/25} \log(T)$$

which is of the order of

$$T^{(24+s_{i-1})/25} \log(T)$$

since $s_i = 1 - (1/2)^i$ is smaller than 1. Note that the number of sampled pairs at step 0 is much smaller. Hence, the T pairs are all sampled in I iterations if the following equality holds

$$\sum_{i=1}^{I} T^{(24+s_{i-1})/25} \log(T) = T,$$

that is,

$$\sum_{i=1}^{I} T^{-(1/2)^{i-1}/25} \log(T) = 1,$$

which is satisfied for a number of iteration I of the order of

$$I \asymp \log\log(T).$$

Combining the above inequality with (3.41), we get the following bound on the total dispersion

$$D(\text{Pair-matching algorithm}) \leq T^{4/5} \log(T)^2 \log\log(T).$$

Note that the algorithm picks $N_i = T^{3/5}\log(T)$ new nodes at each step $i \geq 1$ for a number of iterations $I \asymp \log\log(T)$. The total number $T^{3/5}\log(T)\log\log(T)$ of nodes picked by the algorithm must be smaller than the number of nodes available in the graph, i.e.:

$$n \geq T^{3/5}\log(T)\log\log(T),$$

which is the regime assumed in Conjecture 3. □

3.D Appendix: Probabilistic inequalities

We also recall some classical controls on the Kullback-Leibler divergence between two Bernoulli distribution.

Claim 3.D.1 *For any $p_1, p_2 \in (0,1)$,*

$$\frac{(p_1-p_2)^2}{p_1 \vee p_2} \leq kl(p_1,p_2) \leq \frac{(p_1-p_2)^2}{p_1(1-p_1) \wedge p_2(1-p_2)}.$$

In particular, for any $0 < q \leq p \leq 1-c$ and $p/q \leq \rho^*$ for constants $c \in [1/2, 1)$ and $\rho^* > 1$,

$$\frac{1}{\rho^*} s \leq \frac{(p-q)^2}{p} \leq kl(p,q) \vee kl(q,p) \leq \frac{(p-q)^2}{qc} \leq \frac{1+\rho^*}{2c} s.$$

Chapter 4

Optimal embedding of interaction data: applications to random geometric graphs and statistical seriation

Motivated by the conjectures from the previous chapter about pair-matching, we study the problem of positioning a set of items in a one-dimensional space, using noisy observations of pairwise affinities. The underlying assumption is that items with higher affinity should be closer in the latent space. Such a task is an instance of the seriation problem which seeks to recover a hidden order from unsorted information. Here we address this problem from a statistical point of view and give upper and matching lower bounds on the reconstruction error of the positions. Crucially, this performance is measured in uniform-norm, i.e. corresponds to the worst error over all the estimated positions. In the particular case where the affinities are determined by the distances in the latent space, we study a computationally efficient alternative which exhibits the optimal estimation rate.

In the previous chapter, we have sketched a road-map towards optimal regret for pair-matching problems in geometric random graphs. The proposed strategy relies on the conjectural existence of a procedure recovering the positions of the latent points with a guarantee on its uniform error, see eq.(3.18) in Section 3.3.3. This chapter is dedicated to the latter problem of uniform latent points recovery in latent affinity model that generalize the random geometric graph considered in the previous chapter.

Contents

4.1 Introduction

4.1.1 Pairwise affinity model

Many real-life data arise in the form of pairwise measurements $\{A_{ij}\}_{1\leq i<j\leq n}$, where $A_{ij} \in \mathbb{R}$ is the outcome of some interaction between i and j. For instance, in online video games, A_{ij} may be the number of parties between the i^{th} and j^{th} players. In criminal organizations, A_{ij} may be the number of phone calls between two potential suspects i and j. Interactions data also include the important case of networks where A_{ij} encodes the existence of an edge between i and j, which arises in various fields such as physics, biology or social sciences.

From a data scientist perspective, one usually aims at finding informative visualization of the data. This allows for instance to easily identify groups of users with high-affinity. In that respect, latent space models have been proved especially useful [Hoff et al., 2002]. Given some metric space (X, d), such models amount to assuming the existence of an affinity function $f : X \times X \mapsto \mathbb{R}$ and unobserved positions $(x_1^*, \dots, x_n^*)$ such that the expected interaction between i and j is the affinity between x_i^* and x_j^*, that is $\mathbb{E}[A_{ij}] = f(x_i^*, x_j^*)$. The affinity $f(x_i^*, x_j^*)$ is typically assumed to decrease as the metric distance $d(x_i^*, x_j^*)$ increases. In particular, close points x_i^* and x_j^* share a high affinity whereas distant points share a small affinity. This latent space formulation encompasses many models as exemplified in the next paragraphs.

Example 1: Random Geometric Graph. Consider the unit circle $\mathcal{C}$ in $\mathbb{R}^2$ endowed with the geodesic distance d and a non-increasing function $\tilde{f} : [0, \pi] \mapsto [0, 1]$. Given latent positions $(x_1^*, \dots, x_n^*)$, the corresponding size n random geometric graph is defined as follows: for any $1 \leq i < j \leq j$, the edge are sampled independently with $\mathbb{P}[A_{ij} = 1] = \tilde{f}(d(x_i^*, x_j^*))$.

Example 2: Stochastic Block Models. For a positive integer K, we take $X = [K]$ and consider the function $f(x, y) = a\mathbf{1}_{x=y} + b\mathbf{1}_{x\neq y}$ where $0 \leq b < a \leq 1$. For $(x_1, \dots, x_n)$, an edge is sampled between i and j with probability a if $x_i^* = x_j^*$ and with probability b if $x_i^* \neq x_j^*$, that is $\mathbb{P}[A_{ij} = 1] = f(x_i^*, x_j^*)$. This is a specific instance, sometimes called the affinity model, of the stochastic block model (SBM) [Holland et al., 1983]. Alternatively, we can define the same distribution on A taking the latent space $\mathcal{C}$, partitioning it into K connected subsets $V_1, \dots, V_K$, and setting $f(x, y) = b + (a - b) \sum_{j=1}^{K} \mathbf{1}_{x\in V_j} \mathbf{1}_{y\in V_j}$.

Example 3: R-matrices and Statistical Seriation. A symmetric matrix $B \in \mathbb{R}^{n\times n}$ is called a Robinson matrix (henceforth R-matrix) if, for any $1 \leq j \leq i \leq n$, we have $B_{i,j} \geq B_{i+1,j}$ and $B_{i,j} \geq B_{i,j-1}$. In other words, the entries of B are decreasing when one goes away from its diagonal. Let Σ_n denote the permutation group. For a matrix $F = (F_{ij})$ and a permutation $\sigma \in \Sigma_n$, we write $F_\sigma = (F_{\sigma(i),\sigma(j)})$. A matrix F is called a pre-R matrix if there exists a permutation $\sigma \in \Sigma_n$ such that F_σ is a R-matrix. Given a noisy observation A of a pre-R matrix F, the noisy seriation problem [Fogel et al., 2013] amounts to finding a permutation σ^* such that F_{σ^*} is an R-matrix. Many applications involve recovering a latent order from similarity information, such as genomic sequencing [Garriga et al., 2011], identifying interval graphs [Fulkerson and Gross, 1965], and envelope reduction for sparse matrices [Barnard et al., 1995]. This problem can be recast in the latent space terminology using $X = [n]$, $x_i^* = \sigma^*(i)$, and the affinity function $f(x_i^*, x_j^*) = F_{\sigma^*(i),\sigma^*(j)}$. Since F_{σ^*} is a R-matrix, this enforces that $f(x, y)$ is decreasing with the distance $|x - y|$.

Example 4: Toroidal R-matrices. Consider the set $[n]$ as a torus and the corresponding distance $d(i, j) = \min(|j - i|, |n - i + j|)$ for any $1 \leq i \leq j \leq n$. Then, a symmetric matrix B is a toroidal R-matrix if $B_{i,j} \geq B_{i+1,j}$ when $d(i, j) < d(i + 1, j)$ and $B_{i,j} \geq B_{i,j+1}$ when $d(i, j) < d(i, j + 1)$. In other words, the entries of B decrease as one moves away from the diagonal with respect to the toroidal distance. As previously, a pre-toroidal R-matrix is defined as a permutation of a toroidal R matrix and the statistical seriation model is defined analogously [Recanati et al., 2018]. Again, we can recast this model as a latent model on the space $X = [n]$ endowed with the toroidal distance. Alternatively, we can also rewrite it as a latent space model on the size n grid $\mathcal{C}_n$ of the unit circle $\mathcal{C}$ corresponding to the n-th unit roots.

4.1.2 Localization problem and our contribution

Given these interaction data $(A_{ij})_{1\leq i<j\leq n}$ that are modelled according to a latent space model, that is $\mathbb{E}[A_{ij}] = f(x_i^*, x_j^*)$, we are interested in building an estimator $(\hat{x}_1, \dots, \hat{x}_n)$ of the latent positions $(x_1^*, \dots, x_n^*)$. Throughout this chapter, we will restrict our attention to the latent space $\mathcal{C}$ (unit circle) endowed with the geodesic distance d. Note that this framework encompasses the examples of random geometric graph and the toroidal R matrices. Our main assumption is that f is (almost) bi-Lipschitz with respect to d. The formal definition

is written in section 4.2.1. In particular, this shape constraint excludes the SBM formulation of example 2.

Importantly, we will assume that we are given two independent copies $A^{(1)}$ and $A^{(2)}$ instead of a single observation A. The data will thus consist into matrices $A^{(1)}$ and $A^{(2)}$ with the same underlying structure $\mathbb{E}[A^{(1)}_{ij}] = \mathbb{E}[A^{(2)}_{ij}] = f(x^*_i, x^*_j)$. The purpose of this assumption is to greatly simplify the presentation of the chapter, while retaining the substantive elements of our study.

Motivated by a conjecture in the previous chapter, our objective is to recover the latent points with a guarantee in uniform distance, that is, with respect to the metric d_∞ defined as $d_\infty(\mathbf{x}, \mathbf{x}') = \max_i d(x_i, x'_i)$, for all vectors $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{x}' = (x'_1, \dots, x'_n)$ in $\mathcal{C}^n$. Unfortunately, the latent positions $\mathbf{x}^* = (x^*_1, \dots, x^*_n)$ are not identifiable from the data. In fact, they are not even identifiable in the simple situation where f is known and equal to the affine function $f_0(x, x') = 1 - d(x, x')/(2\pi)$. In this example, a simple remedy to the non-identifiability is the following pseudo-metric:

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) = \min_{Q \in \mathcal{O}} \max_{i \in [n]} \quad d\big(\hat{x}_i, Qx^*_i\big),$$

which measures the performance of an estimator up to transformations in the orthogonal group $\mathcal{O}$ of $\mathbb{R}^2$. However, since f is unknown in the setting of interest, the situation is more involved and other identifiability issues come into play. The formal discussion about identifiability is in section 4.2.2.

The core of the algorithm is a refinement scheme for latent points localization. Our estimation procedure takes a two-step approach: it first computes a good enough estimator $\hat{\mathbf{x}}'$ of $\mathbf{x}^*$ and then builds on this preliminary estimator to locally improve the position estimates, which gives a new estimator $\hat{\mathbf{x}}$. Thanks to the two independent samples $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ at our disposal, the analysis of both steps can be handled independently. For simplicity, we first present our result in a particular case where the vector of positions $\mathbf{x}^*$ is balanced on $\mathcal{C}$ (this will be formalized later). For the first step, we prove that the preliminary estimator which is based on $\mathbf{A}^{(1)}$ satisfies the following d_1-bound with high probability

$$d_{1,\mathcal{O}}(\hat{\mathbf{x}}', \mathbf{x}^*) := \min_{Q \in \mathcal{O}} \sum_i d\big(\hat{x}'_i, Qx^*_i\big) \le c\sqrt{n \log(n)} \quad ,$$

for some numerical constant $c > 0$. Then, the refinement step which relies on $\mathbf{A}^{(2)}$ and $\hat{\mathbf{x}}'$ allows to get a uniform localization with high probability

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \le c' \sqrt{\frac{\log(n)}{n}} \quad .$$

In the following, a vector $\mathbf{y} = (y_1, y_2, \dots, y_n)$ is said to be balanced if it is of the form $(e^{\iota 2\pi\sigma(j)/n})_{1 \le j \le n}$, where σ is a permutation in Σ_n. The $\sqrt{\log(n)/n}$ uniform rate actually holds for any latent positions $\mathbf{x}$ that can be well approximated by a balanced vector. Conversely, we prove a matching lower bound for this rate, thus implying the optimality of our procedure on this class of positions.

In the more general case of arbitrary latent positions, we establish error bounds involving the above rate and a bias-type term. This bias grows linearly with the smallest d_∞-difference

between the latent positions and the set of balanced vectors. This shows that robustness is possible even in uniform norm. To the best of our knowledge, it is the first statistical guarantees in uniform norm for latent points localization.

While the above analysis relies on two independent samples, we sketch a road-map towards a one-sample analysis in section 4.5. This mainly consists of splitting the data A in two independent samples and then replicating the same scheme as above.

In terms of time complexity, the second step of our procedure is polynomial in n, and it can be parallelizable in n linear tasks. In contrast, our preliminary estimator has an exponential-time complexity, which motivates us to consider a spectral alternative for this first step. Building upon the work of [Recanati et al., 2018], we can analyse the spectral procedure in the particular setting of geometric functions, i.e. when $f(x, x') = \tilde{f}(d(x, x'))$. Besides, if $\tilde{f}$ is an affine function, then we recover the same optimal rate $\sqrt{\log(n)/n}$ as in the exponential-time procedure.

4.1.3 Related work

Seriation Given a pre-R matrix F, the seriation problem seeks to find the latent order σ^* such that F_{σ^*} is an R-matrix. For this noiseless version of example 3, efficient algorithms have been proposed using convex optimization [Fogel et al., 2013], or spectral methods [Atkins et al., 1998]. Closer to our setting, the seriation problem has been solved on toroidal R-matrices in the noiseless case [Recanati et al., 2018], by using a spectral algorithm. However, there is still little information about noise robustness and fundamental limits of estimation. Our work partially covers this gap under some regularity assumptions on F_{σ^*}.

Latent points estimation in random geometric graphs [Diaz et al., 2020] considers the problem of estimating latent positions $x_1^*, \dots, x_n^*$ in a square of $\mathbb{R}^2$. The authors assume that the link function f is known and takes the specific form: $f(x, y) = 1$ if $\|x - y\| \leq r$, and $f(x, y) = 0$ otherwise. In our work, f is unknown and belongs to a class of (almost) bi-Lipschitz functions, and the observations A_{ij} are random variables (conditionally to the x_i). Other related work includes [Sussman et al., 2013], which establishes that the latent positions in a random dot-product graph can be consistently estimated. In the dot product model, the latent points x_i are vectors in the unit ball of $\mathbb{R}^d$, and the link function is defined as the dot product $f(x, y) = \langle x, y \rangle$ between vectors $x, y \in \mathbb{R}^d$. While the authors show non-asymptotic bounds on the reconstruction error of $\mathbf{x}^*$ in l_2 norm, our goal is to have non-asymptotic bounds in uniform norm (i.e. with respect to $d_{\infty,\mathcal{O}}$). However, their setting is different from ours, and the results are not comparable.

Two-step method Statistical optimality is sometimes proved using exhaustive search over the parameter space, whereas such combinatorial optimizations are computationally intractable [Zhang et al., 2016]. The global to local scheme allows to dramatically reduce this computational complexity, by building two-step procedures [Gao et al., 2017, Zhang et al., 2016]. The idea is to use an initial estimator that satisfies a certain (weak) consistency condition, and then use a refinement step to obtain an improved estimator that achieves the statistical optimal performance. A popular approach for the initialization step is to use

spectral methods, and then improve the performance with splitting techniques such as [Lei and Zhu, 2014, Chen and Lei, 2018].

4.1.4 Notation and organization of the chapter

In the sequel, $c, c', c'', c''' > 0$ denote numerical constants that may change from line to line. For two functions or sequences x and y, we write $x \lesssim y$ (resp. $x \gtrsim y$) if, for some numerical constant $c > 0$, we have $x \leq cy$ (resp. $x \geq cy$). Denote $x \vee y$ and $x \wedge y$ the maximum (resp. minimum) of x and y. Given any $x > 0$, we denote its integer part by $\lfloor x \rfloor$, and the set of integers from 1 to $\lfloor x \rfloor$ by $[x]$. Given a matrix $F = (f_{ij})$ and $q \geq 1$ we write $\|F\|_q$ its entry-wise l_q norm, that is $\|F\|_q = (\sum_{ij} |f_{ij}|^q)^{1/q}$. Besides, the entry-wise inner product between two matrices F and G is denoted by $\langle F, G \rangle$. The i^{th} row is denoted by F_i.

The collection of permutations of $[n]$ is denoted by Σ_n. For any $\sigma \in \Sigma_n$ of $[n]$ and any vector $\mathbf{x}$ of size n, the permuted vector $(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ is written $\mathbf{x}_\sigma$.

In Section 4.2 we formalize the problem and discuss identifiability issues. The main method UTS and its analysis is studied in Section 4.3. Section 4.4 is dedicated to the two-step spectral method and its application to geometric models. Finally, we summarize our findings and discuss some extensions in Section 4.5. All the proofs are postponed to the end of the chapter.

4.2 Problem formulation and identifiability

4.2.1 Model

We recall that $\mathcal{C}$ denote the unit circle in $\mathbb{R}^2$. For $x \in \mathcal{C}$, we denote its argument by $\underline{x} \in [0, 2\pi)$, $x = e^{\iota \underline{x}}$. Besides, the geodesic distance d on $\mathcal{C}$ is given by $d(x, y) = |\underline{x} - \underline{y}| \wedge (2\pi - |\underline{x} - \underline{y}|)$. For any positive integer k, let $\mathcal{C}_k = \{1, e^{\iota 2\pi/k}, \ldots, e^{\iota 2\pi(k-1)/k}\}$ stand be the regular grid of size k on $\mathcal{C}$.

As explained in the introduction, we shall work with regular affinity functions f. Fix any constant $c_e > 0$ and $0 < c_l \leq c_L$. We define below the class $\mathcal{BL}[c_l, c_L, c_e]$ of (nearly) bi-Lipschitz functions. For short, we write henceforth $\epsilon_n = c_e\sqrt{\log(n)/n}$.

Definition 4.2.1 *The collection $\mathcal{BL}[c_l, c_L, c_e]$ is made all functions $f : \mathcal{C}^2 \to [0, 1]$ that are symmetric ($f(x, y) = f(y, x)$ for all x, y in $\mathcal{C}$) and satisfy the two following conditions for all $(x, y, y') \in \mathcal{C}$,*

$$|f(x, y) - f(x, y')| \leq c_L d(y, y') + \epsilon_n \ ; \tag{4.1}$$

$$f(x, y') - f(x, y) \geq c_l\big(d(x, y) - d(x, y')\big) - \epsilon_n \quad \text{if } d(x, y) \geq d(x, y') \ . \tag{4.2}$$

If $c_e = 0$, Condition (4.2) enforces that, for any x, $f(x, y)$ is a decreasing function of $d(x, y)$, as required for affinity functions in the introduction. In (4.1–4.2), the term ϵ_n interprets as a possible small relaxation of the usual bi-Lipschitz condition.

Let f be a function of $\mathcal{BL}[c_l, c_L, c_e]$ and let $x_1^*, \dots, x_n^*$ be latent points on the two-dimensional sphere $\mathcal{C}$. The two independent interaction matrices, denoted by $A^{(s)} = \{A_{ij}^{(s)}\}_{1 \le i < j \le n}$ for $s = 1, 2$ follow the generating process

$$A_{ij}^{(s)} = F_{ij} + E_{ij}^{(s)}, \qquad 1 \le i < j \le n \ ,$$

where $F_{ij} = f(x_i^*, x_j^*)$ and the two independent centered noise matrices $E^{(s)}$ follow a sub-Gaussian distribution. Namely, for $s = 1, 2$ and all matrices β such that $\sum_{1 \le i < j \le n} \beta_{ij}^2 = 1$, we have

$$\mathbb{P}\left[\sum_{1 \le i < j \le n} \beta_{ij} E_{ij}^{(s)} > t \right] \le e^{-t^2/2} \ , \qquad \forall t > 0 \ .$$

By convention, the diagonal of $A^{(s)}$ is equal to zero. Henceforth, the probability distribution of $A^{(s)} = F + E^{(s)}$ is denoted by $\mathbb{P}_{(\mathbf{x}^*, f)}$.

4.2.2 Identifiability issues and localization problem

Given any vectors $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{x}' = (x_1', \dots, x_n')$ in $\mathcal{C}^n$, let d_∞ be the distance defined as

$$d_\infty(\mathbf{x}, \mathbf{x}') = \max_{i \in [n]} \ d(x_i, x_i') \ . \tag{4.3}$$

As discussed in the introduction, the latent vector $\mathbf{x}^*$ is not identifiable from A with respect to (4.3), even in simple situations where the function f is known and affine with respect to the distance d. In this, we have seen that the data distribution is invariant by orthogonal transformations of the latent positions, that is $\mathbb{P}_{(\mathbf{x}^*, f)} = \mathbb{P}_{(Q\mathbf{x}^*, f)}$ for any transformations Q in the orthogonal group $\mathcal{O}$ of $\mathbb{R}^2$. Accordingly, the performance of any estimator $\hat{\mathbf{x}}$ will be measured up to transformations in $\mathcal{O}$:

$$d_{\infty, \mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) = \min_{Q \in \mathcal{O}} \ d_\infty(\hat{\mathbf{x}}, Q\mathbf{x}^*) \ . \tag{4.4}$$

However, we are interested in the more general setting where the bi-Lipschitz function is unknown. In this situation, $\mathbf{x}^*$ is not identifiable even with respect to (4.4) as exemplified in the next proposition.

Proposition 4.2.2 (Non-identifiability w.r.t. (4.4)) *Consider the simple model $f(x, y) = 1 - d(x, y)/(2\pi)$ and $x_k = e^{\iota k 2\pi/n}$ for all $k \in [n]$. Note that $f \in \mathcal{BL}[(3\pi)^{-1}, \pi^{-1}, 0]$ and $\mathbf{x} \in S_{ev}$. Let us construct another representation $(\tilde{f}, \tilde{\mathbf{x}})$ such that $\tilde{f} \in \mathcal{BL}[(3\pi)^{-1}, \pi^{-1}, 0]$ and $\tilde{\mathbf{x}} \in S_{ev}$ too, and such that $d_{\infty, \mathcal{O}}(\mathbf{x}, \tilde{\mathbf{x}}) \ge \pi/8$ and*

$$\tilde{f}(\tilde{x}_i, \tilde{x}_j) = f(x_i, x_j) \ , \tag{4.5}$$

for all $i, j \in [n]$. It then readily follows from (4.5) that $\mathbb{P}_{\mathbf{x}, f} = \mathbb{P}_{\tilde{\mathbf{x}}, \tilde{f}}$.

The proof of this result is postponed to Section 4.B. It mainly amounts to building a function $\tilde{f}$ that is dilated at some regions of $\mathcal{C}$ and contracted at some other regions of $\mathcal{C}$ so that the corresponding $\tilde{\mathbf{x}}$ are respectively contracted and dilated. Observe that the positions

$\mathbf{x}$ and $\tilde{\mathbf{x}}$ are so different that $d_{\infty,\mathcal{O}}(\mathbf{x},\tilde{\mathbf{x}}) \geq \pi/8$. Hence, for any estimator $\hat{\mathbf{x}}$, the reconstruction error $d_{\infty,\mathcal{O}}(\hat{\mathbf{x}},\mathbf{x}^*)$ cannot be smaller than $\pi/16$ with high probability. This entails that no consistent estimator exists for the estimation problem with respect to the loss function(4.4). One could think that this bad scenario is due to pathological forms of the latent vector $\mathbf{x}^*$ and so it can be avoided by adding some mild assumptions on the form of $\mathbf{x}^*$ (i.e. on the distribution of the latent points in $\mathcal{C}$). This is not the case: the two vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ of the example below are in fact evenly distributed in the latent space; more precisely, they belong to the set $S_{ev} := \{\mathbf{y} \in \mathcal{C}^n : \sup_{z\in\mathcal{C}} \min_{i\in[n]} d(y_i,z) \leq 3\pi/n\}$. So, restricting the analysis to the set of positions S_{ev} is not a solution to the identifiability issue.

