



HAL
open science

Near-Optimal Comparison Based Clustering

Michaël Perrot, Pascal Mattia Esser, Debarghya Ghoshdastidar

► **To cite this version:**

Michaël Perrot, Pascal Mattia Esser, Debarghya Ghoshdastidar. Near-Optimal Comparison Based Clustering. Neural Information Processing Systems, Dec 2020, Vancouver (virtual), Canada. pp.19388 - 19399, 10.5555/3495724.3497350 . hal-03106482

HAL Id: hal-03106482

<https://hal.science/hal-03106482v1>

Submitted on 11 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Near-Optimal Comparison Based Clustering

Michaël Perrot*

Univ Lyon, UJM-Saint-Etienne, CNRS, IOGS,
LabHC UMR 5516, F-42023, SAINT-ETIENNE, France
michael.perrot@univ-st-etienne.fr

Pascal Mattia Esser*, Debarghya Ghoshdastidar*

Department of Informatics
Technical University of Munich
{esser, ghoshdas}@in.tum.de

Abstract

The goal of clustering is to group similar objects into meaningful partitions. This process is well understood when an explicit similarity measure between the objects is given. However, far less is known when this information is not readily available and, instead, one only observes ordinal comparisons such as “*object i is more similar to j than to k .*” In this paper, we tackle this problem using a two-step procedure: we estimate a pairwise similarity matrix from the comparisons before using a clustering method based on semi-definite programming (SDP). We theoretically show that our approach can exactly recover a planted clustering using a near-optimal number of passive comparisons. We empirically validate our theoretical findings and demonstrate the good behaviour of our method on real data.

1 Introduction

In clustering, the objective is to group together objects that share the same semantic meaning, that are similar to each other, into k disjoint partitions. This problem has been extensively studied in the literature when a measure of similarity between the objects is readily available, for example when the examples have a Euclidean representation or a graph structure (Shi and Malik, 2000; Arthur and Vassilvitskii, 2007; von Luxburg, 2007). However, it has attracted less attention when the objects are difficult to represent in a standard way, for example cars or food. A recent trend to tackle this problem is to use comparison based learning (Ukkonen, 2017; Emamjomeh-Zadeh and Kempe, 2018) where, instead of similarities, one only observes comparisons between the examples:

Triplet comparison: Object x_i is more similar to object x_j than to object x_k ;

Quadruplet comparison: Objects x_i and x_j are more similar to each other than objects x_k and x_l . There are two ways to obtain these comparisons. On the one hand, one can adaptively query them from an oracle, for example a crowd. This is the active setting. On the other hand, they can be directly given, with no way to make new queries. This is the passive setting. In this paper, we study comparison based learning for clustering using passively obtained triplets and quadruplets.

Comparison based learning mainly stems from the psychometric and crowdsourcing literature (Shepard, 1962; Young, 1987; Stewart et al., 2005) where the importance and robustness of collecting ordinal information from human subjects has been widely discussed. In recent years, this framework has attracted an increasing amount of attention in the machine learning community and three main learning paradigms have emerged. The first one consists in obtaining an Euclidean embedding of the data that respects the comparisons as much as possible and then applying standard learning

*Equal contribution.

techniques (Borg and Groenen, 2005; Agarwal et al., 2007; Jamieson and Nowak, 2011; Tamuz et al., 2011; van der Maaten and Weinberger, 2012; Terada and von Luxburg, 2014; Zhang et al., 2015; Amid and Ukkonen, 2015; Arias-Castro, 2017). The second paradigm is to directly solve a specific task from the ordinal comparisons, such as data dimension or density estimation (Kleindessner and von Luxburg, 2015; Ukkonen et al., 2015), classification and regression (Haghiri et al., 2018), or clustering (Vikram and Dasgupta, 2016; Ukkonen, 2017; Ghoshdastidar et al., 2019). Finally, the third paradigm is an intermediate solution where the idea is to learn a similarity or distance function, as in embedding approaches, but, instead of satisfying the comparisons, the objective is to solve one or several standard problems such as classification or clustering (Kleindessner and von Luxburg, 2017). In this paper, we focus on this third paradigm and propose two new similarities based on triplet and quadruplet comparisons respectively. While these new similarities can be used to solve any machine learning problem, we show that they are provably good for clustering under a well known planted partitioning framework (Abbe, 2017; Yan et al., 2018; Xu et al., 2020).