Due to the identifiability issue, it is natural to work with the equivalence classes of representations. Given a data distribution $\mathbb{P}$, an equivalence class is the set of all representations $(f,\mathbf{x})$ such that $\mathbb{P}_{\mathbf{x},f} = \mathbb{P}$. In each of these equivalence classes, we then choose a particular representative to estimate. The representative of interest in our work is such that the latent vector is well approximated by a balanced vector of the form $(e^{\iota 2\pi\sigma(j)/n})_{1\leq j\leq n}$, for some permutation $\sigma \in \Sigma_n$. This situation occurs in particular when $x_1^*,\dots,x_n^*$ is a uniform sample of $\mathcal{C}$, which is a classic assumption in latent space models, for instance in SBM, graphons or geometric graphs [Klopp et al., 2017, Araya and De Castro, 2019]. Hereafter, the collection of balanced vectors $\{(e^{\iota 2\pi\sigma(j)/n})_{1\leq j\leq n}, \quad \sigma \in \Sigma_n\}$ is denoted by $\mathcal{P}_n$.

Formally, for any $f \in \mathcal{BL}[c_l,c_L,c_e]$ and $\mathbf{x}^* \in \mathcal{C}^n$, we write $[(\mathbf{x}^*,f)]_{bilip}$ the collection of representations $(\mathbf{x},\tilde{f})$ such that $\tilde{f} \in \mathcal{BL}[c_l,c_L,c_e]$, $\mathbf{x} \in \mathcal{C}^n$ and $\mathbb{P}_{(\mathbf{x},\tilde{f})} = \mathbb{P}_{(\mathbf{x}^*,f)}$. In each (bi-Lipschitz) equivalence class $[(\mathbf{x}^*,f)]_{bilip}$, we sometimes consider a specific representative as follows:

$$(\mathbf{x}_u, f_u) \in \underset{(\mathbf{x},\tilde{f})\in[(\mathbf{x}^*,f)]_{bilip}}{\arg\inf} \min_{\mathbf{y}\in\mathcal{P}_n} d_\infty(\mathbf{x},\mathbf{y}) \ . \tag{4.6}$$

In other words, $\mathbf{x}_u$ is the closest representation to a balanced vector while maintaining f_u bi-Lipschitz. Our objective is to estimate $\mathbf{x}_u$ in $d_{\infty,\mathcal{O}}$ distance.

Remark: In the special case where we assume that $\mathbf{x}^* \in \mathcal{P}_n$ (as in Example 4 in the introduction), the latent vector $\mathbf{x}^*$ is fully encoded by a permutation σ^*. Then, the problem of estimating $\mathbf{x}^*$ up to an orthogonal transformation is equivalent to the seriation problem of recovering σ^* up to a subgroup of permutations induced by the circular permutation and the reverse permutation $(1,n,n-1,n-2,\dots,2)$.

4.3 Main results

4.3.1 Uniform-Two-Steps (UTS) Algorithm

The general approach amounts to first computing a good enough estimator $\hat{\mathbf{x}}'$ of $\mathbf{x}^*$ and then building on this preliminary estimator to locally improve $\hat{\mathbf{x}}$. The combination of these two steps will allow us to get uniform localization bound.

Sketch of UTS algorithm

Input: adjacency matrices $\{A^{(1)}_{ij}\}$ and $\{A^{(2)}_{ij}\}$

Step 1: Initialization from $A^{(1)}$

Estimate the positions $x^*_1, \dots, x^*_n$; output $\hat{x}'_1, \dots, \hat{x}'_n$.

Step 2: Local improvement from $A^{(2)}$

For $i = 1, \dots, n$: estimate the position x^*_i with respect to the reference positions $\hat{x}'_1, \dots, \hat{x}'_n$; output $\hat{x}_i$.

Output: $\hat{x}_1, \dots, \hat{x}_n$.

Since we have two independent samples $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ at our disposal, this much simplifies the analysis as both steps can be handled independently. See Section 4.5 for a tentative roadmap toward a one-sample analysis.

4.3.1.1 Uniform estimation based on local improvements

We first focus on the local improvement step. In this section, we therefore assume that we are given a vector $\mathbf{x} \in \mathcal{P}_n$ of estimated positions, which is independent of $A^{(2)}$. For any $x \in \mathcal{C}$, we define the vector of distances between x and the vector $\mathbf{x}$:

$$D(x, \mathbf{x}) = \big(d(x, x_1), \dots, d(x, x_n)\big) \ .$$

Then for any $i = 1, \dots, n$, we compute $\hat{x}_i$ by

$$\hat{x}_i = \operatorname*{argmin}_{x \in \mathcal{C}_n} \langle A^{(2)}_i, D(x, \mathbf{x}) \rangle \ , \tag{4.7}$$

and write $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$. From (4.7), we see that $\hat{x}_i$ is chosen in such a way that $d(\hat{x}_i, x_j)$ should be small for large values of F_{ij}, and conversely, it should be large for small values of F_{ij}. Had the given vector $\mathbf{x}$ been equal to the true position $\mathbf{x}^*$ and the function $f(x, y)$ been proportional to $d(x, y)$ (that is, $c_l = c_L$ and $c_e = 0$ in the bi-Lipschitz conditions (4.1–4.2)), then one could readily check that the criterion (4.7) perfectly recovers the positions from noiseless observations, that is

$$\operatorname*{argmin}_{x \in \mathcal{C}_n} \langle \mathbb{E}[A^{(2)}_i], D(x, \mathbf{x}^*) \rangle = \operatorname*{argmin}_{x \in \mathcal{C}_n} \langle c_l D(x, \mathbf{x}^*), D(x, \mathbf{x}^*) \rangle = x^*_i \ .$$

The time complexity for computing $\mathbf{x}_i$ is linear in n and the algorithm is easily parallelizable. The next proposition established that the uniform error of $\hat{\mathbf{x}}$ is controlled, even though $\mathbf{x}$ is not exact, $\mathbf{A}^{(2)}$ is a noisy version of F, and the function f is not necessarily affine.

Given two vectors $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{x}' = (x'_1, \dots, x'_n)$ in $\mathcal{C}^n$, we define their d_1-distance by $d_1(\mathbf{x}, \mathbf{x}') = \sum_i d(x_i, x'_i)$, and their d_1-difference up to orthogonal transformation by

$$d_{1,\mathcal{O}}(\mathbf{x}, \mathbf{x}') = \min_{Q \in \mathcal{O}} d_1(\mathbf{x}, Q\mathbf{x}') \ .$$

For two functions a and b, we write $a \lesssim_{c_l, c_L, c_e} b$ when there exists some quantity C only depending on c_l, c_L, and c_e such that $a \leq Cb$.

Proposition 4.3.1 *Consider $f \in \mathcal{BL}[c_l, c_L, c_e]$, $\mathbf{x}^* \in \mathcal{C}^n$, and $\mathbf{x} \in \mathcal{P}_n$ independent of $A^{(2)}$. Then, conditionally to $\mathbf{x}$, the estimator $\hat{\mathbf{x}}$ defined in (4.7) satisfies*

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l,c_L,c_e} \min_{\mathbf{y}\in\mathcal{P}_n} d_\infty(\mathbf{y}, \mathbf{x}^*) + \frac{d_{1,\mathcal{O}}(\mathbf{x}, \mathbf{x}^*)}{n} + \sqrt{\frac{\log(n)}{n}} \ , \tag{4.8}$$

with probability at least $1 - 1/n^2$.

The uniform bound (4.8) contains three terms. The first one is a bias-type term which stems from the fact that we aim at estimating positions in $\mathcal{P}_n$. Also, the second term accounts for the estimation error of the preliminary estimator $\mathbf{x}$ in l_1 type distance. The last term is of the order of $\sqrt{\log(n)/n}$ and turns out to be optimal (see Section 4.3.3).

Since the definition of $\hat{\mathbf{x}}$ does not not depend on the choice of the latent representation $(\mathbf{x}^*, f)$, the conclusion (4.8) remains true for any representatives $(\mathbf{x}_b^*, f_b)$ of the equivalence class $[(\mathbf{x}^*, f)]_{bilip}$, that is

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}_b^*) \lesssim_{c_l,c_L,c_e} \min_{\mathbf{y}\in\mathcal{P}_n} d_\infty(\mathbf{y}, \mathbf{x}_b^*) + \frac{d_{1,\mathcal{O}}(\mathbf{x}, \mathbf{x}_b^*)}{n} + \sqrt{\frac{\log(n)}{n}} \ .$$

This simple observation remains true for all the reconstruction error bounds considered in section 4.3 and we may consider a representative $(\mathbf{x}_u^*, f_u)$ achieving the minimum in the right hand side term of the above bound.

To decrease the time complexity, it is possible to restrict the $\hat{x}_i$ in (4.7) to belong to $\mathcal{C}_{\sqrt{n}}$ instead of $\mathcal{C}_n$. In that case, one can check from the proof of Proposition 4.3.1 that the result (4.8) remains true with different constants.

For the purpose of conciseness, we write in the sequel the uniform approximation error of $\mathbf{x}^*$ by a vector in $\mathcal{P}_n$ as

$$\alpha(\mathbf{x}^*) = \min_{\mathbf{z}\in\mathcal{P}_n} d_\infty(\mathbf{z}, \mathbf{x}^*) \ . \tag{4.9}$$

In view of the above proposition, we seek to building a preliminary estimator $\hat{\mathbf{x}}'$ such that $d_{1,\mathcal{O}}(\mathbf{x}, \mathbf{x}^*)$ is smaller than $n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)}$.

4.3.1.2 Preliminary estimator

Given any vector of points $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{C}^n$, define the $n \times n$ matrix of distances $D(\mathbf{x}) = \big(d(x_i, x_j)\big)_{i,j\le n}$. We consider the following estimator

$$\hat{\mathbf{x}}' \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{P}_n} \langle A^{(1)}, D(\mathbf{x})\rangle \ . \tag{4.10}$$

This estimation looks for a vector $\hat{\mathbf{x}}' = (\hat{x}'_1, \dots, \hat{x}'_n)$ such that the distances $d(\hat{x}'_i, \hat{x}'_j)$ are small when the signals $\mathbb{E}\, A^{(1)}_{ij} = f(x_i^*, x_j^*)$ are large, or equivalently (since f belongs to the class of bi-Lipschitz functions), when the true distances $d(x_i^*, x_j^*)$ are small. For the purpose of understanding (4.10), let us assume that the positions $\mathbf{x}^*$ are evenly spread (that is $\mathbf{x}^* \in \mathcal{P}_n$), that the matrix F is observed (noiseless data), and that $f(x, z)$ is proportional to $d(x, z)$. Then, one can check that the minimum of $\langle D(\mathbf{x}^*), D(\mathbf{x})\rangle$ is achieved at all $Q\mathbf{x}^*$ where $Q \in \mathcal{O}$

is an orthogonal transformation in the plane. In other words, the criterion (4.10) exactly recovers (up to distance preserving transformations) the respective positions in this simplified setting. The following proposition establishes a $d_{1,\mathcal{O}}$ bound in the general case.

Unfortunately, (4.10) amounts to optimizing a criterion over the space of permutations (up to some symmetries) and we are not aware of any polynomial-time algorithm for computing $\hat{\mathbf{x}}'$. In the next section, we shall introduce a polynomial time alternative to this estimator under further model assumptions.

Proposition 4.3.2 *Consider $f \in \mathcal{BL}[c_l, c_L, c_e]$ and $\mathbf{x}^* \in \mathcal{C}^n$. With probability higher than $1 - 1/n^2$, the estimator $\hat{\mathbf{x}}'$ defined in (4.10) satisfies*

$$d_{1,\mathcal{O}}(\hat{\mathbf{x}}', \mathbf{x}^*) \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*) + \sqrt{n \log n} \quad .$$

As for the previous proposition, this $d_{1,\mathcal{O}}$ bound is actually valid for all representatives in the equivalence class $[(\mathbf{x}^*, f)]_{bilip}$. In particular, we readily deduce from the previous bound that

$$d_{1,\mathcal{O}}(\hat{\mathbf{x}}', \mathbf{x}_u^*) \lesssim_{c_l,c_L,c_e} \inf_{(\mathbf{x}_b^*, f_b) \in [(\mathbf{x}^*, f)]_{bilip}} n\alpha(\mathbf{x}_b^*) + \sqrt{n \log n} \ .$$

In the proof of Proposition 4.3.2, we shall establish that the estimator $\hat{\mathbf{x}}'$ is such that $\|D(\mathbf{x}^*) - D(\hat{\mathbf{x}}')\|_2$ is small, meaning that the distances between $\hat{x}'_1, \dots, \hat{x}'_n$ are close to the true distances between $x_1^*, \dots, x_n^*$. Then, relying on a recent result on matrix perturbation from [Arias-Castro et al., 2020], we shall deduce that $\|\hat{\mathbf{x}}' - Q\mathbf{x}^*\|_2$ is small, where Q is a distance preserving transformation. Using the equivalence between distances in $\mathbb{R}^2$, this will lead to the bound of Proposition 4.3.2.

4.3.2 Uniform localization with UTS

To conclude, we gather the preliminary estimator and the local refinement estimator.

UTS algorithm

Inputs: adjacency matrices $\{A_{ij}^{(1)}\}$ and $\{A_{ij}^{(2)}\}$

Preliminary estimator: Compute $\hat{\mathbf{x}}'$ based on (4.10),

$$\hat{\mathbf{x}}' = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{P}_n} \langle A^{(1)}, D(\mathbf{x}) \rangle \ .$$

Local refinement: For $i = 1, \dots, n$, compute $\hat{x}_i$ as in (4.7)

$$\hat{x}_i = \operatorname*{argmin}_{x \in \mathcal{C}_n} \langle A_i^{(2)}, D(x, \hat{\mathbf{x}}') \rangle \ .$$

Output: $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$

The following theorem is a straightforward consequence of Propositions 4.3.1 and 4.3.2 by taking $\mathbf{x} = \hat{\mathbf{x}}'$ in the local refinement step.

Theorem 4.3.3 *Consider $f \in \mathcal{BL}[c_l, c_L, c_e]$ and $\mathbf{x}^* \in \mathcal{C}^n$. With probability higher than $1 - 2/n^2$, the UTS estimator $\hat{\mathbf{x}}$ satisfies*

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l,c_L,c_e} \min_{\mathbf{x}\in\mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}} \ . \tag{4.11}$$

This oracle inequality holds for any vector of positions $\mathbf{x}^* \in \mathcal{C}^n$ even if it is not an element of $\mathcal{P}_n$. Note that the bound (4.11) has two error terms. The first one corresponds to the approximation error between $\mathbf{x}^*$ and the targeted set $\mathcal{P}_n$. Choosing the representation $(\mathbf{x}_u^*, f_u)$ with minimum bias $\alpha(\mathbf{x}_u^*)$ in the equivalence class $[(\mathbf{x}^*, f)]_{bilip}$, we arrive at

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}_u^*) \lesssim_{c_l,c_L,c_e} \inf_{(\mathbf{x}_b,f_b)\in[(\mathbf{x}^*,f)]_{bilip}} \min_{\mathbf{x}\in\mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}_b) + \sqrt{\frac{\log(n)}{n}} \ .$$

In particular, the bias term in Theorem 4.3.3 becomes negligible if, for some constant $c_a > 0$, $\mathbf{x}^*$ satisfies

$$\alpha(\mathbf{x}^*) = \min_{\mathbf{y}\in\mathcal{P}_n} d_\infty(\mathbf{y}, \mathbf{x}^*) \le c_a \sqrt{\frac{\log(n)}{n}} \ , \tag{4.12}$$

or more generally if the minimum bias $\alpha(\mathbf{x}_u^*)$ of the equivalence class satisfies this bound. For instance, if we assume that the positions $x_1^*, \dots, x_n^*$ have been sampled independently and uniformly on $\mathcal{C}$, then Assumption (4.12) is satisfied with high probability.

Corollary 4.3.4 *If there exists a representation $(\mathbf{x}^*, f)$ with $f \in \mathcal{BL}[c_l, c_L, c_e]$ and $\mathbf{x}^*$ satisfying (4.12), then*

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l,c_L,c_e,c_a} \sqrt{\frac{\log(n)}{n}} \ ,$$

with probability at least $1 - 2/n^2$.
Consider a function $f \in \mathcal{BL}[c_l, c_L, c_e]$ and assume that the latent positions (x_i^*) have been uniformly and independently sampled on $\mathcal{C}$. Then, with probability higher than $1 - 3/n^2$, we have

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l,c_L,c_e} \sqrt{\frac{\log(n)}{n}} \ .$$

4.3.3 Minimax lower bounds

In this subsection, we assess that the rate $\sqrt{\log(n)/n}$ in Theorem 4.3.3 is unimprovable, that is, no estimator is able to uniformly recover the latent positions at a rate faster than that.

Let us consider the single observation model $A = F + E$ where we assume that, for $i < j$, the A_{ij} follow independent Bernoulli distributions with parameters $f(x_i^*, x_j^*)$. We focus on this specific case of sub-Gaussian distributions in the lower bound because we have in mind random graph applications, but the following impossibility result also holds for Gaussian noise and when the statistician is given two independent observations $A^{(1)}$ and $A^{(2)}$.

Define the class $S = \{\mathbf{x}^* \in \mathcal{C}^n \text{ satisfying (4.12)}\}$ of positions that are nearly evenly spread. We consider here a simpler setting where f_0 is known to the statistician and is an affine function of d,

$$f_0(x, y) = (3/4) - d(x, y)/(4\pi) \ ,$$

for all x, $y \in \mathcal{C}$. Obviously, this function f_0 corresponds to a random geometric graph as discussed in the introduction and it satisfies the bi-Lipschitz assumption with $c_e = 0$ and $c_l = c_L = (4\pi)^{-1}$. Recall that we write $\mathbb{P}_{(\mathbf{x}^*, f_0)}$ for the distribution of A with $(\mathbf{x}^*, f_0)$.

Theorem 4.3.5 *There exist an integer $n_0 \geq 1$, and a positive constant C only depending on c_a in (4.12), such that the following holds for any $n \geq n_0$. For any estimator $\hat{\mathbf{x}}$, we have*

$$\sup_{\mathbf{x}^* \in S} \mathbb{P}_{(\mathbf{x}^*, f_0)} \left[d_{\infty, \mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \geq C \sqrt{\frac{\log(n)}{n}} \right] \geq \frac{1}{2} \ .$$

The above lower bound holds in the simpler situation where the affinity function is known. In this case $f = f_0$, the positions are identifiable up to orthogonal transformations, and accordingly, the pseudo-distance $d_{\infty,\mathcal{O}}$ takes the minimum over all orthogonal transformations. This lower bound matches the upper bound in Corollary 4.3.4, and thus implies the optimality of the UTS estimator (in the minimax sense).

Since f_0 is an affine function of the distance $d(.,.)$ and therefore $f_0 \in \mathcal{BL}[(4\pi)^{-1}, (4\pi)^{-1}, 0]$, this entails that the $\sqrt{\log(n)/n}$ rate is not driven by the slack $\epsilon_n = c_e\sqrt{\log(n)/n}$ in the bi-Lipschitz assumption (4.1,4.2). In fact, we precisely allowed this slack of $c_e\sqrt{\log(n)/n}$ in the bi-Lipschitz conditions because this slight generalization does not lead to slower estimation rates than for the case of pure bi-Lipschitz functions ($c_e = 0$).

4.4 Spectral localization in the geometric case

Since the preliminary estimator of UTS exhibits exponential-time complexity, we study here a spectral alternative of UTS. Unfortunately, spectral methods require further assumptions to work. Throughout this section, we additionally assume that the function f is *geometric*, that is, there exists a function $\tilde{f} : \mathbb{R}^+ \to [0, 1]$ such that

$$\forall x, y \in \mathcal{C}, \qquad f(x, y) = \widetilde{f}(d(x, y)) \ . \tag{4.13}$$

In other words, $f(x, y)$ only depends on the positions x and y through the distance $d(x, y)$. When the observations are binary random variables, this corresponds to geometric random graph model of Example 1 in the introduction.

4.4.1 Spectral algorithm in a geometric model

The spectral algorithm described below amounts to estimating x_j^* using the j-th coordinates of the second and third eigenvector of $A^{(1)}$ (see the algorithm below). This procedure is sometimes used in the seriation literature. See e.g. Recanati et al., 2018 for an analysis in the noiseless case. In the following proposition, we rigorously extend their work to the noisy setting and to non exact pre-R matrices. Then, combined with our local refinement step, we establish uniform localisation bounds in the next subsection.

Let $\hat{\lambda}_0 \geq \ldots \geq \hat{\lambda}_{n-1}$ denote the eigenvalues of the adjacency matrix $A^{(1)}$.

Spectral algorithm

Inputs: adjacency matrix $\{A^{(1)}_{ij}\}$

1. Compute the 2nd and 3rd eigenvectors of $A^{(1)}$ (i.e. associated with $\hat{\lambda}_1, \hat{\lambda}_2$); output orthonormal vectors $(\hat{u}_1, \dots, \hat{u}_n)$ and $(\hat{v}_1, \dots, \hat{v}_n)$.
2. Set $\hat{x}'_i = \sqrt{\frac{n}{2}}(\hat{u}_i, \hat{v}_i)$ for all $i \in [n]$.

Output: $\hat{\mathbf{x}}' = (\hat{x}'_1, \dots, \hat{x}'_n) \in \mathbb{R}^{2\times n}$.

Note that the estimated positions $\hat{x}'_i$ do not lie on the unit circle $\mathcal{C}$. As a consequence, the quantity $d(\hat{x}'_i, x^*_i)$ is not defined. Considering the positions $\mathbf{x}^*$ as a $2 \times n$ matrix, recall that $\|\hat{\mathbf{x}}' - \mathbf{x}^*\|_1 = \sum_{i=1}^n \sum_{j=1}^2 |\hat{x}'_{ij} - x^*_{ij}|$. Since x^* can be recovered at best up to orthogonal transformation, we therefore consider the loss function $\|\hat{\mathbf{x}}' - \mathbf{x}^*\|_{1,\mathcal{O}}$ defined by

$$\|\mathbf{x} - \mathbf{y}\|_{1,\mathcal{O}} = \min_{Q\in\mathcal{O}} \|\mathbf{x} - Q\mathbf{y}\|_1 \ ,$$

for any two matrices x, $y \in \mathbb{R}^{2\times n}$.

Let $\lambda^*_0 \geq \ldots \geq \lambda^*_{n-1}$ denote the eigenvalues of the matrix $F_{(\mathbf{x}^*,f)} = \{f(x^*_i, x^*_j)\}$. Then denote the two spectral gaps of interest by $\Delta_1 = \lambda^*_0 - \lambda^*_1$ and $\Delta_2 = \lambda^*_2 - \lambda^*_3$.

Proposition 4.4.1 *Assume that the noise random variables $E^{(1)}_{ij}$ for $i < j$ are independent. Consider a geometric function f in $\mathcal{BL}[c_l, c_L, c_e]$ and latent positions satisfying (4.12). The spectral algorithm $\hat{\mathbf{x}}'$ satisfies*

$$\|\hat{\mathbf{x}}' - \mathbf{x}^*\|_{1,\mathcal{O}} \lesssim_{c_l,c_L,c_e,c_a} \frac{n\sqrt{n\log(n)}}{(\Delta_1 \wedge \Delta_2) \vee 1}$$

with probability at least $1 - 1/n^2$.

This proposition is based on the fact that the signal F is well approximated by the set of circulant and circular-R matrices, which benefits from nice spectral properties. See e.g. appendix 4.C.2 for definitions of these matrices and their spectrum. This type of R-matrices were already studied in [Recanati et al., 2018] to derive similar error bounds on the reconstruction of positions [Recanati et al., 2018, Proposition D.1]. Here, Proposition 4.4.1 extends their result by giving a more explicit stochastic bound and also considering a more general signal F which is not assumed to be a circulant and circular-R matrix.

Note that the bound of Proposition 4.4.1 is uninformative if the spectral gaps $\Delta_1 \wedge \Delta_2$ are small compared to $\sqrt{n\log(n)}$. We add the term $\vee 1$ just in case $\Delta_1 \wedge \Delta_2 = 0$. At the end of the section, we shall provide examples of geometric functions f such that the spectral gap is large enough.

4.4.2 Generalization of local refinement step

Since $\hat{\mathbf{x}}'$ does not lie in $\mathcal{C}^n$, we cannot directly plug-it to the local refinement step defined in (4.7). In this subsection, we add a new step which somewhat projects $\hat{\mathbf{x}}'$ onto $\mathcal{P}_n$. More

generally, consider any matrix $\mathbf{x} \in \mathbb{R}^{2\times n}$. The uniform approximation (UA) algorithm defined below outputs a vector $\tilde{\mathbf{x}} \in \mathcal{P}_n$ which is close to $\mathbf{x}$ with respect to $\|.\|_{1,\mathcal{O}}$ pseudo distance.

For a vector $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{C}^n$, we say that $\mathbf{x}$ is ordered, if these points are consecutive when one walks on the circle using the trigonometric direction. In the following algorithm, $\tau = \big(1, \dots, n\big)$ stands for the circular permutation.

Uniform Approximation (UA) algorithm

Input: $\mathbf{x} \in \mathbb{R}^{2\times n}$

1. For $i = 1, \dots, n$, compute the projection of x_i onto $\mathcal{C}$, called $x_i^{(p)}$.
2. Let σ be any permutation such that $x_{\sigma(1)}^{(p)}, \dots, x_{\sigma(n)}^{(p)}$ is ordered.
3. Pick $\widetilde{x}_1$ a closest point to $x_{\sigma(1)}^{(p)}$ in $\mathcal{C}_n$. For $i = 1, \dots, n$, set $\widetilde{x}_{i+1}$ such that the angle $\underline{\widetilde{x}}_{i+1}$ satisfies $\underline{\widetilde{x}}_{i+1} = \underline{\widetilde{x}}_1 + 2\pi i/n \,(\mathrm{mod}\, 2\pi)$.
4. For $\widetilde{\mathbf{x}}_{\tau^j} = (\widetilde{x}_{\tau^j(1)}, \dots, \widetilde{x}_{\tau^j(n)})$, compute j the minimizer of $\|\widetilde{\mathbf{x}}_{\tau^j} - \mathbf{x}_\sigma^{(p)}\|_1$ over $j = 0, \dots, n-1$.
   Do $\widetilde{\mathbf{x}} \leftarrow \widetilde{\mathbf{x}}_{\tau^j \circ \sigma^{-1}}$.

Output: $\widetilde{\mathbf{x}} \in \mathcal{P}_n$.

Lemma 4.4.2 *For any $\mathbf{x} \in \mathbb{R}^{2\times n}$ and $\mathbf{x}^* \in \mathcal{C}^n$, the vector $\widetilde{\mathbf{x}} \in \mathcal{P}_n$ returned by UA satisfies*

$$\|\widetilde{\mathbf{x}} - \mathbf{x}^*\|_{1,\mathcal{O}} \lesssim \|\mathbf{x} - \mathbf{x}^*\|_{1,\mathcal{O}} + n\alpha(\mathbf{x}^*) + 1 \ .$$

Then, we use the same estimator as in (4.7) where we plug $\tilde{\mathbf{x}}$ instead of $\mathbf{x}$. In other words, we compute, for $i = 1, \dots, n$,

$$\hat{x}_i = \operatorname*{argmin}_{x \in \mathcal{C}_n} \langle A_i^{(2)}, D(x, \tilde{\mathbf{x}}) \rangle \ . \tag{4.14}$$

Putting everything together, we arrive at the following spectral UTS procedure.

Spectral-UTS algorithm

Inputs: adjacency matrix $\{A_{ij}^{(1)}\}$ and $\{A_{ij}^{(2)}\}$

Spectral initialization from $A^{(1)}$

1. Apply spectral algorithm to $A^{(1)}$; output $\hat{\mathbf{x}}' = (\hat{x}'_1, \dots, \hat{x}'_n)$.

Finding an approximation in $\mathcal{P}_n$

2. Apply UA algorithm to $\hat{\mathbf{x}}'$; output $\widetilde{\mathbf{x}}$.

Local refinement from $A^{(2)}$

3. For $i = 1, \dots, n$, compute $\hat{x}_i := \operatorname*{argmin}_{x \in \mathcal{C}_n} \langle A_i^{(2)}, D(x, \widetilde{\mathbf{x}}) \rangle$.

Output: $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$.

As a direct consequence of Propositions 4.3.1 and 4.4.1 and Lemma 4.4.2, we arrive at the following uniform bound.

Theorem 4.4.3 *Assume that the noise random variables $E_{ij}^{(1)}$ for $i < j$ are independent. Consider a geometric function f in $\mathcal{BL}[c_l, c_L, c_e]$ and latent positions satisfying (4.12). If $\Delta_1 \wedge \Delta_2 \geq c_b n$ for some positive constant c_b, then the spectral-UTS estimator $\hat{\mathbf{x}}$ satisfies*

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l, c_L, c_e, c_a, c_b} \sqrt{\frac{\log(n)}{n}}$$

with probability at least $1 - 2/n^2$.

Since the function f_0 in the minimax lower bound of Theorem 4.3.5 is geometric, the latter theorem implies that the rate $\sqrt{\log(n)/n}$ is optimal.

For affine functions such as $f(x, y) = c - c' \, d(x, y)/(2\pi)$ for some constants c, $c' > 0$, one checks that $\Delta_1 \wedge \Delta_2$ is of the order of n, thus fulfilling the condition of Theorem 4.4.3. For more details, see the spectrums of the matrices F in appendix 4.C.2.

4.5 Discussion

4.5.1 Summary

Under bi-Lipschitz assumptions on the underlying structure, we have established the minimax rate for uniform localization of balanced vectors $\mathbf{x}^* \in \mathcal{P}_n$. Besides, when the structure is non-evenly spaced (that is, when the latent points do not form a balanced vector), we prove that non-trivial estimation error is still possible. Specifically, such a perturbation yields an additional bias in the error bound, which is linear with the uniform difference between $\mathbf{x}^*$ and the set $\mathcal{P}_n$. It thus shows that robustness is achievable for the uniform norm. We think that this bias is not optimal and only reflects the shortcomings of our proofs.

A major defect of our procedure is an exponential-time complexity. Accordingly, we provide a polynomial-time alternative based on a spectral method. Its analysis is performed in the particular case of geometric functions, where we show the same minimax rate as in the bi-Lipschitz case.

4.5.2 Road-map toward a one-sample analysis

For simplicity, we have assumed that we are given two independent copies $A^{(1)}$ and $A^{(2)}$. However, we are mostly concerned with the original scenario of a single observation A. Below, we describe a possible procedure to deal with this case. This is not a proof, but rather a conjectural road-map for future work, with supporting arguments. The procedure has two steps. Firstly, it splits the data A in two independent parts $A^{(1)}$ and $A^{(2)}$ and then uses the UTS algorithm on each part as described earlier (up to minor modifications); this results in uniform estimates for the positions associated with $A^{(2)}$. Secondly, it repeats the same process by switching the roles of $A^{(1)}$ and $A^{(2)}$, thus giving uniform estimates for $A^{(1)}$. At the

end of the procedure, we have uniform estimates for all latent positions of A, with the same theoretical guarantees as in the previous sections.

Let us give a more precise description of the splitting procedure. We randomly partition the set of indices in two subsets $S^{(1)}$ and $S^{(2)}$ of respective size $n/4$ and $3n/4$. Then, let $\mathbf{x}^*_{S^{(1)}}$ and $\mathbf{x}^*_{S^{(2)}}$ denote the latent points respectively associated with $S^{(1)}$ and $S^{(2)}$. We split A by setting $A^{(1)} = \{A_{ij}\}_{i,j\in S^{(1)}}$ and $A^{(2)} = \{A_{ij}\}_{i\in S^{(2)},j\in S^{(1)}}$.

The d_1-estimation step uses the data $A^{(1)}$ to get position estimates $\hat{\mathbf{x}}'_{S^{(1)}}$ of $\mathbf{x}^*_{S^{(1)}}$. Proposition 4.3.2 ensures that, with high probability,

$$d_{1,\mathcal{O}}(\hat{\mathbf{x}}'_{S^{(1)}}, \mathbf{x}^*_{S^{(1)}}) \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*_{S^{(1)}}) + \sqrt{n\log n}\ , \tag{4.15}$$

where $\mathcal{P}_{n/4}$ is the set of balanced vectors of size $n/4$ and $\alpha(\mathbf{x}^*_{S^{(1)}}) := \min_{\mathbf{y}\in\mathcal{P}_{n/4}} d_\infty(\mathbf{x}^*_{S^{(1)}}, \mathbf{y})$.

Conditionally to $\hat{\mathbf{x}}'_{S^{(1)}}$, the refinement step uses the data $A^{(2)}$ to compute (almost) the same estimator as in the previous sections:

$$\hat{x}_i = \operatorname*{argmin}_{x\in\mathcal{C}_{n/4}} \langle A^{(2)}_i, D(x, \hat{\mathbf{x}}'_{S^{(1)}})\rangle, \tag{4.16}$$

for all $i \in S_2$, where $A^{(2)}_i$ denotes the i^{th}-row of $A^{(2)} = \{A_{ij}\}_{i\in S^{(2)},j\in S^{(1)}}$. By independence between samples $A^{(1)}$ and $A^{(2)}$, and following the proof of Proposition 4.3.1, we conjecture that (a version of) Proposition 4.3.1 is valid, and thus the combination with (4.15) leads to the following d_∞-guarantees for the position estimators associated with $S^{(2)}$:

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}_{S^{(2)}}, \mathbf{x}^*_{S^{(2)}}) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*_{S^{(1)}}) + \sqrt{\frac{\log(n)}{n}}. \tag{4.17}$$

Indeed, the bias term $\alpha(\mathbf{x}^*_{S^{(1)}})$ comes from the fact that the local estimator (4.16) uses the positions associated with $S^{(1)}$ as reference positions. In particular, the positions $\mathbf{x}^*_{S^{(2)}}$ are not involved in the local estimator (4.16), so it is normal that the corresponding bias term $\alpha(\mathbf{x}^*_{S^{(2)}})$ does not appear in the error bound (4.17).

Finally, we need to switch the roles of $A^{(1)}$ and $A^{(2)}$ to get uniform estimates for the remaining positions (those associated with $S^{(1)}$). We do so by randomly sampling a subset $\widetilde{S}^{(1)}$ of $[n]\backslash\{S^{(1)}\}$ with size $n/4$. Write $\widetilde{S}^{(2)} = [n]\backslash\{\widetilde{S}^{(1)}\}$. As above, we run the UTS algorithm on $A^{(1)} = \{A_{ij}\}_{i,j\in\widetilde{S}^{(1)}}$ and $A^{(2)} = \{A_{ij}\}_{i\in\widetilde{S}^{(2)},j\in\widetilde{S}^{(1)}}$, and therefore obtain d_∞-estimates for positions associated with $\widetilde{S}^{(2)}$:

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}_{\widetilde{S}^{(2)}}, \mathbf{x}^*_{\widetilde{S}^{(2)}}) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*_{\widetilde{S}^{(1)}}) + \sqrt{\frac{\log(n)}{n}}.$$

Since $S^{(1)} \subset \widetilde{S}^{(2)}$, we get uniform estimates for positions associated with $S^{(1)}$.

It remains a technical detail to handle: since the pseudo-metric $d_{\infty,\mathcal{O}}$ minimizes over the set $\mathcal{O}$ of orthogonal transformations, the above $d_{\infty,\mathcal{O}}$-bounds for $\hat{\mathbf{x}}_{S^{(2)}}$ and $\hat{\mathbf{x}}_{\widetilde{S}^{(2)}}$, may hold for two different orthogonal transformations. Hence, the position estimates for $S^{(2)}$ and $\widetilde{S}^{(2)}$ may not be "aligned" correctly. We conjecture that this minor issue can be settled by using

the non-empty intersection between $S^{(2)}$ and $\widetilde{S}^{(2)}$, as a reference set to align the two vectors of estimates. After this fix, we would obtain

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*_{S^{(1)}}) + \alpha(\mathbf{x}^*_{\widetilde{S}^{(1)}}) + \sqrt{\frac{\log(n)}{n}} \ .$$

We conjecture that the next (technical) inequality holds true

$$\alpha(\mathbf{x}^*_{S^{(1)}}) + \alpha(\mathbf{x}^*_{\widetilde{S}^{(1)}}) \lesssim \alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}} \ ,$$

with probability larger than $1 - c/n^2$. If this conjecture is true, it then follows from the above a uniform bound:

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}} \ .$$

This informal discussion about the one sample situation will be the subject of future work.

4.5.3 Open questions

A natural question is to know whether there exists a polynomial-time algorithm that achieves the minimax rate $\sqrt{\log(n)/n}$ in the bi-Lipschitz case with uniform sample $\mathbf{x}^*$. We conjecture that progress can be made towards this direction, by exploiting the uni-dimensional aspect of the latent space. For example, one could estimate the latent distances and then build on these estimates to embed the latent points in $\mathcal{C}$.

As discussed in Section 4.5.1, the bias-type term in Theorem 4.3.3 is perhaps pessimistic. A next challenge will be to determine if bias-terms in problem of position estimation can be removed or at least refined.

Another direction left open due to lack of time is the toroidal seriation problem and its connections with the current work. For now, we only state a direct consequence. Let F corresponds to a toroidal pre R-matrix and that we want to estimate a permutation $\sigma^* \in \Sigma_n$ such $F_{\sigma^*(i),\sigma^*(j)}$ is a R-matrix. Let us use the representation of F on the latent space $\mathcal{C}_n$ and let us assume that the corresponding function f on $\mathcal{C}_n \times \mathcal{C}_n$ can be extended in a bi-Lipschitz function of $\mathcal{C} \times \mathcal{C}$. Then, transforming the estimator $\widehat{x}$ from UTS into map $\widehat{\sigma} : [n] \to [n]$, one readily derives from Corollary 4.3.4 that, with high probability

$$\min_{\tau \in \mathcal{O}(\Sigma_n)} \max_{i \in [n]} \frac{|\tau \circ \sigma^*(i) - \hat{\sigma}(i)|}{n} \lesssim_{c_l,c_L,c_e} \sqrt{\frac{\log(n)}{n}} \ ,$$

where $\mathcal{O}(\Sigma_n)$ is the set of 'orthogonal transformations' on Σ_n (i.e. the permutations corresponding to rotations and reflections). Unfortunately, our minimax lower bound from Theorem 4.3.5 does not apply to this setting. Still, we conjecture that the above rate is optimal.

Appendices

4.A Proofs for UTS

Throughout the proofs, C stands for a positive function that may depend on other quantities such as c_l, c_L, c_e and may change from line to line.

Given an orthogonal matrix $Q \in \mathcal{O}$, we introduce the d_1-loss relative to Q by

$$\mu_{\mathbf{x},\mathbf{x}^*}(Q) := \frac{d_1(\mathbf{x}, Q\mathbf{x}^*)}{n} \quad . \tag{4.18}$$

Before proving Proposition 4.3.1, we study the simpler situation where the latent positions $\mathbf{x}^*$ belong to $\mathcal{P}_n$. In this case, $\alpha(\mathbf{x}^*) = 0$ and the result Proposition 4.3.1 follows directly from the next lemma.

Lemma 4.A.1 *For $\mathbf{x}^* \in \mathcal{P}_n$, $\mathbf{x} \in \mathcal{P}_n$ and $Q \in \mathcal{O}$, the estimator (4.7) satisfies the following uniform-bound*

$$d_\infty(\hat{\mathbf{x}}, Q\mathbf{x}^*) \lesssim_{c_l, c_L, c_e} \mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{\frac{\log(n)}{n}}$$

with probability at least $1 - 1/n^2$.

The proof of Proposition 4.3.1 for general x^* follows the same scheme as that of Lemma 4.A.1, but also requires some slight refinements. We first prove Lemma 4.A.1 before turning to the general case.

4.A.1 Proof of Lemma 4.A.1

First, we claim that suffices to restrict our attention to transformation $Q \in \mathcal{O}$ that let $\mathcal{C}_n$ invariant. Indeed, for general Q, there exists an orthogonal transformation Q', letting $\mathcal{C}_n$ invariant, and such that $\max_{z \in \mathcal{C}_n} d(Qz, Q'z) \lesssim 1/n$. Replacing Q' by Q in the statement of the lemma only entails an additional term of order $1/n$ which is negligible compared to the $\sqrt{\log(n)/n}$ term.

Let $i \in [n]$. In the two next lemmas, we bound

$$L_i := \langle F_i,\, D(Q^{-1}\hat{x}_i, \mathbf{x}^*) - D(x_i^*, \mathbf{x}^*) \rangle \tag{4.19}$$

from above and below.

Lemma 4.A.2 *With probability at least $1 - 1/n^3$, we have*

$$L_i \leq C d(\hat{x}_i, Qx_i^*) \left(n\mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{n \log(n)} \right) \quad ,$$

for some constant $C > 0$ that only depends on c_L and c_e.

Lemma 4.A.3 *We have*

$$L_i \geq c' n d(\hat{x}_i, Qx_i^*) \Big(c_l d(\hat{x}_i, Qx_i^*) - \epsilon_n \Big) - \frac{c_e^3}{\pi c_l^2} \sqrt{\frac{\log^3(n)}{n}} \quad ,$$

for some numerical constant $c' > 0$ and all n larger than C (only depending on c_l and c_e).

The two lemmas imply that, for n large enough, and $d(\hat{x}_i, Qx_i^*) \geq 2C'\sqrt{\log(n)/n}$ with C' a large enough constant that only depends on c_l and c_e,

$$C'nd^2(\hat{x}_i, Qx_i^*) \leq L_i \leq Cd(\hat{x}_i, Qx_i^*)\left(n\mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{n\log(n)}\right) \ .$$

Hence, we conclude that the error bound $d(\hat{x}_i, Qx_i^*) \lesssim_{c_l,c_L,c_e} \mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{\log(n)/n}$ holds with probability at least $1 - 1/n^3$. Since the bound holds for an arbitrary $i \in [n]$, Lemma 4.A.1 follows from an union bound over all $i \in [n]$.

4.A.1.1 Proof of Lemma 4.A.2

First, we decompose L_i as follows

$$L_i = \sum_{j=1}^{n} f(x_i^*, x_j^*)\Big(d(Q^{-1}\hat{x}_i, x_j^*) - d(x_i^*, x_j^*)\Big).$$

Since $\mathcal{C}_n$ is invariant by Q and since $\mathbf{x}^*$ belongs to $\mathcal{P}_n$, we have $\{x_j^*;\, j \in [n]\} = \mathcal{C}_n = \{Q^{-1}x_j;\, j \in [n]\}$. Hence, we can reorder the last sum as follows

$$L_i = \sum_{j=1}^{n} f(x_i^*, Q^{-1}x_j)\Big(d(Q^{-1}\hat{x}_i, Q^{-1}x_j) - d(x_i^*, Q^{-1}x_j)\Big) \ .$$

To alleviate the notation, we rewrite $\hat{z}_i := Q^{-1}\hat{x}_i$ and $z_i := Q^{-1}x_i$. so that

$$L_i = \sum_{j=1}^{n} f(x_i^*, z_j)\Big(d(\hat{z}_i, z_j) - d(x_i^*, z_j)\Big) \ .$$

Lemma 4.A.4 *We have*

$$\sum_{j=1}^{n}\Big(f(x_i^*, z_j) - f(x_i^*, x_j^*)\Big)\Big(d(\hat{z}_i, z_j) - d(x_i^*, z_j)\Big) \lesssim_{c_L,c_e} d(\hat{x}_i, Qx_i^*)(n\mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{n\log(n)}) \ .$$

Gathering this lemma with the definition of L_i leads us to

$$L_i \lesssim_{c_L,c_e} d(\hat{x}_i, Qx_i^*)(n\mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{n\log(n)}) + \sum_{j=1}^{n} f(x_i^*, x_j^*)\Big(d(\hat{z}_i, z_j) - d(x_i^*, z_j)\Big) \ . \quad (4.20)$$

The orthogonal transformation Q preserves the distances, hence the last term of (4.20) is equal to

$$\sum_{j=1}^{n} f(x_i^*, x_j^*)\Big(d(Q^{-1}\hat{x}_i, Q^{-1}x_j) - d(x_i^*, Q^{-1}x_j)\Big) = \langle F_i,\, D(\hat{x}_i, \mathbf{x}) - D(Qx_i^*, \mathbf{x})\rangle \ . \quad (4.21)$$

To handle this term, we come back to the definition (4.7) of $\hat{x}_i$. Since $Qx_i^* \in \mathcal{C}_n$, we have

$$\langle A_i^{(2)}, D(\hat{x}_i, \mathbf{x})\rangle \leq \langle A_i^{(2)}, D(Qx_i^*, \mathbf{x})\rangle \ .$$

Since $A^{(2)}_{ii} = 0$, this yields

$$\langle F_i, D(\hat{x}_i, \mathbf{x}) - D(Qx^*_i, \mathbf{x})\rangle_{-i} \leq \langle E^{(2)}_i, D(Qx^*_i, \mathbf{x}) - D(\hat{x}_i, \mathbf{x})\rangle_{-i} \ ,$$

where $\langle ., . \rangle_{-i}$ denotes the inner product between the vectors whose i^{th} coordinate has been removed. Since $\hat{x}_i \in \mathcal{C}_n$, we simultaneously control this expression for all $z \in \mathcal{C}_n$. The expression $\langle E^{(2)}_i, D(Qx^*_i, \mathbf{x}) - D(z, \mathbf{x})\rangle_{-i}$ is a mean zero sub-Gaussian random variable with norm at most $c\|D(Qx^*_i, \mathbf{x}) - D(z, \mathbf{x})\|_2$. Applying the union bound over all $z \in \mathcal{C}_n$ leads us to

$$\langle E^{(2)}_i, D(Qx^*_i, \mathbf{x}) - D(\hat{x}_i, \mathbf{x})\rangle_{-i} \ \leq c'\sqrt{\log(n)}\|D(Qx^*_i, \mathbf{x}) - D(\hat{x}_i, \mathbf{x})\|_2 \ ,$$

with probability higher than $1 - 1/n^3$. Invoking the triangular inequality for the distance d, we deduce that $\|D(Qx^*_i, \mathbf{x}) - D(\hat{x}_i, \mathbf{x})\|_2 \leq d(\hat{x}_i, Qx^*_i)\sqrt{n}$. It follows that, with probability at least $1 - 1/n^3$,

$$\langle F_i,\ D(\hat{x}_i, \mathbf{x}) - D(Qx^*_i, \mathbf{x})\rangle_{-i} \leq c' d(\hat{x}_i, Qx^*_i)\sqrt{n\log(n)} \ .$$

The missing i^{th}-term in the above inner product satisfies

$$F_{ii}\left(d(\hat{x}_i, x_i) - d(Qx^*_i, x_i)\right) \leq d(\hat{x}_i, Qx^*_i) \ ,$$

since F is uniformly bounded by 1. We conclude that

$$\langle F_i,\ D(\hat{x}_i, \mathbf{x}) - D(Qx^*_i, \mathbf{x})\rangle \leq c'' d(\hat{x}_i, Qx^*_i)\sqrt{n\log(n)} \ .$$

Gathering this bound with (4.20) and (4.21) concludes the proof.

Proof of Lemma 4.A.4. Bi-lipschitz condition (4.1) ensures that $|f(x^*_i, z_j) - f(x^*_i, x^*_j)| \leq c_L d(z_j, x^*_j) + \epsilon_n$. By triangular inequality, we have $|d(\hat{z}_i, z_j) - d(x^*_i, z_j)| \leq d(\hat{z}_i, x^*_i)$ so that

$$\sum_{j=1}^{n}\Big(f(x^*_i, z_j) - f(x^*_i, x^*_j)\Big)\Big(d(\hat{z}_i, z_j) - d(x^*_i, z_j)\Big) \leq d(\hat{z}_i, x^*_i)\sum_{j=1}^{n}(c_L d(z_j, x^*_j) + \epsilon_n) \ ,$$

Since $d(z_j, x^*_j) = d(x_j, Qx^*_j)$ for any orthogonal transformation Q, it follows that

$$\sum_{j=1}^{n}\Big(f(x^*_i, z_j) - f(x^*_i, x^*_j)\Big)\Big(d(\hat{z}_i, z_j) - d(x^*_i, z_j)\Big) \leq d(\hat{x}_i, Qx^*_i)\left(c_L n\mu_{\mathbf{x},\mathbf{x}^*}(Q) + c_e\sqrt{n\log(n)}\right) \ ,$$

as $n\epsilon_n = c_e\sqrt{n\log(n)}$. □

4.A.1.2 Proof of Lemma 4.A.3.

An interval $I = [a, b]$ denotes the set of points lying between a and b in the one-dimensional torus $\mathbb{R}/(2\pi)$, when following the trigonometric direction from a to b. The length of I is denoted by $|I|$.

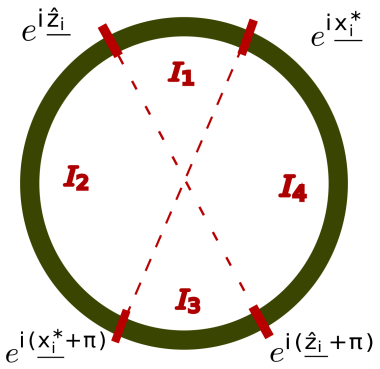


Figure 1 : partition of the circle

Let $x_i^*, \hat{x}_i$ be two points of $\mathcal{C}_n$, and denote $\hat{z}_i := Q^{-1}\hat{x}_i$. Since $d(x_i^*, \hat{z}_i) \leq \pi$, we can assume without loss of generality that $\underline{x}_i^* = 0$ and $\underline{\hat{z}}_i \in (0, \pi]$, so that we have the equality $d(x_i^*, \hat{z}_i) = |\underline{x}_i^* - \underline{\hat{z}}_i|$. If $\hat{z}_i = x_i^*$, Lemma 4.A.3 is trivial. We therefore assume in the following that $\underline{\hat{z}}_i \in (0, \pi]$. Below, we introduce a partition of $[n]$ according to the relative positions of x_j^*, x_i^* and $\hat{z}_i$. This partition is represented in Figure 4.1.

$$I_1 = \{j \in [n] : \underline{x}_j^* \in [\underline{x}_i^*, \underline{\hat{z}}_i)\} \ ; \qquad I_2 = \{j \in [n] : \underline{x}_j^* \in [\underline{\hat{z}}_i, \underline{x}_i^* + \pi)\} \ ;$$
$$I_3 = \{j \in [n] : \underline{x}_j^* \in [\underline{x}_i^* + \pi, \underline{\hat{z}}_i + \pi)\} \ ; \qquad I_4 = \{j \in [n] : \underline{x}_j^* \in [\underline{\hat{z}}_i + \pi, \underline{x}_i^*)\} \ .$$

Although I_s stands for a subset of indices, with a slight abuse of notation, we still write $|I_s|$ for the length of the corresponding interval in $\mathbb{R}/(2\pi)$. For instance, $|I_1| := |\underline{x}_i^* - \underline{\hat{z}}_i|$.

We decompose L according to this partition of indices

$$L_i = L_i^{(1)} + L_i^{(2)} + L_i^{(3)} + L_i^{(4)},$$

where $L_i^{(s)}$ is the restriction of L_i to the set I_s. In particular, if $\underline{\hat{z}}_i = \pi$, then the intervals I_2 and I_4 are empty, and $L_i^{(2)} = L_i^{(4)} = 0$.

We heavily rely on the fact that the elements of $\mathbf{x}^*$ are evenly spaced on the disc. Using the symmetry of the set $\mathcal{C}_n$, we establish below that the sums $L_i^{(2)}$ and $L_i^{(4)}$ nearly compensate so that $L_i^{(2)} + L_i^{(4)}$ admits a positive lower bound.

Lemma 4.A.5 *We have*

$$L_i^{(2)} + L_i^{(4)} \geq \frac{n|I_4|}{2\pi}\, d(\hat{x}_i, Qx_i^*)\Big(c_l d(\hat{x}_i, Qx_i^*) - \epsilon_n\Big) \ .$$

As for $L_i^{(1)}$ (resp. $L_i^{(4)}$), we rely on the symmetry of I_1 (resp. I_4) around the point of $\mathcal{C}$ whose argument is $(\underline{x}_i^* + \underline{z}_i)/2$ (resp. $((\underline{x}_i^* + \underline{z}_i)/2) + \pi$).

Lemma 4.A.6 *For some numerical constant $c > 0$, we have*

$$L_i^{(1)} + L_i^{(3)} \geq cn\big(\frac{|I_1|}{4} - c_l^{-1}\epsilon_n\big)\frac{d(\hat{x}_i, Qx_i^*)}{2}\big(c_l\frac{d(\hat{x}_i, Qx_i^*)}{2} - \epsilon_n\big) - (\pi c_l^2)^{-1}c_e^3\sqrt{\log^3(n)/n} \ .$$

By definition, $|I_1| + |I_4| = \pi$, which yields

$$L_i \geq c' n d(\hat{x}_i, Qx_i^*)\Big(c_l d(\hat{x}_i, Qx_i^*) - \epsilon_n\Big) - \frac{c_e^3}{\pi c_l^2}\sqrt{\frac{\log^3(n)}{n}} \ ,$$

for n large enough. □

Proof of Lemma 4.A.5. In Figure 4.1 of the intervals I_s, $s \in [4]$, we can see that the difference $d(\hat{z}_i, x_j^*) - d(x_i^*, x_j^*)$ is equal to $-d(\hat{z}_i, x_i^*)$ for all $j \in I_2$, whereas it is equal to the opposite, $d(\hat{z}_i, x_i^*)$, on I_4. Thus, we obtain

$$\begin{aligned} L_i^{(2)} &= \sum_{j \in I_2} f(x_i^*, x_j^*)\Big(d(\hat{z}_i, x_j^*) - d(x_i^*, x_j^*)\Big) = -d(\hat{z}_i, x_i^*) \sum_{j \in I_2} f(x_i^*, x_j^*) \ ; \\ L_i^{(4)} &= \sum_{j \in I_4} f(x_i^*, x_j^*)\Big(d(\hat{z}_i, x_j^*) - d(x_i^*, x_j^*)\Big) = d(\hat{z}_i, x_i^*) \sum_{j \in I_4} f(x_i^*, x_j^*) \ . \end{aligned}$$

Let ϕ denote the symmetry with respect to the line going through the two points of $\mathcal{C}$ of arguments

$$a = \frac{\underline{x}_i^* + \underline{\hat{z}}_i}{2} \quad \text{and} \quad b = \frac{(\underline{x}_i^* + \pi) + (\underline{\hat{z}}_i + \pi)}{2}.$$

As can be checked in Figure 4.1, for any $l \in I_2$, we have $\phi(x_j^*) = x_l^*$ for some j in I_4. Hence,

$$L_i^{(2)} + L_i^{(4)} = d(\hat{z}_i, x_i^*) \sum_{j \in I_4} \big(f(x_i^*, x_j^*) - f(x_i^*, \phi(x_j^*))\big) \ .$$

To lower bound the difference in the sum , we invoke the bi-Lipschitz condition (4.2), which gives

$$f(x_i^*, x_j^*) - f(x_i^*, \phi(x_j^*)) \geq c_l(d(x_i^*, \phi(x_j^*)) - d(x_i^*, x_j^*)) - \epsilon_n \ ,$$

since x_j^* is closer to x_i^* than $\phi(x_j^*)$ (see again Figure 4.1). Again, we can check from Figure 4.1 that $d(x_i^*, \phi(x_j^*)) - d(x_i^*, x_j^*) = d(\hat{z}_i, x_i^*)$ for all $j \in I_4$. Since $\mathcal{C}_n$ is evenly spaced, the number of indices j in I_4 is larger than $n|I_4|/(2\pi)$. This leads us to

$$L_i^{(2)} + L_i^{(4)} \geq \frac{n|I_4|}{2\pi} d(\hat{z}_i, x_i^*)\big(c_l d(\hat{z}_i, x_i^*) - \epsilon_n\big) \ .$$

Since $d(\hat{z}_i, x_i^*) = d(\hat{x}_i, Qx_i^*)$, this concludes the proof. □

Proof of Lemma 4.A.6. From Figure 4.1, we see that, for all $j \in I_1$,

$$d(\hat{z}_i, x_j^*) - d(x_i^*, x_j^*) = |\underline{\hat{z}}_i - \underline{x}_j^*| - |\underline{x}_i^* - \underline{x}_j^*| = \underline{\hat{z}}_i + \underline{x}_i^* - 2\underline{x}_j^* \ .$$

For $\alpha \in (0, 1)$, denote by $I_1^{(\alpha)}$ the sub-interval of I_1 defined by

$$I_1^{(\alpha)} = \big\{j \in [n] : \underline{x}_j^* \in [\underline{x}_i^*, (1-\alpha)\underline{x}_i^* + \alpha\underline{\hat{z}}_i)\big\} \ .$$

In particular, for all $j \in I_1^{(1/2)}$, the above expression leads us to

$$d(\hat{z}_i, x_j^*) - d(x_i^*, x_j^*) = -\big(\underline{\hat{z}}_i + \underline{x}_i^* - 2\underline{\phi(x_j^*)}\big) \ ,$$

where ϕ is the symmetry introduced in the proof of Lemma 4.A.5. Hence, we get

$$\begin{aligned} L_i^{(1)} &= \sum_{j\in I_1} f(x_i^*,x_j^*)\big(d(\hat{z}_i,x_j^*)-d(x_i^*,x_j^*)\big) \\ &= \sum_{j\in I_1^{(1/2)}} \big(f(x_i^*,x_j^*)-f(x_i^*,\phi(x_j^*))\big)\big(\hat{\underline{z}}_i+\underline{x}_i^*-2\underline{x}_j^*\big). \end{aligned}$$

For any $j\in I_1^{(1/2)}$, we have $d(x_i^*,\phi(x_j^*))\geq d(x_i^*,x_j^*)$. As a consequence, it follows from the bi-lipschitz condition (4.2) that

$$f(x_i^*,x_j^*)-f(x_i^*,\phi(x_j^*))\geq c_l\left[d(x_i^*,\phi(x_j^*))-d(x_i^*,x_j^*)\right]-\epsilon_n \ .$$

Since $d(x_i^*,\phi(x_j^*))-d(x_i^*,x_j^*)=|\underline{\phi(x_j^*)}-\underline{x}_j^*|$ for all $j\in I_1^{(1/2)}$, we get

$$L_i^{(1)}\geq \sum_{j\in I_1^{(1/2)}} (c_l|\underline{\phi(x_j^*)}-\underline{x}_j^*|-\epsilon_n)(\hat{\underline{z}}_i+\underline{x}_i^*-2\underline{x}_j^*) \ . \tag{4.22}$$

To control (4.22), we split the interval $I_1^{(1/2)}$ according to the sign of the term $(c_l|\underline{\phi(x_j^*)}-\underline{x}_j^*|-\epsilon_n)$. That is, we write $I_1^{(1/2)}=I_1^{(1/2)-}\cup I_1^{(1/2)+}$ where $I_1^{(1/2)-}$ is the set of indices j such that $c_l|\underline{\phi(x_j^*)}-\underline{x}_j^*|<\epsilon_n$.

Claim 4.A.7 $\sum_{j\in I_1^{(1/2)-}}(c_l|\underline{\phi(x_j^*)}-\underline{x}_j^*|-\epsilon_n)(\hat{\underline{z}}_i+\underline{x}_i^*-2\underline{x}_j^*)\geq -(c_l^2 2\pi)^{-1}c_e^3\sqrt{\log^3(n)/n}$.

Claim 4.A.8

$$\sum_{j\in I_1^{(1/2)+}} (c_l|\underline{\phi(x_j^*)}-\underline{x}_j^*|-\epsilon_n)(\hat{\underline{z}}_i+\underline{x}_i^*-2\underline{x}_j^*)\geq cn\big(\frac{|I_1|}{4}-c_l^{-1}\epsilon_n\big)\frac{|\hat{\underline{z}}_i-\underline{x}_i^*|}{2}\big(c_l\frac{|\hat{\underline{z}}_i-\underline{x}_i^*|}{2}-\epsilon_n\big) \ .$$

Gathering these two claims leads us to

$$L_i^{(1)}\geq cn\big(\frac{|I_1|}{4}-c_l^{-1}\epsilon_n\big)\frac{|\hat{\underline{z}}_i-\underline{x}_i^*|}{2}\big(c_l\frac{|\hat{\underline{z}}_i-\underline{x}_i^*|}{2}-\epsilon_n\big)-(c_l^2 2\pi)^{-1}c_e^3\sqrt{\log^3(n)/n} \ ,$$

which is the desired bound since $|\hat{\underline{z}}_i-\underline{x}_i^*|=d(\hat{z}_i,x_i^*)=d(\hat{x}_i,Qx_i^*)$. By symmetry, the term $L_i^{(3)}$ is handled as $L_i^{(1)}$ and admits the same lower bound. □

Proof of Claim 4.A.7. For simplicity, the notation $\underline{x}$ is dropped out in the proof of Claim 4.A.7, and $\underline{x}$ is denoted by x. By definition of ϕ, we know that $(\hat{z}_i+x_i^*)/2=(\phi(x_j^*)+x_j^*)/2$ for all $j\in I_1^{(1/2)}$, which gives the equality $\hat{z}_i+x_i^*-2x_j^*=\phi(x_j^*)-x_j^*$. Since $0\leq\phi(x_j^*)-x_j^*<c_l^{-1}\epsilon_n$ for all $j\in I_1^{(1/2)-}$, we have

$$0\leq \hat{z}_i+x_i^*-2x_j^*\leq c_l^{-1}\epsilon_n \ .$$

Since $c_l|\phi(x_j^*)-x_j^*|-\epsilon_n<0$ for j in $I_1^{(1/2)-}$, we obtain the following inequality

$$(c_l|\phi(x_j^*)-x_j^*|-\epsilon_n)(\hat{z}_i+x_i^*-2x_j^*)\geq (c_l|\phi(x_j^*)-x_j^*|-\epsilon_n)c_l^{-1}\epsilon_n\geq -c_l^{-1}\epsilon_n^2 \ .$$

Since the number of indices in $I_1^{(1/2)-}$ is at most $n/(2\pi)$ times the arc length $|I_1^{(1/2)-}|$, and the length of this arc is at most $c_l^{-1}\,\epsilon_n$, we conclude that

$$\sum_{j\in I_1^{(1/2)-}} (c_l|\phi(x_j^*)-x_j^*|-\epsilon_n)(\hat{z}_i+x_i^*-2x_j^*) \geq -\frac{n}{2\pi}c_l^{-2}\epsilon_n^3 = -\frac{c_l^{-2}}{2\pi}c_e^3\sqrt{\frac{\log^3(n)}{n}}\ .$$

□

Proof of Claim 4.A.8. Again, for convenience the notation $\underline{x}$ is dropped out here: $\underline{x}$ is denoted by x in the proof of Claim 4.A.8. Since all the terms of the sum are non-negative, we can simply consider indices j in $I_1^{(1/2)+}\cap I_1^{(1/4)}$. Using $\phi(x_j^*)-x_j^* = \hat{z}_i + x_i^* - 2x_j^*$ for all $j\in I_1^{(1/2)}$ and $x_i^*=0$, we obtain that for $j\in I_1^{(1/4)}$, $\phi(x_j^*)-x_j^*\geq \frac{\hat{z}_i}{2}$. This gives

$$(c_l|\phi(x_j^*)-x_j^*|-\epsilon_n)(\hat{z}_i+x_i^*-2x_j^*) \geq (c_l\frac{\hat{z}_i}{2}-\epsilon_n)\frac{\hat{z}_i}{2}\ ,$$

and for some numerical constant $c>0$:

$$\sum_{j\in I_1^{(1/2)+}} (c_l|\phi(x_j^*)-x_j^*|-\epsilon_n)(\hat{z}_i+x_i^*-2x_j^*) \geq cn\big|I_1^{(1/2)+}\cap I_1^{(1/4)}\big|\frac{\hat{z}_i}{2}(c_l\frac{\hat{z}_i}{2}-\epsilon_n)\ . \tag{4.23}$$

Since either $I_1^{(1/2)+}\subset I_1^{(1/4)}$ or $I_1^{(1/4)}\subset I_1^{(1/2)+}$ and $|I_1^{(1/4)}| = |I_1|/4$ and $|I_1^{(1/2)+}| = |I_1^{(1/2)}| - |I_1^{(1/2)-}| \geq |I_1^{(1/2)}| - c_l^{-1}\epsilon_n = \frac{|I_1|}{2} - c_l^{-1}\epsilon_n$, we deduce that

$$|I_1^{(1/2)+}\cap I_1^{(1/4)}| \geq \frac{|I_1|}{4} - c_l^{-1}\epsilon_n\ .$$

Thus, we have

$$\sum_{j\in I_1^{(1/2)+}} (c_l|\phi(x_j^*)-x_j^*|-\epsilon_n)(\hat{z}_i+x_i^*-2x_j^*) \geq cn\big(\frac{|I_1|}{4}-c_l^{-1}\epsilon_n\big)\frac{\hat{z}_i}{2}(c_l\frac{\hat{z}_i}{2}-\epsilon_n)\ .$$

Since $\hat{z}_i = |\hat{z}_i - x_i^*|$, this concludes the proof. □

4.A.2 Proof of Proposition 4.3.1

Let $\mathbf{x}^{**}\in\mathcal{P}_n$ be a closest approximation of $\mathbf{x}^*$ in $\mathcal{P}_n$, that is, such that $d_\infty(\mathbf{x}^*,\mathbf{x}^{**}) = \alpha(\mathbf{x}^*)$.

To prove Proposition 4.3.1, it suffices to establish a variant of Lemmas 4.A.2 and 4.A.3, in which we replace L_i by

$$\widetilde{L}_i := \langle F_i,\, D(Q^{-1}\hat{x}_i,\mathbf{x}^{**}) - D(x_i^*,\mathbf{x}^{**})\rangle, \tag{4.24}$$

which is equal to L_i after substituting $\mathbf{x}^*$ with $\mathbf{x}^{**}$ in the right entry of $D(.,.)$. The following variants of Lemmas 4.A.2 and 4.A.3 hold.

Lemma 4.A.9 *with probability at least* $1-1/n^3$*, we have* $\tilde{L}_i \lesssim_{c_L,c_e} 1 + d(\hat{x}_i, Qx_i^*)\big(n\alpha(\mathbf{x}^*) + n\mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{n\log(n)}\,\big)$.

Lemma 4.A.10 *For all* $i \in [n]$*, there exists a positive constant* C *only depending on* c_L *and* c_e *and positive constants* C' *and* C'' *only depending on* c_l *and* c_e *such that*

$$\begin{aligned}\widetilde{L}_i \;\geq\; & C' n d(\hat{x}_i, Qx_i^*)\Big(d(\hat{x}_i, Qx_i^*) - \frac{\sqrt{\log(n)}}{n}\Big) - C''\sqrt{\frac{\log^3(n)}{n}} \\ & -C\left\{1 + n\left[\alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}}\right] d(\hat{x}_i, Qx_i^*)\right\}\;,\end{aligned}$$

for n *large enough.*

These two lemmas enforce that, with probability higher than $1-1/n^3$,

$$d(\hat{x}_i, Qx_i^*) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*) + \mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{\log(n)/n}\;. \tag{4.25}$$

Indeed, assume that $d(\hat{x}_i, Qx_i^*) \geq C'\big(\alpha(\mathbf{x}^*) + \sqrt{\log(n)/n}\big)$ where the constant C' (only depending on c_l, c_L, c_e) is large enough, then Lemma 4.A.10 implies that $\tilde{L}_i \gtrsim_{c_l,c_L,c_e} nd^2(\hat{x}_i, Qx_i^*)$. Together with Lemma 4.A.9, we deduce that

$$d(\hat{x}_i, Qx_i^*) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*) + \mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{\log(n)/n}\;.$$

In any case, we conclude that

$$d(\hat{x}_i, Qx_i^*) \lesssim_{c_l,c_L,c_e} \alpha(\mathbf{x}^*) + \mu_{\mathbf{x},\mathbf{x}^*}(Q) + \sqrt{\log(n)/n}\;.$$

Taking the union bound over all i for (4.25) concludes the proof.

4.A.2.1 Proof of Lemma 4.A.9

We start from

$$\widetilde{L}_i = \sum_{j=1}^{n} f(x_i^*, x_j^*)\Big(d(Q^{-1}\hat{x}_i, x_j^{**}) - d(x_i^*, x_j^{**})\Big)\;. \tag{4.26}$$

In order to come back to the setting of Lemma 4.A.1, we replace $f(x_i^*, x_j^*)$ by $f(x_i^*, x_j^{**})$ by using the bi-Lipschitz condition (4.1), so that

$$f(x_i^*, x_j^*) - f(x_i^*, x_j^{**}) \lesssim_{c_L,c_e} \alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}}\;.$$

By triangular inequality, we have $d(Q^{-1}\hat{x}_i, x_j^{**}) - d(x_i^*, x_j^{**}) \leq d(Q^{-1}\hat{x}_i, x_i^*)$ which implies

$$\sum_{j=1}^{n}\Big(f(x_i^*, x_j^*) - f(x_i^*, x_j^{**})\Big)\Big(d(Q^{-1}\hat{x}_i, x_j^{**}) - d(x_i^*, x_j^{**})\Big) \lesssim_{c_L,c_e} d(\hat{x}_i, Qx_i^*)r_n\;,$$

where $r_n := \big(n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)}\big)$. We have

$$\widetilde{L}_i \lesssim_{c_L,c_e} \sum_{j=1}^{n} f(x_i^*, x_j^{**})\Big(d(Q^{-1}\hat{x}_i, x_j^{**}) - d(x_i^*, x_j^{**})\Big) + d(\hat{x}_j, Qx_j^*)r_n.$$

Since x_j^{**} now runs over $\mathcal{P}_n$, we can replace the sum over x_j^{**} by a sum over $Q^{-1}x_j$, by using a permutation.

The remainder of the proof follows the same lines as for Lemma 4.A.2, except for a small difference: in Lemma 4.A.2, we had $Qx_i^* \in \mathcal{C}_n$, which is not the case here. We explain below how to handle this minor change.

In order to use the minimality of the estimator $\hat{x}_i$ over $\mathcal{C}_n$, that is,

$$\langle A_i^{(2)}, D(\hat{x}_i, \mathbf{x})\rangle \leq \langle A_i^{(2)}, D(Qx_i^*, \mathbf{x})\rangle \ ,$$

we need $Qx_i^* \in \mathcal{C}_n$. Since $Qx_i^* \notin \mathcal{C}_n$ here, we replace x_i^* with a closest element y_i^* in $\mathcal{C}_n$. It satisfies $d(x_i^*, y_i^*) \leq 2\pi/n$ and $Qy_i^* \in \mathcal{C}_n$. This leads us to

$$\langle A_i^{(2)}, D(\hat{x}_i, \mathbf{x})\rangle \leq \langle A_i^{(2)}, D(Qy_i^*, \mathbf{x})\rangle \leq \langle A_i^{(2)}, D(Qx_i^*, \mathbf{x})\rangle + \langle A_i^{(2)}, D(Qy_i^*, \mathbf{x}) - D(Qx_i^*, \mathbf{x})\rangle \ .$$

Since $|d(Qy_i^*, x_i) - d(Qx_i^*, x_i)| \leq 2\pi/n$ and $|f(x,y)| \leq 1$ the above additional error satisfies

$$\begin{aligned}\langle A_i^{(2)}, D(Qy_i^*, \mathbf{x}) - D(Qx_i^*, \mathbf{x})\rangle &\leq \langle F_i, D(Qy_i^*, \mathbf{x}) - D(Qx_i^*, \mathbf{x})\rangle + \langle E_i^{(2)}, D(Qy_i^*, \mathbf{x}) - D(Qx_i^*, \mathbf{x})\rangle \\ &\leq 2\pi + c'\sqrt{\log(n)}\|D(Qy_i^*, \mathbf{x}) - D(Qx_i^*, \mathbf{x})\|_2 \ , \\ &\leq c'' \ ,\end{aligned}$$

with probability at least $1 - 1/(2n^3)$. The remainder of the proof is the same as for Lemma 4.A.2.

4.A.2.2 Proof of Lemma 4.A.10.

We define $y_i^* \in \mathcal{C}_n$ (respectively $\hat{y}_i \in \mathcal{C}_n$) as a closest point to x_i^* (respectively $\hat{x}_i$). Although the introduction of $\hat{y}_i$ is superfluous for the current estimator $\hat{x}_i$ (since it already belongs to $\mathcal{C}_n$), we still use $\hat{y}_i$ to show that it is possible to keep our theoretical result even with a change in the domain of the estimator.

Introduce the quantity

$$L_i' = \sum_{j=1}^{n} f(y_i^*, x_j^{**})\big(d(Q^{-1}\hat{y}_i, x_j^{**}) - d(y_i^*, x_j^{**})\big) \ ,$$

which has the same properties as the L_i used in Lemma 4.A.1, since each point involved in the expression of L_i' is an element of $\mathcal{C}_n$, and the sum runs over a vector in $\mathcal{P}_n$. This allows us to invoke Lemma 4.A.3 (from the proof of Lemma 4.A.1) which gives

$$L_i' \geq c'nd(\hat{y}_i, Qy_i^*)\Big(c_l d(\hat{y}_i, Qy_i^*) - \epsilon_n\Big) - \frac{c''c_e^3}{c_l^2}\sqrt{\frac{\log^3(n)}{n}} \ ,$$

for some numerical constant $c' > 0$.

By definition of y_i^* and $\hat{y}_i$, we know that $d(y_i^*, x_i^*) \vee d(\hat{y}_i, \hat{x}_i) \leq 2\pi/n$. Hence, by triangular inequality, $d(\hat{y}_i, Qy_i^*) \geq d(\hat{x}_i, Qx_i^*) - 4\pi/n$, where we use the fact that Q preserves the distances. Then, we derive that

$$L_i' \geq C' n d(\hat{x}_i, Qx_i^*)\Big(d(\hat{x}_i, Qx_i^*) - \sqrt{\frac{\log(n)}{n}}\Big) - C''\sqrt{\frac{\log^3(n)}{n}} \ ,$$

where C' and C'' only depend c_l and c_e. Next, we rely on the following lemma to replace L_i' by $\widetilde{L}_i$.

Lemma 4.A.11 *We have*

$$|\widetilde{L}_i - L_i'| \lesssim_{c_L, c_e} 1 + \big(n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)}\big) d(\hat{x}_i, Qx_i^*) \ .$$

Gathering these two bounds completes the proof. □

Proof of Lemma 4.A.11. From the definition of y_i^* and $\hat{y}_i$, and the triangular inequality, we have

$$\Big|\big(d(Q^{-1}\hat{x}_i, x_j^{**}) - d(x_i^*, x_j^{**})\big) - \big(d(Q^{-1}\hat{y}_i, x_j^{**}) - d(y_i^*, x_j^{**})\big)\Big| \leq 4\pi/n \ .$$

This allows us to deduce that the quantity

$$L_i'' = \sum_{j=1}^{n} f(x_i^*, x_j^*)\big(d(Q^{-1}\hat{y}_i, x_j^{**}) - d(y_i^*, x_j^{**})\big) \tag{4.27}$$

satisfies $|\widetilde{L}_i - L_i''| \leq 4\pi$. Besides, we deduce from the Bi-lipschitz condition (4.1) and the triangular inequality that

$$|f(x_i^*, x_j^*) - f(y_i^*, x_j^{**})| \leq c_L\left(d(x_i^*, y_i^*) + d(x_j^*, x_j^{**})\right) + 2\epsilon_n \lesssim_{c_L} \alpha(\mathbf{x}^*) + \epsilon_n \ .$$

Then, we deduce that

$$|L_i' - L_i''| \lesssim_{c_L} \big(\alpha(\mathbf{x}^*) + \epsilon_n\big)\sum_{j=1}^{n}\big(d(Q^{-1}\hat{y}_i, x_j^{**}) - d(y_i^*, x_j^{**})\big) \lesssim_{c_L} n d(Q^{-1}\hat{y}_i, y_i^*)\big(\alpha(\mathbf{x}^*) + \epsilon_n\big) \ .$$

All in all, we have

$$|\widetilde{L}_i - L_i'| \leq |\widetilde{L}_i - L_i''| + |L_i'' - L_i'| \lesssim_{c_L, c_e} 1 + d(Q^{-1}\hat{y}_i, y_i^*)\big(n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)}\big) \ .$$

□

4.A.3 Proof of Proposition 4.3.2

4.A.3.1 Main Arguments

We shall establish that the estimator $\hat{\mathbf{x}}'$ is such that $\hat{\mathbf{x}}'^T\hat{\mathbf{x}}'$ is close to $\mathbf{x}^{*T}\mathbf{x}^*$. In other words, the distances between the $\hat{x}'_1, \dots, \hat{x}'_n$ are close to the true distances between the $x^*_1, \dots, x^*_n$. Then, relying on a recent matrix perturbation result from [Arias-Castro et al., 2020], we deduce that $\|\hat{\mathbf{x}}' - \mathbf{x}^*\|_2$ is small. Let us first state this perturbation result. Given any matrix M with real coefficients, denote its transpose by M^T, and the Moore-Penrose pseudo-inverse by $M^\dagger$, and the usual operator norm by $\|M\|_{op}$. We recall that $\mathcal{O}$ stands for the collection of orthogonal matrices of size 2×2.

Proposition 4.A.12 (Theorem 1 in [Arias-Castro et al., 2020]) *For two tall matrices M and N of same size, with N having full rank, let $\nu = \|MM^T - NN^T\|_2$. Then we have*

$$\min_{Q \in \mathcal{O}} \|M - NQ\|_2 \lesssim \nu \|N^\dagger\|_{op} \ ,$$

as soon as $2\nu\|N^\dagger\|^2_{op} \le 1$.

Let $\mathbf{x}^{**} \in \mathcal{P}_n$ denote a closest approximation of $\mathbf{x}^*$ in $\mathcal{P}_n$, that is, such that $d_\infty(\mathbf{x}^*, \mathbf{x}^{**}) = \alpha(\mathbf{x}^*)$. In order to invoke the above theorem for $M = \hat{\mathbf{x}}'^T$ and $N = \mathbf{x}^{**T}$ in $\mathbb{R}^{n\times 2}$, we check that the condition $2\nu\|N^\dagger\|^2_{op} \le 1$ is fulfilled. First, we work out

$$N^\dagger = (N^T N)^{-1} N^T = \left(\mathbf{x}^{**}\mathbf{x}^{**T}\right)^{-1} \mathbf{x}^{**} = \frac{\mathbf{x}^{**}}{n} \ ,$$

so that $\|N^\dagger\|^2_{op} \le \|N^\dagger\|^2_2 = 1/n$. The main part of the proof consists in bounding the perturbation $\nu = \|\hat{\mathbf{x}}'^T\hat{\mathbf{x}}' - \mathbf{x}^{**T}\mathbf{x}^{**}\|_2$.

Lemma 4.A.13 *With probability at least $1 - 1/n^2$, we have*

$$\nu \lesssim_{c_l, c_L, c_e} n\alpha(\mathbf{x}^*) + \sqrt{n \log(n)} \ .$$

Then we have

$$2\nu\|N^\dagger\|^2_{op} \le C\left(\alpha(\mathbf{x}^{**}) + \sqrt{\frac{\log(n)}{n}}\right) \ ,$$

where C only depends on c_l, c_L, and c_e. If $\alpha(\mathbf{x}^{**}) + \sqrt{\log(n)/n} \ge 1/C$, then Proposition 4.3.2 becomes trivial since $d_{1,\mathcal{O}}(\hat{\mathbf{x}}', \mathbf{x}^*) \le n$. Hence, it suffices to consider the case where $\alpha(\mathbf{x}^{**}) + \sqrt{\log(n)/n} \le 1/C$ so that the condition of Proposition 4.A.12 is fulfilled, which implies

$$\min_{Q \in \mathcal{O}} \|\hat{\mathbf{x}}'^T - \mathbf{x}^{**T}Q\|_2 \le C\left(\sqrt{n}\alpha(\mathbf{x}^*) + \sqrt{\log(n)}\right) \ ,$$

and so

$$\min_{Q \in \mathcal{O}} \|\hat{\mathbf{x}}' - Q\mathbf{x}^{**}\|_1 \le C\left(n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)}\right) \ .$$

Besides, the orthogonal transformation Q preserves the distances and the l_1 and l_2 distances are equivalent on $\mathbb{R}^2$, so we have $\|Q\mathbf{x}^{**} - Q\mathbf{x}^*\|_1 \lesssim d_1(Q\mathbf{x}^{**}, Q\mathbf{x}^*) = d_1(\mathbf{x}^{**}, \mathbf{x}^*)$. Then, using

the definition of $\alpha(\mathbf{x}^*)$, we get $\|Q\mathbf{x}^{**} - Q\mathbf{x}^*\|_1 \lesssim n\alpha(\mathbf{x}^*)$. Together with the triangular inequality, this leads us to

$$\min_{Q\in\mathcal{O}} \|\hat{\mathbf{x}}' - Q\mathbf{x}^*\|_1 \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)} \ .$$

Using again the equivalence between norms, that is $d(x,y) \lesssim \|x-y\|_1$ for all x and y in $\mathcal{C}$, we conclude that

$$d_{1,\mathcal{O}}(\mathbf{x},\mathbf{y}) \lesssim_{c_l,c_L,c_e} \min_{Q\in\mathcal{O}} \|\hat{\mathbf{x}}' - Q\mathbf{x}^*\|_1 \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)} \ ,$$

and the proof of Proposition 4.3.2 is complete.

4.A.3.2 Proof of Lemma 4.A.13.

The vectors of points $\hat{\mathbf{x}}'$ and $\mathbf{x}^{**}$ are elements of $\mathcal{P}_n \subset \mathbb{R}^{2\times n}$, hence they both satisfy $\hat{\mathbf{x}}'1 = 0$ and $\mathbf{x}^{**}1 = 0$ (where 1 denotes the vector of ones). We can then invoke the next lemma to bound $\nu = \|\hat{\mathbf{x}}'^T\hat{\mathbf{x}}' - \mathbf{x}^{**T}\mathbf{x}^{**}\|_2$.

Lemma 4.A.14 *For any vectors of centered points $Z = (z_1,\dots,z_n)$ and $Z' = (z'_1,\dots,z'_n)$ in $\mathbb{R}^{2\times n}$ with $Z1 = Z'1 = 0$, let $D = (D_{ij})$ and $D' = (D'_{ij})$ be their (squared) distance matrices, that is $D_{ij} = \|z_i - z_j\|_2^2$ and $D'_{ij} = \|z'_i - z'_j\|_2^2$ for all $i,j \in [n]$. Then we have*

$$\|Z^TZ - Z'^TZ'\|_2 \le \|D - D'\|_2 \ .$$

For $Z = \mathbf{x}^{**}$ and $Z' = \hat{\mathbf{x}}'$, and accordingly $D_{ij} = \|x_i^{**} - x_j^{**}\|_2^2$ and $D'_{ij} = \|\hat{x}'_i - \hat{x}'_j\|_2^2$, it follows from Lemma 4.A.14 that $\nu \le \|D - D'\|_2$. Since all square distances D_{ij} and D'_{ij} are at most equal to 4, we get

$$\nu \le 4\|\sqrt{D} - \sqrt{D'}\|_2 \ ,$$

where $\sqrt{D}$ and $\sqrt{D'}$ denote the matrices of coefficients $\sqrt{D_{ij}} = \|x_i^{**} - x_j^{**}\|_2$ and $\sqrt{D'_{ij}} = \|\hat{x}'_i - \hat{x}'_j\|_2$.

For any $x,y \in \mathcal{C}$, elementary geometry gives $\|x-y\|_2 = 2\sin(d(x,y)/2)$. Since the sinus function is 1-Lipschitz, we have

$$\big|\, \|x-y\|_2 - \|x'-y'\|_2 \big| \le |d(x,y) - d(x',y')| \ ,$$

for any $x,y,x',y' \in \mathcal{C}$. Hence, we deduce that $\nu \le 4\|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\|_2$, where $D(\mathbf{x}^{**})$ and $D(\hat{\mathbf{x}}')$ respectively denote the matrices of coefficients $d(x_i^{**}, x_j^{**})$ and $d(\hat{x}'_i, \hat{x}'_j)$. As a consequence, we mainly have to control with high probability the deviations of this matrix norm.

Lemma 4.A.15 *With probability at least $1 - 1/n^2$, we have*

$$\|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\|_2 \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)} \ .$$

Hence $\nu \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*) + \sqrt{n\log(n)}$ and the proof of Lemma 4.A.13 is complete. □

Proof of Lemma 4.A.14. Let $H = I - J/n$, where I is the identity and J the matrix of ones. Since $Z1 = 0$, we have $ZH = Z$, so that

$$Z^T Z = HZ^T ZH = -\frac{1}{2}HDH,$$

since D is the matrix of distances associated with Z. Then we have

$$\|Z^T Z - Z'^T Z'\|_2 = \frac{1}{2}\|H(D - D')H\|_2 \leq \frac{1}{2}\|D - D'\|_2 \ ,$$

where the last inequality derives from the general relation $\|AB\|_2 \leq \|A\|_{op}\|B\|_2$ for any matrices A, B, and the fact that $\|H\|_{op} = 1$ (because H is an orthogonal projection). Lemma 4.A.14 is proved. □

Proof of Lemma 4.A.15. First, we come back to the definition of the estimation $\hat{\mathbf{x}}'$ defined in (4.10). We have $\langle A^{(1)}, D(\hat{\mathbf{x}}')\rangle \leq \langle A^{(1)}, D(\mathbf{x}^{**})\rangle$, which in turn implies that

$$\langle F, D(\hat{\mathbf{x}}') - D(\mathbf{x}^{**})\rangle \leq \langle E^{(1)}, D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\rangle.$$

As in the last lines of the proof of Lemma 4.A.2, we bound the term $\langle E^{(1)}, D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\rangle$ by an union bound over all possible vectors $\hat{\mathbf{x}}'$. Hence, we get

$$\langle F, D(\hat{\mathbf{x}}') - D(\mathbf{x}^{**}) \leq \langle E^{(1)}, D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\rangle \lesssim \sqrt{n\log(n)}\|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\|_2 \ , \qquad (4.28)$$

with probability at least $1 - 1/n^2$. Conversely, we shall lower bound $\langle F, D(\hat{\mathbf{x}}') - D(\mathbf{x}^{**})\rangle$.

$$\langle F, D(\hat{\mathbf{x}}') - D(\mathbf{x}^{**})\rangle = \sum_{i,j=1}^{n} f(x_i^*, x_j^*)\Big(d(\hat{x}_i', \hat{x}_j') - d(x_i^{**}, x_j^{**})\Big) \ .$$

Using the bi-Lipschitz property of the function f, we deduce that

$$\big|f(x_i^*, x_j^*) - f(x_i^{**}, x_j^{**})\big| \leq c_L\left(d(x_i^*, x_i^{**}) + d(x_j^*, x_j^{**})\right) + 2c_e\sqrt{\log(n)/n} \leq 2c_L\alpha(\mathbf{x}^*) + 2c_e\sqrt{\log(n)/n} \ ,$$

and

$$\begin{aligned}\sum_{i,j=1}^{n} \big|f(x_i^*, x_j^*) - f(x_i^{**}, x_j^{**})\big|\big|d(\hat{x}_i', \hat{x}_j') - d(x_i^{**}, x_j^{**})\big| \\ \lesssim_{c_L,c_e} \left(\alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}}\right)\|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\|_1 \\ \lesssim_{c_L,c_e} n\left(\alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}}\right)\|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\|_2 \ ,\end{aligned}$$

where we applied Cauchy-Schwarz inequality on $\mathbb{R}^{n\times n}$. As a consequence,

$$\begin{aligned}\langle F, D(\hat{\mathbf{x}}') - D(\mathbf{x}^{**})\rangle \geq \sum_{i,j=1}^{n} f(x_i^{**}, x_j^{**})\Big(d(\hat{x}_i', \hat{x}_j') - d(x_i^{**}, x_j^{**})\Big) \\ - Cn\left(\alpha(\mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}}\right)\|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}')\|_2 \ ,\end{aligned} \qquad (4.29)$$

where C only depends on c_L and c_e.

Lemma 4.A.16 *We have*

$$\sum_{i,j=1}^{n} f(x_i^{**},x_j^{**})\Big(d(\hat{x}_i',\hat{x}_j')-d(x_i^{**},x_j^{**})\Big) \geq \frac{c_l}{2}\|D(\mathbf{x}^{**})-D(\hat{\mathbf{x}}')\|_2^2-c_e\sqrt{n\log(n)}\|D(\mathbf{x}^{**})-D(\hat{\mathbf{x}}')\|_2 \,.$$

We conclude from (4.28) and the above lemma that

$$c\sqrt{n\log(n)}\|D(\mathbf{x}^{**})-D(\hat{\mathbf{x}}')\|_2 \geq \frac{c_l}{2}\|D(\mathbf{x}^{**})-D(\hat{\mathbf{x}}')\|_2^2 - C\left(n\alpha(\mathbf{x}^*)+\sqrt{n\log(n)}\right)\|D(\mathbf{x}^{**})-D(\hat{\mathbf{x}}')\|_2 \;,$$

where C only depends on c_L and c_e whereas c is a numerical constant. This leads us to

$$\|D(\mathbf{x}^{**})-D(\hat{\mathbf{x}}')\|_2 \lesssim_{c_l,c_L,c_e} n\alpha(\mathbf{x}^*)+\sqrt{n\log(n)} \;.$$

Lemma 4.A.15 is proved. □

Proof of Lemma 4.A.16. To alleviate the notation, we introduce $\gamma_i = \sum_{j=1}^n f(x_i^{**},x_j^{**})[d(\hat{x}_i',\hat{x}_j')-d(x_i^{**},x_j^{**})]$ so that we aim at establishing a lower bound for each γ_i and in turn for $\gamma=\sum_{i=1}^n\gamma_i$. To simplify the arguments, we only consider the case where n is odd, the case of n even being almost similar. To control this quantity, we shall rely on the second Bi-lipschitz condition (4.2) to get a lower bound of differences of the form $f(x_i^{**},x_j^{**})-f(x_i^{**},x_k^{**})$. Unfortunately, this Bi-lipschitz condition only allows to control these differences for $d(x_i^{**},x_k^{**})\geq d(x_i^{**},x_j^{**})$. This is why we need to carefully match indices j and k in order to only consider such differences.

Since both $\mathbf{x}^{**}$ and $\hat{\mathbf{x}}'$ belongs to $\mathcal{P}_n$, we shall heavily rely on the symmetry of the problem. Assume without loss of generality that $i=1$ and $x_j^{**}=e^{\iota 2\pi(j-1)/n}$. Then, $d(x_i^{**},x_j^{**})=\frac{2\pi}{n}[|j-1|\wedge|n-j+1|]$. Since $\hat{\mathbf{x}}'$ also belongs to $\mathcal{P}_n$, there exists a permutation σ of $[n]$ such that $\sigma(1)=1$ and $d(\hat{x}_i',\hat{x}_j')=\frac{2\pi}{n}[|\sigma(j)-1|\wedge|n+1-\sigma(j)|]$. Recall that we consider the case where n is odd. Therefore, there exists a surjective map $\overline{\sigma}:[n-1]\mapsto[\lfloor n/2\rfloor]$ such that $|\overline{\sigma}^{-1}(\{z\})|=2$ for any $z\in[\lfloor n/2\rfloor]$ and $d(\hat{x}_i',\hat{x}_j')=\frac{2\pi}{n}\overline{\sigma}(j-1)$ for any $j=2,\dots n$. Finally, we write $\psi_j=f(1,e^{\iota 2\pi j/n})$ and $\psi_j'=f(1,e^{-\iota 2\pi j/n})$ for $j=1,\dots\lfloor n/2\rfloor$. Equipped with this new notation, we arrive at

$$\gamma_i=\frac{2\pi}{n}\sum_{j=1}^{\lfloor n/2\rfloor}\psi_j(\overline{\sigma}(j)-j)+\psi_j'(\overline{\sigma}(n-j)-j)$$

Finally, we denote $a_j=\overline{\sigma}(j)-j$ and $a_j'=\overline{\sigma}(n-j)-j$ for $j=1,\dots,\lfloor n/2\rfloor$. Obviously, we have $\sum_{j=1}^{\lfloor n/2\rfloor}a_j+a_j'=0$. More generally, one easily check that for any positive integer $s\leq\lfloor n/2\rfloor$, the sum $\sum_{j=1}^s(a_j+a_j')$ is non-negative. Starting from

$$\gamma_i=\frac{2\pi}{n}\sum_{j=1}^{\lfloor n/2\rfloor}\psi_j a_j+\psi_j' a_j'$$

we partition the indices according to the signs of a_j and a_j'. Define $A_+=\{j\in[\lfloor n/2\rfloor]:a_j\geq 0\}$, $A_-=\{j\in[\lfloor n/2\rfloor]:a_j<0\}$, $A_+'=\{j\in[\lfloor n/2\rfloor]:a_j'\geq 0\}$, and $A_-'=\{j\in[\lfloor n/2\rfloor]:a_j'<$

$0\}$. Intuitively, we want to group indices j such that $a_j > 0$ with indices k such that $a_k < 0$. This can be done by recursion. First, consider the smallest index $k \in A_- \cup A'_-$. By symmetry, suppose that $a_k < 0$. Since $\sum_{j=1}^{k}(a_j + a'_j) \geq 0$, this implies that $\sum_{j=1}^{k} \mathbf{1}_{j\in A_+} a_j + \mathbf{1}_{j\in A'_+} a'_j \geq |a_k| + \mathbf{1}_{k\in A'_-}|a'_k|$, hence it is possible to build non-negative numbers $b_{j,k,1} \leq a_j$ for $j \in A_+ \cap [k]$ and $b'_{j,k,1} \leq a'_j$ for $j \in A'_+ \cap [k]$ such that $\sum_{j=1}^{k} \mathbf{1}_{j\in A_+} b_{j,k,1} + \mathbf{1}_{j\in A'_+} b'_{j,k,1} = |a_k|$. Iterating the construction we obtain the following decomposition

$$\begin{aligned}\frac{n}{2\pi}\gamma_i &= \sum_{j\in A_+}\left(\sum_{k\in A_-}(\psi_j-\psi_k)b_{j,k,1}+\sum_{k\in A'_-}(\psi_j-\psi'_k)b_{j,k,2}\right)\\ &\quad+\sum_{j\in A'_+}\left(\sum_{k\in A_-}(\psi'_j-\psi_k)b'_{j,k,1}+\sum_{k\in A'_-}(\psi'_j-\psi'_k)b'_{j,k,2}\right)\ ,\end{aligned}$$

where all $b_{j,k,t}$'s are non-negative, $b_{j,k,t} = 0$ for $k < j$, and

$$\left\{\begin{array}{l}\sum_{k\in A_-} b_{j,k,1}+\sum_{k\in A_+} b_{j,k,2} = a_j \text{ for } j\in A_+\ ;\\ \sum_{k\in A_-} b'_{j,k,1}+\sum_{k\in A_+} b'_{j,k,2} = a'_j \text{ for } j\in A'_+\ ;\\ \sum_{j\in A_+} b_{j,k,1}+\sum_{j\in A'_+} b'_{j,k,1} = -a_k \text{ for } k\in A_-\ ;\\ \sum_{j\in A_+} b_{j,k,2}+\sum_{j\in A'_+} b'_{j,k,2} = -a'_k \text{ for } k\in A'_-\ .\end{array}\right.$$

In the above decomposition each term $b_{j,k,1}$, $b_{j,k,2}$, $b'_{j,k,1}$, and $b'_{j,k,2}$ is non-negative. Besides, it is not equal to zero only when $k \geq j$, so that we can use the bi-Lipschitz condition (4.2)

$$(\psi_j - \psi'_k) = f(1, e^{\iota 2\pi(j-1)/n}) - f(1, e^{\iota 2\pi(k-1)/n}) \geq c_l \frac{2\pi(k-j)}{n} - c_e\sqrt{\frac{\log(n)}{n}}\ ,$$

We obtain similarly the same lower bound for $\psi_j - \psi_k$, $\psi'_j - \psi_k$, and $\psi'_j - \psi'_k$. Coming back to the expression of γ_j and the definition of the $b_{i,j,t}$ with $t = 1, 2$ yields

$$\begin{aligned}\frac{n}{2\pi}\gamma_i &\geq c_l\frac{2\pi}{n}\sum_{j=1}^{\lfloor n/2\rfloor} -j\left[a_j + a'_j\right] - c_e\sqrt{\frac{\log(n)}{n}}\sum_{j=1}^{\lfloor n/2\rfloor}|a_j| + |a'_j|\\ &\geq -c_l\frac{2\pi}{n}\sum_{j=1}^{\lfloor n/2\rfloor}\left[j(\overline{\sigma}(j)-j) + j(\overline{\sigma}(n-j)-j)\right]\\ &\quad - c_e\sqrt{\frac{\log(n)}{n}}\sum_{j=1}^{\lfloor n/2\rfloor}|\overline{\sigma}(j)-j| + |\overline{\sigma}(n-j)-j|\ .\end{aligned}$$

Let us work out these two expressions in the rhs. By symmetry and definition $\overline{\sigma}$ and σ we get

$$\begin{aligned}\sum_{j=1}^{\lfloor n/2\rfloor} -j(\overline{\sigma}(j)-j) + j(\overline{\sigma}(n-j)-j) &= \frac{1}{2}\sum_{j=1}^{\lfloor n/2\rfloor}(\overline{\sigma}(j)-j)^2 + (\overline{\sigma}(n-j)-j)^2\\ &= \frac{n^2}{8\pi^2}\sum_{j=1}^{n}\left[d(\hat{x}'_i,\hat{x}'_j) - d(x_i^{**},x_j^{**})\right]^2\ .\end{aligned}$$

Similarly, we get

$$\sum_{j=1}^{\lfloor n/2 \rfloor} |\sigma(j) - j| + |\overline{\sigma}(n-j) - j| = \frac{n}{2\pi} \sum_{j=1}^{n} |d(\hat{x}'_i, \hat{x}'_j) - d(x_i^{**}, x_j^{**})| \ .$$

Putting everything together yields

$$\gamma_i \geq \frac{c_l}{2} \sum_{j=1}^{n} \left[d(\hat{x}'_i, \hat{x}'_j) - d(x_i^{**}, x_j^{**}) \right]^2 - c_e \sqrt{\frac{\log(n)}{n}} \sum_{j=1}^{n} |d(\hat{x}'_i, \hat{x}'_j) - d(x_i^{**}, x_j^{**})| \ ,$$

which in turn allows us to conclude

$$\begin{aligned} \gamma &\geq \frac{c_l}{2} \|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}^{**})\|_2^2 - c_e \sqrt{\log(n)/n} \|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}^{**})\|_1 \\ &\geq \frac{c_l}{2} \|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}^{**})\|_2^2 - c_e \sqrt{n \log(n)} \|D(\mathbf{x}^{**}) - D(\hat{\mathbf{x}}^{**})\|_2 \ . \end{aligned}$$

□

4.A.4 Proof of Corollary 4.3.4

We only focus on the case of random latent points $\mathbf{x}^*$ as the case of deterministic points is a straightforward Corollary of Theorem 4.3.3. Assume henceforth that $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ is a random sample drawn from the uniform distribution on $\mathcal{C}$. Conditionally to $\mathbf{x}^*$, note that Theorem 4.3.3 gives

$$d_{\infty,\mathcal{O}}(\hat{\mathbf{x}}, \mathbf{x}^*) \lesssim_{c_l, c_L, c_e} \min_{\mathbf{x} \in \mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}^*) + \sqrt{\frac{\log(n)}{n}}$$

with probability at least $1 - 2/n^2$. Hence, it suffices to show that, with high probability,

$$\alpha(\mathbf{x}^*) := \min_{\mathbf{x} \in \mathcal{P}_n} d_\infty(\mathbf{x}, \mathbf{x}^*) \lesssim \sqrt{\frac{\log(n)}{n}}.$$

In other words, we need to prove that $\mathbf{x}^*$ is almost evenly distributed on the circle. This could be done relying on Dvoretzky–Kiefer–Wolfowitz inequality but we use a more direct approach here.

Recall that an interval $I = [a, b]$ denotes the set of points lying between a and b in the one-dimensional torus $\mathbb{R}/(2\pi)$, when following the trigonometric direction from a to b. The length of I is denoted by $|I|$. For any interval I, let N_I be the number of points x_i^* whose argument lies in I. Then, the centered random variablse associated with N_I, that is

$$V_I = N_I - \frac{n|I|}{2\pi} \ ,$$

satisfies the two next claims. Denote $\mathcal{I}$ the set of all intervals $I \subset \mathcal{C}$ of the form $[a, b]$.

Claim 4.A.17 *We have* $n\alpha(\mathbf{x}^*) \leq 4\pi + 2\pi \sup_{I \in \mathcal{I}} |V_I|$, *almost surely.*

Claim 4.A.18 *The following inequality holds with probability at least* $1 - 1/n^2$,

$$\sup_{I \in \mathcal{I}} |V_I| \lesssim \sqrt{n \log(n)} \ .$$

It follows from these claims that $\alpha(\mathbf{x}^*) \lesssim \sqrt{\log(n)/n}$ with probability larger than $1-1/n^2$. The proof is complete.

4.A.4.1 Proof of Claim 4.A.17

Recall that for a vector $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{C}^n$, we say that $\mathbf{x}$ is ordered, if these points are consecutive when one walks on the circle with the trigonometric direction.

Assume that $x_1^*, \dots, x_n^*$ are all distinct points. Otherwise, if some $x_{j_1}^*, \dots, x_{j_s}^*$ are equal, it suffices to consider proxy points $y_{j_1}, \dots, y_{j_s}$ that are all distinct and satisfy $d(y_{j_r}, x_{j_r}^*) \leq 1/n^2$ for all $r \in [s]$.

Without loss of generality, assume that the identity permutation is a latent order, that is $x_1^*, \dots, x_n^*$ is ordered.

Let $\mathbf{x}^{**} = (x_1^{**}, \dots, x_n^{**})$ be a vector of $\mathcal{P}_n$ defined as follows. The $x_1^{**} \in \mathcal{C}_n$ is a closest point to x_1^* with respect to the distance d, and the successive points x_{j+1}^{**} are elements of $\mathcal{C}_n$ with the following arguments

$$\underline{x}_{j+1}^{**} = \underline{x}_1^{**} + j\frac{2\pi}{n} \pmod{2\pi} \ ,$$

for $j = 1, \dots, n-1$.

Let $i \in \{1, \dots, n\}$, and consider the intervals $I_i = [\underline{x}_1^*, \underline{x}_i^*]$ and $I_i' = [\underline{x}_1^*, \underline{x}_i^{**}]$. The difference between their lengths gives the following bound

$$d(x_i^*, x_i^{**}) \leq \big| |I_i| - |I_i'| \big| \ . \tag{4.30}$$

Observe that $N_{I_i} = i$ since $x_1^*, \dots, x_n^*$ are ordered and all distinct. Hence,

$$\left| \frac{2\pi i}{n} - |I_i| \right| = \left| \frac{2\pi N_{I_i}}{n} - |I_i| \right| = 2\pi \frac{|V_{I_i}|}{n} \leq 2\pi \sup_{I \in \mathcal{I}} \frac{|V_I|}{n} \ . \tag{4.31}$$

On the other hand, we know that the length of I_i' is bounded by

$$\big|[\underline{x}_1^{**}, \underline{x}_i^{**}]\big| - d(x_1^*, x_1^{**}) \leq |I_i'| \leq \big|[\underline{x}_1^{**}, \underline{x}_i^{**}]\big| + d(x_1^*, x_1^{**}).$$

By construction of the x_j^{**}, we have $d(x_1^*, x_1^{**}) \leq 2\pi/n$ and $\big|[\underline{x}_1^{**}, \underline{x}_i^{**}]\big| = 2\pi(i-1)/n$. Then, it follows from the above line that

$$\left| \frac{2\pi i}{n} - |I_i'| \right| \leq \frac{4\pi}{n} \ .$$

Finally, using the triangular inequality, the last display and (4.31), we get

$$\big| |I_i| - |I_i'| \big| \leq 2\pi \sup_{I \in \mathcal{I}} \frac{|V_I|}{n} + \frac{4\pi}{n} \ .$$

Coming back to (4.30) and taking the supremum over all $i \in [n]$ concludes the proof.

4.A.4.2 Proof of Claim 4.A.18

The proof is based on an ϵ-net method. Let $\mathcal{I}_n$ be a discretization of $\mathcal{I}$, defined as the collection of intervals $I_n = [a_n, b_n]$ with $a_n, b_n \in \{2\pi i/n;\ i \in [n]\}$. Observe that for any $I \in \mathcal{I}$, there exists $I_n \in \mathcal{I}_n$ such that the following decomposition holds,

$$I = I^{(l)} \cup I_n \cup I^{(r)}, \tag{4.32}$$

where $I^{(l)}$ and $I^{(r)}$ are two sub-intervals of $I \setminus I_n$, with lengths smaller than $2\pi/n$.

It follows that $V_I = V_{I^{(l)}} + V_{I_n} + V_{I^{(r)}}$ with respect to the decomposition (4.32). Then, the triangular inequality gives

$$\sup_{I\in\mathcal{I}} |V_I| \leq \sup_{I_n\in\mathcal{I}_n} |V_{I_n}| + 2 \sup_{\substack{I\in\mathcal{I} \\ |I|\leq 2\pi/n}} |V_I| \ . \tag{4.33}$$

For any fixed interval $I_n \in \mathcal{I}_n$, we derive from Hoeffing inequality that $|V_{I_n}| \geq nt$ holds with probability at most $2\exp[-nt^2]$. Since $|\mathcal{I}_n| \leq n^2$, we apply an union bound to derive that

$$\sup_{I_n\in\mathcal{I}_n} |V_{I_n}| \lesssim \sqrt{n\log(n)} \ , \tag{4.34}$$

with probability higher than $1 - 1/n^2$.

Regarding the second term in the bound (4.33), the triangular inequality gives $|V_I| \leq N_I + (n|I|/(2\pi))$, thus leading to

$$\sup_{\substack{I\in\mathcal{I} \\ |I|\leq 2\pi/n}} |V_I| \leq \sup_{\substack{I\in\mathcal{I} \\ |I|\leq 2\pi/n}} N_I + 1 \ .$$

For any interval I of length smaller than $2\pi/n$, there exists $I_n \in \mathcal{I}_n$ such that $I \subset I_n$ and $|I_n| = 4\pi/n$. In particular, $N_I \leq N_{I_n}$, so that

$$\sup_{\substack{I\in\mathcal{I} \\ |I|\leq 2\pi/n}} |V_I| \leq 1 + \sup_{\substack{I_n\in\mathcal{I}_n \\ |I_n|\leq 4\pi/n}} N_{I_n} N_I \leq 3 + \sup_{\substack{I_n\in\mathcal{I}_n \\ |I_n|\leq 4\pi/n}} V_{I_n} \ .$$

Gathering the latter inequality with (4.33) and (4.34) concludes the proof.

4.B Proof of the identifiability results and minimax lower bound

4.B.1 Proof of Proposition 4.2.2

For simplicity, we assume that $n/8$ is an integer in the rest of the example. The construction mainly amounts to contracting the function f in some regions and dilating it in other regions which allows to contracting and dilating the position $\mathbf{x}$.

Consider a partition of the latent space $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$ in three arcs $\mathcal{C}_1 = [\underline{x}_n, \underline{x}_{n/8}] = [0, \pi/4]$, $\mathcal{C}_2 = [\underline{x}_{n/8}, \underline{x}_{n/2}] = [\pi/4, \pi]$ and $\mathcal{C}_3 = [\underline{x}_{n/2}, \underline{x}_n] = [\pi, 2\pi]$. In $\mathcal{C}_1$, set $\tilde{f}_1(x, y) =$

$1 - d(x,y)/\pi$. Let $\tilde{x}_k = e^{\iota k\pi/n}$ for $k \in [n/4]$ and $\tilde{x}_n = x_n = 1$. In other words, we contract the positions. Note that $d_{\infty,\mathcal{O}}(\mathbf{x}, \tilde{\mathbf{x}}) \geq \pi/8$, whatever the definition of other coordinates in $\tilde{\mathbf{x}}$. Then observe that (4.5) is satisfied by $\tilde{f}_1$ for all $i, j \in [n/4] \cup \{0\}$.

In $\mathcal{C}_2$, set $\tilde{f}_2(x,y) = 1 - d(x,y)/(3\pi)$. Let $\tilde{x}_{k+(n/4)} = e^{\iota\pi/4} e^{\iota k 3\pi/n}$ for $k \in [n/4]$. Again, observe that (4.5) holds for $\tilde{f}_2$ and all integers $i, j \in [n/4, n/2]$. Finally in $\mathcal{C}_3$, set $\tilde{f}_3(x,y) = f(x,y)$, and let $\tilde{x}_k = x_k$ for all integers $k \in (n/2, n)$. Obviously, (4.5) is true for $\tilde{f}_3$ and all integers $i, j \in [n/2, n]$.

It remains to deal with the situations where the pairs of points lie in different parts of the partition $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$. In the case where $x \in \mathcal{C}_1$ and $y \in \mathcal{C}_2$, define $\tilde{f}_{1-2}(x,y) = \tilde{f}_1(x, e^{\iota\pi/4}) + \tilde{f}_2(e^{\iota\pi/4}, y) - 1$. For all integers $i \in [0, n/4]$ and $j \in [n/4, n/2]$, one have already seen that $\tilde{f}_1(\tilde{x}_i, e^{\iota\pi/4}) = f(x_i, e^{\iota\pi/2})$ and $\tilde{f}_1(e^{\iota\pi/4}, \tilde{x}_j) = f(e^{\iota\pi/2}, x_j)$. Hence $\tilde{f}_{1-2}(\tilde{x}_i, \tilde{x}_j) = f(x_i, e^{\iota\pi/2}) + f(e^{\iota\pi/2}, x_j) - 1 = f(x_i, x_j)$, that is, $\tilde{f}_{1-2}$ satisfies (4.5) for all integers $i \in [0, n/4]$ and $j \in [n/4, n/2]$.

In the case where $x \in \mathcal{C}_1$ and $y \in \mathcal{C}_3$, define $\tilde{f}_{1-3}(x,y) = \tilde{f}_1(x, e^{\iota 0}) + \tilde{f}_3(e^{\iota 0}, y) - 1$, if the length of the interval $[x, e^{\iota 0}] \cup (e^{\iota 0}, y]$ is less than π; otherwise, $\tilde{f}_{1-3}(x,y) = \tilde{f}_1(x, e^{\iota\pi/4}) + \tilde{f}_2(e^{\iota\pi/4}, e^{\iota\pi}) + \tilde{f}_3(e^{\iota\pi}, y) - 2$. Since f admits similar decompositions, one can deduce from the above that (4.5) is valid for all integers $i \in [0, n/4]$ and $j \in [n/2, n]$.

The remaining cases can be handled in the same manner. Finally, define $\tilde{f}$ as the composition of the above functions $\tilde{f}_1$, $\tilde{f}_2$, $\tilde{f}_3$, $\tilde{f}_{1-2}$, $\tilde{f}_{1-3}$, and $\tilde{f}_{2-3}$ (whose supports are disjoints). Then, we readily check that $f \in \mathcal{BL}[(3\pi)^{-1}, \pi^{-1}, 0]$ and that (4.5) is satisfied for all $i, j \in [n]$. □

4.B.2 Proof of Theorem 4.3.5

We will establish the lower bound $\sqrt{\log(n)/n}$ in the particular setting where the observations A_{ij} are independent Bernoulli random variables of parameters $F_{ij} = f_0(x_i, x_j)$, for the specific function

$$f_0(x_i, x_j) = (3/4) - d(x_i, x_j)/(4\pi), \tag{4.35}$$

with $\mathbf{x} = (x_1, \ldots, x_n) \in S$. The corresponding probability distribution is denoted by $\mathbb{P}_{(\mathbf{x}, f_0)}$. There exists $\mathbf{y} \in \mathcal{P}_n$ such that $\max_{1 \leq j \leq n} d(y_j, x_j^{(p)}) \leq \pi/16$ and

$$\|\mathbf{x}^{(p)} - \mathbf{y}\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) + 1 \ .$$

This minimax lower bound is based on Fano's method as stated below. Given two configuration $\mathbf{x}$ and $\mathbf{x}'$ in S, we denote the Kullback-Leibler divergence of $\mathbb{P}_{(\mathbf{x}, f_0)}$ and $\mathbb{P}_{(\mathbf{x}', f_0)}$ by $KL(\mathbb{P}_{(\mathbf{x}, f_0)} \,\|\, \mathbb{P}_{(\mathbf{x}', f_0)})$. Given $\mathbf{x}$ and $\mathbf{y}$ in S, we consider the pseudometric $\rho(\mathbf{x}, \mathbf{y}) = d_{\infty,\mathcal{O}}(\mathbf{x}, \mathbf{y})$. Given $\epsilon > 0$, the packing number $\mathcal{M}(\epsilon, S', \rho)$ is defined as the largest number of points in S' that are at least ϵ away from each other with respect to ρ. Below, we state a specific version of Fano's lemma.

Proposition 4.B.1 (from [Yu, 1997]) *Consider any subset $S' \subset S$. Define the Kullback-Leibler diameter of S' by*

$$d_{KL}(S') = \sup_{\mathbf{x}, \mathbf{x}' \in S'} KL(\mathbb{P}_{(\mathbf{x}, f_0)} \,\|\, \mathbb{P}_{(\mathbf{x}', f_0)}) \ .$$

Then, for any estimator $\hat{\mathbf{x}}$ *and for any* $\epsilon > 0$*, we have*

$$\sup_{\mathbf{x}\in S'} \mathbb{P}_{(\mathbf{x},f_0)}\left[\rho(\hat{\mathbf{x}},\mathbf{x}) \geq \frac{\epsilon}{2}\right] \geq 1 - \frac{d_{KL}(S') + \log(2)}{\mathcal{M}(\epsilon, S', \rho)} \ .$$

In view of the above proposition, we mainly have to choose a suitable subset S', control its Kullback diameter, and get a sharp lower bound of its packing number. The main difficulty stems from the fact that the loss function $\rho(\mathbf{x},\mathbf{y}) = d_{\infty,\mathcal{O}}(\mathbf{x},\mathbf{y})$ is a minimum over a collection of orthogonal transformations. It is therefore challenging to get a lower bound of this loss.

First, we prove an upper bound of the Kullback discrepansy whose proof is postponed to the end of the section.

Claim 4.B.2 *For any* $\mathbf{x},\mathbf{x}' \in \mathcal{C}$*, we have* $KL(\mathbb{P}_{(\mathbf{x},f_0)} \,\|\, \mathbb{P}_{(\mathbf{x}',f_0)}) \leq 8\sum_{i,j}(f_0(x_i,x_j) - f_0(x'_i,x'_j))^2$.

Fix $\delta_n = C'\, c_a \sqrt{\log(n)/n}$ for a small enough constant $C' \in (0,1]$ that will be set later. For n large enough, let S' be the set of vectors $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}$ defined as follows.
$\forall s \in [n]$:

$$\underline{x}_j^{(s)} = 2\pi\frac{j}{n} + \omega_j^{(s)}\delta_n, \qquad j = 1,\dots,n,$$

where $\omega_j^{(s)} \in \{0,1\}$ satisfies

$$\begin{aligned}
\forall s \in [1, n/3], \qquad & \omega_j^{(s)} = 1 \text{ iff } j = 1, 2, s+2,\\
\forall s \in (n/3, 2n/3] \qquad & \omega_j^{(s)} = 1 \text{ iff } j = 1, 3, s - \lfloor n/3 \rfloor + 4,\\
\forall s \in (2n/3, n], \qquad & \omega_j^{(s)} = 1 \text{ iff } j = 1, 4, s - \lfloor 2n/3 \rfloor + 6.
\end{aligned}$$

We can observe that $S' = \{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}\} \subset S$, and that, $\forall\, 1 \leq s < t \leq n$

$$d_{\infty,\mathcal{O}}(\mathbf{x}^{(t)},\mathbf{x}^{(s)}) \geq \delta_n, \tag{4.36}$$

so that $\mathcal{M}(\delta_n, S', d_{\infty,\mathcal{O}}) \geq n$.

To lower bound the KL diameter of S', we use Claim 4.B.2 together with the definition of f_0 in (4.35), which gives:

$$KL(\mathbb{P}_{(\mathbf{x}^{(t)},f_0)} \,\|\, \mathbb{P}_{(\mathbf{x}^{(s)},f_0)}) \leq c' n \delta_n^2 \leq c'(C' c_a)^2 \log(n)$$

for some numerical constant $c' > 0$. Then, if the constant C' in the definition of δ_n satisfies $C' \leq (2c_a\sqrt{c'})^{-1}$, we have $d_{KL}(S') \leq \log(n)/4$.

Applying Proposition 4.B.1 to this set S', we arrive at

$$\inf_{\hat{\mathbf{x}}} \sup_{\mathbf{x}\in S'} \mathbb{P}_{(\mathbf{x},f_0)}\left[d_{\infty,\mathcal{O}}(\hat{\mathbf{x}},\mathbf{x}) \geq \frac{\delta_n}{2}\right] \geq 1 - \frac{\log(n)/4 + \log(2)}{\log(n)} \geq \frac{1}{2},$$

as soon as n is large enough. □

Proof of Claim 4.B.2. By definition of the Kullback-Leibler divergence, and $F_{ij} := f_0(x_i, x_j)$ and $F'_{ij} := f_0(x'_i, x'_j)$, we have

$$KL(\mathbb{P}_{(\mathbf{x},f_0)} \,\|\, \mathbb{P}_{(\mathbf{x}',f_0)}) = \sum_{i<j} F_{ij} \log \frac{F_{ij}}{F'_{ij}} + (1 - F_{ij}) \log \frac{1 - F_{ij}}{1 - F'_{ij}} \ ,$$

and since $\log(t) \le t - 1$ for all $t > 0$, it follows that

$$KL(\mathbb{P}_{(\mathbf{x},f_0)} \,\|\, \mathbb{P}_{(\mathbf{x}',f_0)}) \le \sum_{ij} \frac{(F_{ij} - F'_{ij})^2}{F'_{ij}(1 - F'_{ij})} \le 8 \sum_{i,j} (F_{ij} - F'_{ij})^2,$$

where the second inequality follows from the fact that $1/4 \le F'_{ij} \le 3/4$.

4.C Proof for the spectral method

4.C.1 Uniform approximation algorithm (Lemma 4.4.2)

Recall that for a vector $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{C}^n$, we say that $\mathbf{x}$ is ordered, if these points are consecutive when one walks on the circle with the trigonometric direction.

We introduce some notation. For any vector $\mathbf{v} \in \mathcal{C}^n$, denote $\mathcal{P}_n(\mathbf{v})$ all elements of $\mathcal{P}_n$ that have the same order as $\mathbf{v}$. In other words, $\mathbf{u} \in \mathcal{P}_n(\mathbf{v})$ if $\mathbf{u} \in \mathcal{P}_n$ and, for any permutation σ such that $v_{\sigma(1)}, \dots, v_{\sigma(n)}$ is ordered, the sequence $u_{\sigma(1)}, \dots, u_{\sigma(n)}$ is ordered.

The next lemma is a key element in the proof; it states that the d_1-error between two vectors cannot get larger after reordering one of the vector.

Lemma 4.C.1 *Consider any $\mathbf{v} \in \mathcal{C}^n$, and $\mathbf{u} \in \mathcal{P}_n$. Then, there exists a permutation σ such that $\mathbf{u}_\sigma \in \mathcal{P}_n(\mathbf{v})$ and*

$$d_1(\mathbf{v}, \mathbf{u}_\sigma) \lesssim d_1(\mathbf{v}, \mathbf{u}) \ .$$

For $\mathbf{x} \in \mathbb{R}^{2\times n}$, $\mathbf{x}^* \in \mathcal{C}^n$, and $Q \in \mathcal{O}$, let us show that the UA algorithm returns an element $\widetilde{\mathbf{x}}$ of $\mathcal{P}_n$ fulfilling the inequality

$$\|\widetilde{\mathbf{x}} - Q\mathbf{x}^*\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) + 1.$$

For any ordered vector $\mathbf{v} \in \mathcal{C}^n$, the set $\mathcal{P}_n(\mathbf{v})$ can be described by a single element $u \in \mathcal{P}_n(\mathbf{v})$ and all circular permutations of u. Given the projection $\mathbf{x}^{(p)}$ of $\mathbf{x}$ on $\mathcal{C}^n$, UA computes in step 3 an element of $\overline{\mathbf{x}} \in \mathcal{P}_n(\mathbf{x}^{(p)}_\sigma)$ and then picks in step 4 a vector $\overline{\mathbf{x}}' \in \mathcal{P}_n(\mathbf{x}^{(p)}_\sigma)$ that has the smallest l_1-error:

$$\overline{\mathbf{x}}' = \operatorname*{argmin}_{\mathbf{u} \in \mathcal{P}_n(\mathbf{x}^{(p)}_\sigma)} \|\mathbf{x}^{(p)}_\sigma - \mathbf{u}\|_1.$$

Finally, UA picks $\tilde{\mathbf{x}} = \overline{\mathbf{x}}'_{\sigma^{-1}}$ so that $\|\mathbf{x}^{(p)}_\sigma - \overline{\mathbf{x}}'\|_1 = \|\mathbf{x}^{(p)} - \tilde{\mathbf{x}}\|_1$. It follows from these definition and the equivalence between the distance d on $\mathcal{C}$ and l_1-norm in $\mathbb{R}^2$ that

$$\|\mathbf{x}^{(p)} - \tilde{\mathbf{x}}\|_1 = \min_{\mathbf{v} \in \mathcal{P}_n(\mathbf{x}^{(p)})} \|\mathbf{x}^{(p)} - \mathbf{v}\|_1 \lesssim \min_{\mathbf{v} \in \mathcal{P}_n(\mathbf{x}^{(p)})} d_1(\mathbf{x}^{(p)}, \mathbf{v}) \ .$$

Gathering this bound with Lemma 4.C.1 we derive that

$$\|\mathbf{x}^{(p)} - \tilde{\mathbf{x}}\|_1 \lesssim \min_{\mathbf{u}\in\mathcal{P}_n} d_1(\mathbf{x}^{(p)}, \mathbf{u}) \lesssim \min_{\mathbf{u}\in\mathcal{P}_n} \|\mathbf{x}^{(p)} - \mathbf{u}\|_1 \ .$$

As a consequence, it suffices to exhibit some $u \in \mathcal{P}_n$ such that its l_1 distance to the projection $\mathbf{x}^{(p)}$ is small. This is precisely the purpose of the next lemma.

Lemma 4.C.2 *There exists* $\mathbf{y} \in \mathcal{P}_n$ *such that* $\max_{1\le j\le n} d(y_j, x_j^{(p)}) \le \pi/16$ *and*

$$\|\mathbf{x}^{(p)} - \mathbf{y}\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) + 1 \ .$$

We conclude that

$$\|\mathbf{x}^{(p)} - \tilde{\mathbf{x}}\|_1 \lesssim \|\mathbf{x}^{(p)} - \mathbf{y}\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) + 1 \ .$$

By triangular inequality, we have

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_1 \le \|\tilde{\mathbf{x}} - \mathbf{x}^{(p)}\|_1 + \|\mathbf{x}^{(p)} - \mathbf{x}\|_1 \ .$$

The minimality associated with a projection (and the equivalence between the l_1-norm and the euclidean norm in $\mathbb{R}^2$) ensures that

$$\|\mathbf{x}^{(p)} - \mathbf{x}\|_1 \lesssim \|Q\mathbf{x}^* - \mathbf{x}\|_1 \ ,$$

since $\mathbf{x}^{(p)}$ is the projection of $\mathbf{x}$ on $\mathcal{C}^n$ and $Q\mathbf{x}^*$ is an element of $\mathcal{C}^n$. The last three displays allow us to conclude that

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) + 1 \ .$$

□

4.C.1.1 Proofs of Lemma 4.C.2

Let $\mathbf{x}^{**} \in \mathcal{P}_n$ be a closest approximation of $\mathbf{x}^*$ in $\mathcal{P}_n$, that is, such that $d_\infty(\mathbf{x}^*, \mathbf{x}^{**}) = \alpha(\mathbf{x}^*)$. The triangular inequality gives

$$\|Q\mathbf{x}^{**} - \mathbf{x}^{(p)}\|_1 \le \|Q\mathbf{x}^{**} - \mathbf{x}\|_1 + \|\mathbf{x} - \mathbf{x}^{(p)}\|_1 \lesssim \|Q\mathbf{x}^{**} - \mathbf{x}\|_1 \ ,$$

where the last inequality comes from the minimality associated with a projection and the equivalence between the l_1-norm and the euclidean norm in $\mathbb{R}^2$. By triangular inequality again,

$$\|Q\mathbf{x}^{**} - \mathbf{x}\|_1 \le \|Q\mathbf{x}^{**} - Q\mathbf{x}^*\|_1 + \|Q\mathbf{x}^* - \mathbf{x}\|_1 \ .$$

An orthogonal transformation preserves the distances, so

$$\|Q\mathbf{x}^{**} - Q\mathbf{x}^*\|_1 = \|\mathbf{x}^{**} - \mathbf{x}^*\|_1 \lesssim d_1(\mathbf{x}^{**}, \mathbf{x}^*) \le n\alpha(\mathbf{x}^*) \ ,$$

where we use the equivalence between the distance d in $\mathcal{C}$ and the l_1-norm in $\mathbb{R}^2$. Putting everything together, we conclude that

$$\|Q\mathbf{x}^* - \mathbf{x}^{(p)}\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) \ .$$

Although $\mathbf{x}^{**}$ belongs to $\mathcal{P}_n$, this is not necessarily the case for $Q\mathbf{x}^{**}$. Nevertheless, it is easy to check that there exists some $Q' \in \mathcal{O}$ such that $Q'\mathbf{x}^{**} \in \mathcal{P}_n$ and $\|Q'\mathbf{x}^{**} - Q\mathbf{x}^{**}\|_1 \lesssim 1$. Setting $\mathbf{y} := Q'\mathbf{x}^{**} \in \mathcal{P}_n$, then we see that that

$$\|\mathbf{y} - \mathbf{x}^{(p)}\|_1 \lesssim \|\mathbf{x} - Q\mathbf{x}^*\|_1 + n\alpha(\mathbf{x}^*) + 1 \ , \tag{4.37}$$

which concludes the proof.

4.C.1.2 Proof of Lemma 4.C.1.

Let τ be a permutation ordering the coordinates of $\mathbf{v}$ on the unit circle, meaning that $v_{\tau(1)}, \dots v_{\tau(n)}$ is ordered. For simplicity and without loss of generality, assume that τ is the identity. Define the set of 'bad' indices $\mathcal{B} = \{i : d(u_i, v_i) \geq \pi/16\}$. If $\mathcal{B} = [n]$, then $d_1(\mathbf{u}, \mathbf{v}) \geq n\pi/16$ and any permutation σ of $[n]$ leads to $d_1(\mathbf{u}_\sigma, \mathbf{v}) \leq 16 d_1(\mathbf{u}, \mathbf{v})$. We can assume henceforth that $|\mathcal{B}| < n$. First, we focus on the set of 'good' indices $\mathcal{G} = [n] \setminus \mathcal{B}$. We establish the following claim at the end of the proof.

Claim 4.C.3 *There exists a permutation σ of $\mathcal{G}$ such that the sequence $(u_\sigma(j))$ with $j \in \mathcal{G}$ is ordered and*

$$\sum_{i \in \mathcal{G}} d(u_{\sigma(i)}, v_i) \leq \sum_{i \in \mathcal{G}} d(u_i, v_i)$$

Hence, it is possible to order the restriction of u to $\mathcal{G}$ without increasing the sum of the distances. It remains to transform σ into a permutation of $[n]$. We iteratively add elements of $\mathcal{B}$ into σ. Consider any $i \in \mathcal{B}$. Let k and l be the two consecutive (modulo n) elements of $\mathcal{G}$ such that $\underline{u}_i$ belongs to the arc $[\underline{u}_{\sigma(k)}, \underline{u}_{\sigma(l)})$. Let r and s be the two consecutive elements of $\mathcal{G}$ such $i \in (r, s)$ (where we work modulo n). Then, we define the permutation σ' of $(\mathcal{G} \cup \{i\})$ as follows.

If $(r, s) = (k, l)$, then we take $\sigma'(j) = \sigma(j)$ if $j \in \mathcal{G}$ and $\sigma'(i) = i$. One readily checks that the sequence $(u_{\sigma'(j)})$ with $j \in \mathcal{G} \cup \{i\}$ is ordered and that $\sum_{j \in \mathcal{G} \cup \{i\}} d(u_{\sigma'(j)}, v_j) \leq \sum_{j \in \mathcal{G} \cup \{i\}} d(u_j, v_j)$.

Otherwise, we set $\sigma'(i) = \sigma(s)$ and $\sigma'(k) = i$. For $j \in \mathcal{G}$, let $\mathrm{succ}_{\mathcal{G}}(j)$ denote the successor of $j \in \mathcal{G}$. For any $j \in \mathcal{G}$ in the segment $[s, k)$, we set $\sigma'(j) = \sigma(\mathrm{succ}_{\mathcal{G}}(j))$. Besides, we set $\sigma'(j) = \sigma(j)$ for all $j \in \mathcal{G}$ in the segment $[l, r]$. In other words, we have shifted all elements in the segment $[s, l]$ to succesfully include i in the permutation σ'. It follows from the definition that the sequence $u_{\sigma'(j)}$ with $j \in \mathcal{G} \cup \{i\}$ is ordered. By triangular inequality, we have

$$\begin{aligned}
\sum_{j \in \mathcal{G} \cup \{i\}} d(u_{\sigma'(j)}, v_j) &= \sum_{j \in \mathcal{G} \cap [l, r]} d(u_{\sigma(j)}, v_j) + d(u_{\sigma(s)}, v_i) + d(u_i, v_k) + \sum_{j \in \mathcal{G} \cap [s, k)} d(u_{\sigma(\mathrm{succ}_{\mathcal{G}}(j))}, v_j) \\
&\leq 2\pi + \sum_{j \in \mathcal{G}} d(u_{\sigma(j)}, v_j) + \sum_{j \in \mathcal{G} \cap [s, k)} d(u_{\sigma(j)}, u_{\sigma(\mathrm{succ}_{\mathcal{G}}(j))}) \\
&\leq 4\pi + \sum_{j \in \mathcal{G}} d(u_{\sigma(j)}, v_j) \leq 4\pi + \sum_{j \in \mathcal{G} \cup \{i\}} d(u_j, v_j) \ ,
\end{aligned}$$

where we used in the third line that $\sum_{j \in \mathcal{G} \cap [s,k)} d(u_{\sigma(j)}, u_{\sigma(\mathrm{succ}_{\mathcal{G}}(j))}) \leq 2\pi$. Indeed, the sequence $u_{\sigma(j)}$ is ordered on the circle and this sum is therefore equal to the length of the arc $[\underline{u}_{\sigma(s)}, \underline{u}_{\sigma(k)}]$. By a straightforward induction, we manage to build a permutation $\overline{\sigma}$ on $[n]$ such that $(u_{\overline{\sigma}(j)})$ is ordered and

$$\begin{aligned}
\sum_{j \in [n]} d(u_{\overline{\sigma}(j)}, v_j) &\leq 4\pi \big| \big\{ i \ , d(u_i, v_i) \geq \frac{\pi}{16} \big\} \big| + \sum_{j \in [n]} d(u_j, v_j) \\
&\leq 65 \sum_{j \in [n]} d(u_j, v_j) \ ,
\end{aligned}$$

where we used Markov Inequality in the last line. We have shown the desired result.

Proof of claim 4.C.3. Without loss of generality, we assume in the proof that $\mathcal{B} = \emptyset$ so that we build a permutation σ of $[n]$.

Recall that $\mathbf{u} \in \mathcal{P}_n$ satisfies $d_\infty(\mathbf{u}, \mathbf{v}) \leq \pi/16$. We shall iteratively build a permutation σ such that $\mathbf{u}_\sigma$ is ordered. Let us first partition the one-dimensional torus $\mathbb{R}/(2\pi)$ into three parts $\mathbb{R}/(2\pi) = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$ where $\mathcal{D}_s = \left[(s-1)\frac{2\pi}{3}, s\frac{2\pi}{3}\right)$ for $s = 1, 2, 3$.

For $s = 1, 2, 3$, define $I_s = \{i : \underline{v}_i \in \mathcal{D}_s\}$. Since $d_\infty(\mathbf{u}, \mathbf{v}) \leq \pi/16$, it follows that $\{\underline{u}_i : i \in I_s\} \subset \left[(s-1)\frac{2\pi}{3} - \frac{\pi}{16}, s\frac{2\pi}{3} + \frac{\pi}{16}\right) = \mathcal{D}'_s$. Note that the diameter of $\mathcal{D}'_s$ is smaller $2\pi/3 + \pi/8 < \pi$. We have the decomposition

$$d_1(\mathbf{v}, \mathbf{u}) = \sum_{s=1}^{3} \sum_{i \in I_s} d(v_i, u_i) \ .$$

For $s = 1, 2, 3$, let σ_s denote the permutation of I_s such that the sequence $u_{\sigma_s(i)}$ is ordered when i describes I_s. Since the diameter of $\mathcal{D}'_s$ is at most π, the sequence $\underline{u}_{\sigma_s(i)}$ in $\mathcal{D}'_s$ is isometric to an increasing sequence of points in $[0, \pi] \subset \mathbb{R}$ endowed with the absolute value distance. It goes the same for the ordered sequence $\underline{v}_i$ in $\mathcal{D}'_s$. Next, we use the following classical property.

Claim 4.C.4 *Let $l \geq 1$ be an integer and $\mathbf{a}, \mathbf{b}$ be two monotonic vectors of $\mathbb{R}^l$, that is, $a_1 \leq a_2 \leq \ldots \leq a_l$ and $b_1 \leq b_2 \leq \ldots \leq b_l$. Then for all permutation τ of the indices $\{1, \ldots, l\}$, we have*

$$\sum_{j=1}^{l} |a_i - b_i| \leq \sum_{j=1}^{l} |a_i - b_{\tau(i)}| \quad \text{and} \quad \max(|a_i - b_i|) \leq \max(|a_i - b_{\tau(i)}|) \ .$$

It follows that, for $s = 1, 2, 3$,

$$\sum_{i \in I_s} d(v_i, u_{\sigma_s(i)}) \leq \sum_{i \in I_s} d(v_i, u_i) \ .$$

Let σ be the denote the permutation such that $\sigma(i) = \sigma_s(i)$ if $i \in I_s$. Obviously, we have $d_1(\mathbf{v}, \mathbf{u}_\sigma) \leq d_1(\mathbf{v}, \mathbf{u})$. Besides, u_σ is ordered except possibly at the indices $J_s = \{i : \underline{u}_{\sigma(i)} \in [(s-1)\frac{2\pi}{3} - \frac{\pi}{16}; (s-1)\frac{2\pi}{3} + \frac{\pi}{16}]\}$ with $s = 1, 2, 3$. Since $\max_i d(u_{\sigma(i)}, v_i) \leq \frac{\pi}{16}$ by the second part of the above claim, all $\underline{u}_{\sigma(i)}$ and $\underline{v}_i$ with $i \in J_s$ belong to a subset of diameter smaller than π. Besides,

$$d_1(\mathbf{v}, \mathbf{u}_\sigma) = \sum_{j \notin (\cup_s J_s)} d(v_i, u_{\sigma_s(i)}) + \sum_{s=1}^{3} \sum_{i \in J_s} d(v_i, u_{\sigma(i)}) \ .$$

Hence, we can build as previously partitions σ'_s of J_s that make $u_{\sigma'_s(\sigma(i))}$ ordered on J_s and so that

$$\sum_{i \in J_s} d(v_i, u_{\sigma'_s(\sigma(i))}) \leq \sum_{i \in J_s} d(v_i, u_{\sigma(i)}) \ .$$

Defining $\overline{\sigma}(i) = \sigma'_s(\sigma(i))$ if $i \in J_s$ for $s = 1, 2, 3$ and $\overline{\sigma}(i) = \sigma(i)$ otherwise, we conclude that $\mathbf{u}_{\overline{\sigma}}$ is ordered and that $d_1(\mathbf{v}, \mathbf{u}_{\overline{\sigma}}) \leq d_1(\mathbf{v}, \mathbf{u})$.

4.C.2 Proof of Proposition 4.4.1

In the geometric model introduced in sub-section 4.4.1, we will show that the estimation error of the spectral algorithm is bounded by $\frac{n\sqrt{n\log(n)}}{(\Delta_1\wedge\Delta_2)\vee 1}$ in l_1-norm. The proof consists in approximating the signal $F_{(\mathbf{x}^*,f)}$ by a circulant and circular-R matrix (Definition 4.C.5) whose spectrum is known (Lemma 4.C.6) and provides information on the latent positions $\mathbf{x}^*$. The difference between the spectrums of $F_{(\mathbf{x}^*,f)}$ and $A^{(1)}$ will be bounded using Davis-Kahan perturbation bound.

4.C.2.1 Preliminaries

Let us start by introducing the notion of circulant matrix (see [Gray, 2006, Recanati et al., 2018]).

Definition 4.C.5 *Assume that n is an odd integer. A symmetric matrix $M \in \mathbb{R}^{n\times n}$ is circulant if there exists a vector $\mathbf{a}$ of size n such that $M_{ij} = a_{|i-j|}$ and*

$$\forall k = 1,\dots,n-1, \qquad a_k = a_{n-k}.$$

Moreover, M is a circulant and circular R-matrix if the above holds and the sequence $(a_j)_{0\le j\le \lfloor n/2\rfloor}$ is non-increasing.

Remark: If n is even, we require that the sequence $(a_j)_{0\le j\le \lfloor n/2\rfloor -1}$ is non-increasing and that $a_k = a_{n-k-1}$ for all $k = 0,\dots,n-1$.

The spectrum of circulant matrices is known (see [Gray, 2006] and the references therein), which allows to deduce easily the spectrum of symmetric circulant matrices, see Proposition C.4 from [Recanati et al., 2018]. For clarity, we recall this result below (up to a small correction on the first coordinate of the eigenvector $v^{(m)}$).

Lemma 4.C.6 (spectrum of symmetric circulant matrices) *Let $M \in \mathbb{R}^{n\times n}$ be any symmetric circulant matrix associated to the vector $\mathbf{a}$.*

- *For $n = 2p+1$, the eigenvalues of M are equal to*

$$\alpha_m = a_0 + 2\sum_{j=1}^{p} a_j \cos\left(j\frac{2\pi m}{n}\right) \quad ,$$

where each α_m, $m = 1,\dots,p$, has multiplicity 2 and is associated with the two eignenvectors

$$\begin{aligned} \mathbf{u}^{(m)} &= (1, \cos(2\pi m/n), \dots, \cos((n-1)2\pi m/n)) \\ \mathbf{v}^{(m)} &= (0, \sin(2\pi m/n), \dots, \sin((n-1)2\pi m/n)). \end{aligned} \tag{4.38}$$

For $m = 0$, α_0 has multiplicity 1 and is associated to $\mathbf{u}^{(0)} = (1,\dots,1)$.

- *For $n = 2p$,*

$$\alpha_m = a_0 + 2\sum_{j=1}^{p-1} a_j \cos(j\frac{2\pi m}{n}) + a_p\cos(\pi m),$$

where each α_m, $m = 1, \dots, p-1$, is associated with the two eigenvectors in (4.38). The eigenvalue α_p is singular with $\mathbf{u}^{(p)} = (1, -1, \dots, 1, -1)$. For $m = 0$, α_0 has multiplicity 1 and is associated to $\mathbf{u}^{(0)} = (1, \dots, 1)$.

If the vector $\mathbf{a}$ has non-negative entries, α_0 is obviously the largest eigenvalue. The next lemma ensures that, for circulant R-matrices, α_1 is the second largest eigenvalue. Its proof can be found in [Recanati et al., 2018, Proposition C.5].

Lemma 4.C.7 (second largest eigenvalue) *For any symmetric and circulant circular-R matrix, with non-negative entries and eigenvalues $\{\alpha_m\}$, $m = 0, \dots, \lfloor n/2 \rfloor$ (as defined in Lemma 4.C.6), we have $\alpha_1 \geq \alpha_j$ for all $j = 2, \dots, \lfloor n/2 \rfloor$.*

4.C.2.2 Proof of Proposition 4.4.1

If $\Delta_1 \wedge \Delta_2 \leq C\sqrt{n \log(n)}$, then the bound in Proposition 4.4.1 trivially holds

$$\|\hat{\mathbf{x}}' - \mathbf{x}^*\|_{1,\mathcal{O}} \leq 4n \leq 4C \frac{n\sqrt{n \log(n)}}{(\Delta_1 \wedge \Delta_2) \vee 1} \ .$$

We assume therefore that $\Delta_1 \wedge \Delta_2 \geq C\sqrt{n \log(n)}$ for a constant C that will be set later. By definition of Δ_1 and Δ_2, this means that

$$|\lambda_0^* - \lambda_1^*| \wedge |\lambda_2^* - \lambda_3^*| \geq C\sqrt{n \log(n)}. \tag{4.39}$$

Let $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ denote eigenvectors of R as described in Lemma 4.C.6.

Lemma 4.C.8 *There exist a permutation σ and a circulant circular R-matrix R with non-negative entries such that the following inequality holds $\|F_{(\mathbf{x}^*,f)} - R_\sigma\|_2 \lesssim_{c_l, c_L, c_e, c_a} \sqrt{n \log(n)}$. Besides, the vector $\mathbf{x}^{**} \in \mathcal{P}_n$ defined by $x_i^{**} := (u^{(1)}_{\sigma(i)}, v^{(1)}_{\sigma(i)})$ for $i = 1, \dots, n$ satisfies*

$$d_{\infty,\mathcal{O}}(\mathbf{x}^{**}, \mathbf{x}^*) \lesssim_{c_a} \sqrt{\frac{\log(n)}{n}} \ . \tag{4.40}$$

Denote $\lambda_0 \geq \ldots \geq \lambda_{n-1}$ the eigenvalues of R. The combination of Lemmas 4.C.6 and 4.C.7 ensures that

$$\lambda_0 = \alpha_0 \geq \lambda_1 = \alpha_1 \geq \lambda_2 = \alpha_1 \geq \lambda_j$$

where $\lambda_j \in \{\alpha_2, \dots, \alpha_{\lfloor n/2 \rfloor}\}$ for all $j = 3, \dots, n-1$.

Let us recall Weyl's inequality (see e.g. [Tao, 2012, page 45]). Let A and B be $n \times n$ symmetric matrices with respective eigenvalues $\lambda_0 \geq \ldots \geq \lambda_{n-1}$ and $\lambda'_0 \geq \ldots \geq \lambda'_{n-1}$. Then, on has $|\lambda_i - \lambda'_i| \leq \|A - B\|_{op}$ for all $0 \leq i \leq n-1$.

Lemma 4.C.8 ensures that there exists a constant C'' only depending on c_l, c_L, c_e, c_a such that $\|F_{(\mathbf{x}^*,f)} - R_\sigma\|_2 \leq C''\sqrt{n \log(n)}$. Since R_σ has the same eigenvalues as R, it follows from Weyl's inequality that

$$|\lambda_i^* - \lambda_i| \leq \|F_{(\mathbf{x}^*,f)} - R_\sigma\|_{op} \leq \|F_{(\mathbf{x}^*,f)} - R_\sigma\|_2 \leq C''\sqrt{n \log(n)} \ , \tag{4.41}$$

for all $i = 0, \dots, n-1$.

If the constant C in (4.39) is chosen as $4C''$ where C'' is introduced in (4.41), it follows that

$$\begin{aligned} \lambda_0 - \lambda_1 &\geq (\lambda_0^* - \lambda_1^*) - (\lambda_0^* - \lambda_0) - (\lambda_1 - \lambda_1^*) \geq (C - 2C'')\sqrt{n\log(n)} \\ &\geq \frac{C}{2}\sqrt{n\log(n)} \end{aligned}$$

and similarly

$$\lambda_2 - \lambda_3 \geq (\lambda_2^* - \lambda_3^*) - (\lambda_2^* - \lambda_2) - (\lambda_3 - \lambda_3^*) \geq \frac{C}{2}\sqrt{n\log(n)} \ .$$

Since the eigenvectors $(\sqrt{2/n})\mathbf{u}^{(1)}$ and $(\sqrt{2/n})\mathbf{v}^{(1)}$ of R are orthonormal (see Lemma 4.C.9 below), the vectors $(\sqrt{2/n})\mathbf{u}_\sigma^{(1)}$ and $(\sqrt{2/n})\mathbf{v}_\sigma^{(1)}$ are orthonormal eigenvectors of R_σ, with the same eigenvalue $\lambda_1 = \lambda_2 = \alpha_1$.

Lemma 4.C.9 *The vectors $(\sqrt{2/n})\mathbf{u}^{(1)}$ and $(\sqrt{2/n})\mathbf{v}^{(1)}$ are orthonormal.*

Next, we state a variant of Davis-Kahan perturbation bound [Yu et al., 2015, see Theorem 2].

Lemma 4.C.10 (Davis-Kahan) *Let $M, \hat{M} \in \mathbb{R}^{n\times n}$ be two symmetric matrices, with eigenvalues $\lambda_0 \geq \ldots \geq \lambda_{n-1}$ and $\hat{\lambda}_0 \geq \ldots \geq \hat{\lambda}_{n-1}$ respectively. Fix $0 \leq r \leq s \leq n-1$ and assume that $(\lambda_{r-1} - \lambda_r) \wedge (\lambda_s - \lambda_{s+1}) > 0$, where $\lambda_{-1} = \infty$ and $\lambda_n = -\infty$. Let $d = s - r + 1$, and let $\mathbf{V} = (\mathbf{v}_r, \mathbf{v}_{r+1}, \dots, \mathbf{v}_s) \in \mathbb{R}^{n\times d}$ and $\hat{\mathbf{V}} = (\hat{\mathbf{v}}_r, \hat{\mathbf{v}}_{r+1}, \dots, \hat{\mathbf{v}}_s) \in \mathbb{R}^{n\times d}$ have orthonormal columns satisfying $M\mathbf{v}_j = \lambda_j\mathbf{v}_j$ and $\hat{M}\hat{\mathbf{v}}_j = \hat{\lambda}_j\hat{\mathbf{v}}_j$ for $j = r, r+1, \dots, s$. Then, there exists an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ such that*

$$\|\hat{\mathbf{V}}Q - \mathbf{V}\|_2 \leq \sqrt{8d}\frac{\|\hat{M} - M\|_{op}}{(\lambda_{r-1} - \lambda_r) \wedge (\lambda_s - \lambda_{s+1})} \ .$$

The assumptions of Lemma 4.C.10 are therefore fulfilled for the orthonormal eigenvectors $(\sqrt{2/n})\mathbf{u}_\sigma^{(1)}$ and $(\sqrt{2/n})\mathbf{v}_\sigma^{(1)}$, and the positive spectral gaps $(\lambda_0 - \lambda_1) \wedge (\lambda_2 - \lambda_3) > 0$. Hence, for $\hat{\mathbf{x}}'$ and $\mathbf{x}^{**} = (\mathbf{u}_\sigma^{(1)}, \mathbf{v}_\sigma^{(1)})$ in $\mathbb{R}^{2\times n}$, Lemma 4.C.10 entails

$$\sqrt{\frac{2}{n}}\|Q\hat{\mathbf{x}}' - \mathbf{x}^{**}\|_2 \lesssim \frac{\|A^{(1)} - R_\sigma\|_{op}}{(\lambda_0 - \lambda_1) \wedge (\lambda_2 - \lambda_3)}$$

for some $Q \in \mathcal{O}$.

It remains to control $\|A^{(1)} - R_\sigma\|_{op}$ and the spectral gap.

$$\|A^{(1)} - R_\sigma\|_{op} \leq \|A^{(1)} - F_{(\mathbf{x}^*, f)}\|_{op} + \|F_{(\mathbf{x}^*, f)} - R_\sigma\|_2 \ , \tag{4.42}$$

using the triangular inequality and the fact that the operator norm is smaller than the Frobenius norm. To control the operator norm of the noise matrix, we shall use the following result [Vershynin, 2018, Corollary 4.4.8]. See the same reference for the definition of sub-Gaussian norms $\|.\|_{\psi_2}$.

Lemma 4.C.11 (norm of symmetric matrices with sub-gaussian entries) *Let A be an $n \times n$ symmetric random matrix whose entries A_{ij} on and above the diagonal are independent mean-zero sub-gaussian random variables. Then, for any $t > 0$, we have*

$$\|A\|_{op} \leq cK(\sqrt{n} + t)$$

with probability at least $1 - 4e^{-t^2}$. Here $K = \max_{i,j} \|A_{i,j}\|_{\psi_2}$.

Applying the above lemma with $t = c'\sqrt{\log(n)}$ (for a large enough numerical constant c) to $A^{(1)} - F_{(\mathbf{x}^*,f)}$, we deduce that $\|A^{(1)} - F_{(\mathbf{x}^*,f)}\|_{op} \lesssim \sqrt{n}$ with probability higher than $1 - 1/n^2$.

Together with Lemma 4.C.8 and the bound (4.42), we deduce that $\|A^{(1)} - R_\sigma\|_{op} \lesssim_{c_l,c_L,c_e,c_a} \sqrt{n\log(n)}$, so that

$$\|Q\hat{\mathbf{x}}' - \mathbf{x}^{**}\|_2 \lesssim_{c_l,c_L,c_e,c_a} \frac{n\sqrt{\log(n)}}{(\lambda_0 - \lambda_1) \wedge (\lambda_2 - \lambda_3)}.$$

Then, we deduce from Cauchy-Schwarz inequality that

$$\|\hat{\mathbf{x}}' - \mathbf{x}^{**}\|_{1,\mathcal{O}} \leq \|Q\hat{\mathbf{x}}' - \mathbf{x}^{**}\|_1 \lesssim_{c_l,c_L,c_e,c_a} \frac{n\sqrt{n\log(n)}}{(\lambda_0 - \lambda_1) \wedge (\lambda_2 - \lambda_3)},$$

taking the minimum over the set $\mathcal{O}$. Besides, the bounds (4.39) and (4.41) together with $C = 4C''$ in these bounds allow us to replace the above spectral gaps by $(\lambda_0^* - \lambda_1^*) \wedge (\lambda_2^* - \lambda_3^*)$. By (4.40), and the equivalence between the distance d in $\mathcal{C}$ and the l_1-norm in $\mathbb{R}^2$, we have $\|\mathbf{x}^* - \mathbf{x}^{**}\|_{1,\mathcal{O}} \lesssim_{c_a} \sqrt{n\log(n)}$. Since all the entries of $F_{(\mathbf{x}^*,f)}$ belong to $[0,1]$, we have $\lambda_0 \leq n$ and it follows from the triangular inequality that

$$\|\hat{\mathbf{x}}' - \mathbf{x}^*\|_{1,\mathcal{O}} \lesssim_{c_l,c_L,c_e,c_a} \frac{n\sqrt{n\log(n)}}{(\lambda_0^* - \lambda_1^*) \wedge (\lambda_2^* - \lambda_3^*)} \ .$$

The result follows. □

4.C.2.3 Proofs of technical lemmas

Proof of Lemma 4.C.8. For $\mathbf{x}^*$ on $\mathcal{C}$ satisfying (4.12), there exists $\mathbf{x} \in \mathcal{P}_n$ satisfying

$$d_\infty(\mathbf{x}^*, \mathbf{x}) \lesssim_{c_a} \sqrt{\log(n)/n} \ . \tag{4.43}$$

Combining this with the bi-Lipschitz condition (4.1), and the equivalence between the distance d in $\mathcal{C}$ and the euclidean norm in $\mathbb{R}^2$, we get

$$\|F_{(\mathbf{x}^*,f)} - F_{(\mathbf{x},f)}\|_2 \lesssim_{c_L,c_e,c_a} \sqrt{n\log(n)} \tag{4.44}$$

for the matrices $F_{(\mathbf{x}^*,f)} = (f(x_i^*, x_j^*))$ and $F_{(\mathbf{x},f)} = (f(x_i, x_j))$.

Let τ be some permutation that orders $x_1, \ldots, x_n$ on the unit circle, that is, such that $x_{\tau(1)}, \ldots, x_{\tau(n)}$ is ordered. Then, $F_{(\mathbf{x},f),\tau}$ is a symmetric circulant matrix since f is symmetric and satisfies the geometric condition (4.13) on the unit circle $\mathcal{C}$.

The matrix $F_{(\mathbf{x},f),\tau}$ is therefore defined by a single vector $\mathbf{a}$ of size n such that $a_s = \tilde{f}(2\pi s/n)$ for $s = 1, \dots, \lfloor n/2 \rfloor$. From the Lipschitz condition (4.2), we deduce that $\mathbf{a}$ satisfies some kind of weak non-increasing condition, that is $a_t \geq a_s \geq 0$ for all $0 \leq t < s \leq \lfloor n/2 \rfloor$ satisfying $s - t \gtrsim_{c_l,c_e} \sqrt{n \log(n)}$.

From the bi-Lipschitz condition (4.1), it is easy to see that $\mathbf{a}$ can be uniformly approximated by a non-increasing vector $\mathbf{a}'$ such that $\max_j |a_j - a'_j| \lesssim_{c_l,c_L,c_e} \sqrt{\log(n)/n}$. Denoting R the circulant circular R-matrix based on the vector $\mathbf{a}'$, this means that $\max_{ij} |R_{ij} - f(x_{\tau(i)}, x_{\tau(j)})| \lesssim_{c_l,c_L,c_e} \sqrt{\log(n)/n}$. Hence,

$$\|F_{(\mathbf{x},f)} - R_{\tau^{-1}}\|_2 = \|F_{(\mathbf{x},f),\tau} - R\|_2 \lesssim_{c_l,c_L,c_e} \sqrt{n \log(n)} \ .$$

The first result of Lemma 4.C.8 is a consequence of (4.44) with the last display, after setting $\sigma = \tau^{-1}$.

Next, it follows from the definition of $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ that the ordered vector $\mathbf{x}_{\sigma^{-1}} = (Q\mathbf{u}^{(1)}, Q\mathbf{v}^{(1)})$ for some orthogonal transformation Q in $\mathbb{R}^2$. Equivalently, we have $\mathbf{x} = (Q\mathbf{u}^{(1)}_\sigma, Q\mathbf{v}^{(1)}_\sigma)$. Thus, the second result of the lemma follows from (4.43) taking $\mathbf{x}^{**} = (\mathbf{u}^{(1)}_\sigma, \mathbf{v}^{(1)}_\sigma)$.

□

Proof of Lemma 4.C.9. Since $\sum_{k=0}^{n-1} e^{\iota 4\pi k/n} = 0$, we have $\sum_{k=0}^{n-1} \cos(4\pi k/n) = 0$ and $\sum_{k=0}^{n-1} \sin(4\pi k/n) = 0$. Then, combining with the trigonometric formulas, $\cos(2x) = 2\cos^2(x) - 1$, and $\sin(2x) = 2\cos(x)\sin(x)$, we get

$$\|\mathbf{u}^{(1)}\|_2^2 = \sum_{k=0}^{n-1} \cos^2(2\pi k/n) = \frac{n}{2} \ , \quad \text{and,}$$
$$\langle \mathbf{u}^{(1)}, \mathbf{v}^{(1)} \rangle = \sum_{k=0}^{n-1} \cos(2\pi k/n)\sin(2\pi k/n) = 0 \ .$$

□

Bibliography

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.*, 18(1):6446–6531, 2017.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 670–688, 2015.

Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Ernesto Araya and Yohann De Castro. Latent distance estimation for random geometric graphs. *arXiv preprint arXiv:1909.06841*, 2019.

Ery Arias-Castro, Antoine Channarond, Bruno Pelletier, and Nicolas Verzelen. On the estimation of latent distances using graph distances. *arXiv preprint arXiv:1804.10611*, 2018.

Ery Arias-Castro, Adel Javanmard, and Bruno Pelletier. Perturbation bounds for procrustes, classical scaling, and trilateration, with applications to manifold learning. *Journal of Machine Learning Research*, 21:15–1, 2020.

Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

Jonathan E Atkins, Erik G Boman, and Bruce Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.

Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

Stephen T Barnard, Alex Pothen, and Horst Simon. A spectral algorithm for envelope reduction of sparse matrices. *Numerical linear algebra with applications*, 2(4):317–334, 1995.

Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

Peter J Bickel, Aiyou Chen, Elizaveta Levina, et al. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

Béla Bollobás. Random graphs. In *Modern graph theory*, pages 215–252. Springer, 1998.

Danail Bonchev and Gregory A Buck. Quantitative measures of network complexity. In *Complexity in chemistry, biology, and ecology*, pages 191–235. Springer, 2005.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs. *Ann. Probab.*, 46(1):1–71, 2018.

Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of pair comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122 – 142, 1996.

Le Can, Elizaveta Levina, and Roman Vershynin. Concentration of random graphs and application to community detection. *Proc. Int. Cong. of Math.*, 3:2913–2928, 2018.

Francois Caron and Arnaud Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems 21*, pages 273–280. 2009.

Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521): 241–251, 2018.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, 17(27):1–57, 2016.

Zhengxing Chen, Su Xue, John Kolen, Navid Aghdaie, Kazi A. Zaman, Yizhou Sun, and Magy Seif El-Nasr. EOMM: An engagement optimized matchmaking framework. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1143–1150, 2017.

Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 391–423, 2015.

V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

Jens Christian Claussen. Offdiagonal complexity: A computationally quick complexity measure for graphs and networks. *Physica A: Statistical Mechanics and its Applications*, 375(1):365–373, 2007.

Gregory M Constantine. Graph complexity and the laplacian matrix in blocked experiments. *Linear and Multilinear Algebra*, 28(1-2):49–56, 1990.

Yohann De Castro, Claire Lacour, and Thanh Mai Pham Ngoc. Minimax adaptive estimation of nonparametric geometric graphs. *arXiv preprint arXiv:1708.02107*, 2017.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, 2011.

Matthias Dehmer and Abbe Mowshowitz. A history of graph entropy measures. *Information Sciences*, 181(1):57–78, 2011.

Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.

Josep Diaz, Colin McDiarmid, and Dieter Mitsche. Learning random points from geometric graphs or orderings. *Random Structures & Algorithms*, 57(2):339–370, 2020.

DLMF. *NIST Digital Library of Mathematical Functions*. URL `http://dlmf.nist.gov/`. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2):99–124, 1998.

Kenneth J Falconer. *Techniques in fractal geometry*, volume 3.

Yingjie Fei and Yudong Chen. Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Transactions on Information Theory*, 65(1):551–571, 2019.

Fajwel Fogel, Rodolphe Jenatton, Francis Bach, and Alexandre d'Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, pages 1016–1024, 2013.

Delbert Fulkerson and Oliver Gross. Incidence matrices and interval graphs. *Pacific journal of mathematics*, 15(3):835–855, 1965.

Chao Gao, Yu Lu, Harrison H Zhou, et al. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.*, 18(1):1980–2024, 2017.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018.

Gemma C Garriga, Esa Junttila, and Heikki Mannila. Banded structure in binary matrices. *Knowledge and information systems*, 28(1):197–226, 2011.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

Christophe Giraud, Yann Issartel, Luc Lehéricy, and Matthieu Lerasle. Pair matching: When bandits meet stochastic block model. *arXiv preprint arXiv:1905.07342*, 2019.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

Sergiu Goschin, Ari Weinstein, Michael L. Littman, and Erick Chastain. Planning in reward-rich domains via pac bandits. In *Proceedings of the Tenth European Workshop on Reinforcement Learning*, volume 24, pages 25–42, 2013.

Robert M Gray. *Toeplitz and circulant matrices: A review*. now publishers inc, 2006.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6), 2019.

Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill(TM): A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 20, 2007.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

Yann Issartel. On the estimation of network complexity: Dimension of graphons. *arXiv preprint arXiv:1909.02900*, 2019.

Yann Issartel. Optimal embedding of interaction data: applications to random geometric graphs and statistical seriation. *In preparation*, 2020a.

Yann Issartel. Pair matching in stochastic block model with k communities: a sequential problem of link prediction. *In progress*, 2020b.

Yann Issartel. Pair-matching in latent space model: link prediction with adaptative queries. *In progress*, 2020c.

Kevin G. Jamieson and Robert Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems 24*, pages 2240–2248. 2011.

Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Top-k ranking from pairwise comparisons: When spectral ranking is optimal. *CoRR*, abs/1603.04153, 2016.

Harry Joe. Extended use of paired comparison models, with application to chess rankings. *Journal of the Royal Statistical Society: Series C*, 39(1):85–93, 1990.

Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, 2016.

Balázs Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in neural information processing systems*, pages 697–704, 2003.

Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax rates for estimating the dimension of a manifold. *arXiv preprint arXiv:1605.01011*, 2016.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, 2004.

Olga Klopp, Alexandre B Tsybakov, Nicolas Verzelen, et al. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Vladimir I Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *Annals of statistics*, pages 591–629, 2000.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, 1985.

Pierre Latouche and Stéphane Robin. Variational bayes model averaging for graphon functions and motif frequencies inference in w-graph models. *Statistics and Computing*, 26(6):1173–1185, 2016.

Antti M Latva-Koivisto. Finding a complexity measure for business process models.

Sylvain Le Corff, Matthieu Lerasle, and Élodie Vernet. A Bayesian nonparametric approach for generalized Bradley-Terry models in random environment. *arXiv preprint arXiv:1808.08104*, 2018.

Guillaume Lecué, Shahar Mendelson, et al. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

Jing Lei and Lingxue Zhu. A generic sample splitting approach for refined community recovery in stochastic block models. *arXiv preprint arXiv:1411.1469*, 2014.

Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems 30*, pages 3074–3083. 2017.

Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

Yu Lu and Harrison H. Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Pascal Massart. Concentration inequalities and model selection. In *Lecture Notes in Mathematics*, volume 1896. Springer, Berlin, 2007.

Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Symposium on Theory of Computing*, pages 694–703, 2014.

Catherine Matias and Stéphane Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, 47:55–74, 2014.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Tom Minka, Ryan Cleven, and Yordan Zaykov. TrueSkill 2: An improved Bayesian skill rating system. *MSR-TR-2018-8*, 2018.

Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *CoRR*, abs/1702.00467, 2017.

Mikołaj Morzy, Tomasz Kajdanowicz, and Przemysław Kazienko. On measuring the complexity of networks: Kolmogorov complexity versus entropy. *Complexity*, 2017, 2017.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. *Electron. J. Probab.*, 21:24 pp., 2016.

Mathew Penrose et al. *Random geometric graphs*. Number 5. Oxford university press, 2003.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

Miklós Z Rácz, Sébastien Bubeck, et al. Basic models and questions in statistical network analysis. *Statistics Surveys*, 11:1–47, 2017.

Antoine Recanati, Thomas Kerdreux, and Alexandre d'Aspremont. Reconstructing latent orderings by spectral clustering. *arXiv preprint arXiv:1807.07122*, 2018.

Wenbo Ren, Jia Liu, and Ness B. Shroff. Exploring k out of top ρ fraction of arms in stochastic bandits. In *Proceedings of Machine Learning Research*, volume 89, pages 2820–2828, 2019.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 1952.

Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. Theoretical justification of popular link prediction heuristics. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *Proceedings of Machine Learning Research*, volume 89, pages 2564–2572, 2019.

Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *J. Mach. Learn. Res.*, 18(1):7246–7283, 2017.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *J. Mach. Learn. Res.*, 17(1):2049–2095, 2016.

Clément Sire and Sidney Redner. Understanding baseball team standings and streaks. *Eur. Phys. J. B*, 67:473–481, 2009.

Karthik Sridharan. A gentle introduction to concentration inequalities. *Dept Comput Sci*, 2002.

Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. *arXiv preprint arXiv:1811.05808*, 2018.

Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2013.

Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for plackett-luce: A dueling bandits approach. In *Advances in Neural Information Processing Systems 28*, pages 604–612. 2015.

Minh Tang, Daniel L Sussman, Carey E Priebe, et al. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009.

Todd L Veldhuizen. Software libraries and their reuse: Entropy, kolmogorov complexity, and zipf's law. *arXiv preprint cs/0508023*, 2005.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J. Whiting, Devi M. Stuart-Fox, O'Connor David, D. Firth, Nigel C. Bennett, and Simon P. Bloomberg. Ultraviolet signals ultra-aggression in a lizard. *Animal behaviour*, 72: 353–363, 2006.

Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. In *Conference on Learning Theory*, pages 903–920, 2014.

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

Jia Yuan Yu and Shie Mannor. Unimodal bandits. In *ICML*, 2011.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Se-Young Yun and Alexandre Proutière. Accurate community detection in the stochastic block model via spectral algorithms. *CoRR*, abs/1412.7335, 2014a.

Seyoung Yun and Alexandre Proutière. Community detection via random and adaptive sampling. In *COLT*, 2014b.

Hector Zenil, Narsis Kiani, and Jesper Tegnér. A review of graph and network complexity from an algorithmic information perspective. *Entropy*, 20(8):551, 2018.

Ernst Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.*, 29(1):436–460, 1929.

Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.

Pedro Zufiria and Iker Barriales-Valbuena. Entropy characterization of random network models. *Entropy*, 19(7):321, 2017.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Inférence sur des graphes aléatoires

Mots clés: Réseaux aléatoires, statistiques non paramétriques, graphes aléatoires à positions latentes, graphon, estimation minimax, prédiction séquentielle de liens, bandits, appariement séquentiel et adaptatif

Résumé: Cette thèse s'inscrit dans les domaines de la statistique non-paramétrique et de la théorie statistique de l'apprentissage non-supervisé. Son objet est la compréhension et la mise en oeuvre de méthodes d'estimation et de décision pour des modèles de graphes aléatoires à espace latent. Ces outils probabilistes rencontrent un succès grandissant pour la modélisation de grands réseaux dans des domaines aussi différents que la biologie, le marketing ou les sciences sociales.
Dans un premier temps, nous définissons un indice identifiable de la dimension de l'espace latent puis un estimateur consistant de cet indice. Plus généralement, ces quantités identifiables et interprétables permettent de palier l'absence d'identifiabilité de l'espace latent lui-même.
Dans la suite, nous introduisons le problème de 'pair-matching'. En partant d'un graphe non-observé, une stratégie choisit de façon séquentielle des paires de noeuds et observe la présence/absence d'arêtes. Son objectif est de découvrir le plus grand nombre possible d'arêtes avec un budget fixé. Pour ce problème de type bandit, nous étudions les regrets optimaux dans un modèle à blocs stochastiques puis dans un graphe aléatoire géométrique.
Enfin, nous estimons les positions des noeuds dans l'espace latent, dans le cas particulier où l'espace est un cercle dans le plan euclidien.
Pour chacun des trois problèmes, nous obtenons des procédures optimales au sens minimax, ainsi que des procédures efficaces satisfaisant certaines garanties théoriques. Ces algorithmes sont analysés d'un point de vue non-asymptotique en s'appuyant, entre autres, sur des inégalités de concentration.

Title: Inference on random graphs

Keywords: Random networks, non-parametric statistics, latent position random graphs, graphon, minimax estimation, sequential link prediction, bandits, adaptive sequential matching

Abstract: This thesis lies at the intersection of the theories of non-parametric statistics and statistical learning. Its goal is to provide an understanding of statistical problems in latent space random graphs. Latent space models have emerged as useful probabilistic tools for modeling large networks in various fields such as biology, marketing or social sciences.
We first define an identifiable index of the dimension of the latent space and then a consistent estimator of this index. More generally, such identifiable and interpretable quantities alleviate the absence of identifiability of the latent space itself.
We then introduce the pair-matching problem. From a non-observed graph, a strategy sequentially queries pairs of nodes and observes the presence/absence of edges. Its goal is to discover as many edges as possible with a fixed budget of queries. For this bandit type problem, we study optimal regrets in stochastic block models and random geometric graphs.
Finally, we are interested in estimating the positions of the nodes in the latent space, in the particular situation where the space is a circle in the Euclidean plane.
For each of the three problems, we obtain procedures that achieve the statistical optimal performance, as well as efficient procedures with theoretical guarantees. These algorithms are analysed from a non-asymptotic viewpoint, relying in particular on concentration inequalities.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France