Motivation of this work. A key bottleneck in comparison based learning is the overall number of available comparisons: given n examples, there exist $\mathcal{O}(n^3)$ different triplets and $\mathcal{O}(n^4)$ different quadruplets. In practice, it means that, in most applications, obtaining all the comparisons is not realistic. Instead, most approaches try to use as few comparisons as possible. This problem is relatively easy when the comparisons can be actively queried and it is known that $\Omega(n \ln n)$ adaptively selected comparisons are sufficient for various learning problems (Haghiri et al., 2017; Emamjomeh-Zadeh and Kempe, 2018; Ghoshdastidar et al., 2019). On the other hand, this problem becomes harder when the comparisons are passively obtained. The general conclusion in most theoretical results on learning from passive ordinal comparisons is that, in the worst case, almost all the $\mathcal{O}(n^3)$ or $\mathcal{O}(n^4)$ comparisons should be observed (Jamieson and Nowak, 2011; Emamjomeh-Zadeh and Kempe, 2018). The focus of this work is to show that, by carefully handling the passively obtained comparisons, it is possible to design comparison based approaches that use almost as few comparisons as active approaches for planted clustering problems.

Near-optimal guarantees for clustering with passive comparisons. In hierarchical clustering, Emamjomeh-Zadeh and Kempe (2018) showed that constructing a hierarchy that satisfies all comparisons in a top-down fashion requires $\Omega(n^3)$ passively obtained triplets in the worst case. Similarly, Ghoshdastidar et al. (2019) considered a planted model and showed that $\Omega(n^{3.5} \ln n)$ passive quadruplets suffice to recover the true hierarchy in the data using a bottom-up approach. Since the main difficulty lies in recovering the small clusters at the bottom of the tree, we believe that this latter result also holds for standard clustering. In this paper, we consider a planted model for standard clustering and we show that, when the number of clusters k is constant, $\Omega(n(\ln n)^2)$ passive triplets or quadruplets are sufficient for exact recovery.² This result is comparable to the sufficient number of active comparisons in most problems, that is $\Omega(n \ln n)$ (Haghiri et al., 2017; Emamjomeh-Zadeh and Kempe, 2018). Furthermore, it is near-optimal. Indeed, to cluster an example, it is necessary to observe it in a comparison at least once as, otherwise, it can only be assigned to a random cluster. Thus, to cluster n objects, it is necessary to have access to at least $\Omega(n)$ comparisons. Finally, to obtain these results, we study a semi-definite programming (SDP) based clustering method and our analysis could be of significant interest beyond the comparison based framework.

General noise model for comparison based learning. In comparison based learning, there are two main sources of noise. First, the observed comparisons can be noisy, that is the observed triplets and quadruplets are not in line with the underlying similarities. This noise stems, for example, from the randomness of the answers gathered from a crowd. It is typically modelled by assuming that each observed comparison is randomly (and independently) flipped (Jain et al., 2016; Emamjomeh-Zadeh and Kempe, 2018). This is mitigated in the active setting by repeatedly querying each comparison, but may have a significant impact in the passive setting where a single instance of each comparison is often observed. Apart from the aforementioned observation errors, the underlying similarities may also have intrinsic noise. For instance, the food data set by Wilber et al. (2014) contains triplet comparisons in terms of which items taste more similar, and it is possible that the taste of a dessert is closer to a main dish than to another dessert. This noise has been considered in Ghoshdastidar et al. (2019) by assuming that every pair of items possesses a latent random similarity, which affects the

²When we write that $\Omega(n(\ln n)^2)$ comparisons are sufficient, we express that any number of comparisons greater than $Cn(\ln n)^2$ with C a constant is sufficient to solve the problem. In other words, it means that having exactly $Cn(\ln n)^2$ comparisons is sufficient but also that having more comparisons is not detrimental. This notation is used in statistics and information theory (Fletcher et al., 2009) and is equivalent to \gtrsim .

responses to comparisons. In this paper, we propose, to the best of our knowledge, the first analysis that considers and shows the impact of both types of noise on the number of passive comparisons.

Scalable comparison based similarity functions. Several similarity and kernel functions have been proposed in the literature (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019). However, computing these similarities is usually expensive as they require up to $\mathcal{O}(n)$ passes over the set of available comparisons. In this paper, we propose new similarity functions whose construction is much more efficient than previous kernels. Indeed, they can be obtained with a single pass over the set of available comparisons. It means that our similarity functions can be computed in an online fashion where the comparisons are obtained one at a time from a stream. The main drawback compared to existing approaches is that we lose the positive semi-definiteness of the similarity matrix, but our theoretical results show that this is not an issue in the context of clustering. We also demonstrate this empirically as our similarities obtain results that are comparable with state of the art methods.

2 Background and theoretical framework

In this section, we present the comparison based framework and our planted clustering model, under which we later show that a small number of passive comparisons suffices for learning. We consider the following setup. There are n items, denoted by $[n] = \{1, 2, \dots, n\}$, and we assume that, for every pair of distinct items $i, j \in [n]$, there is an implicit real-valued similarity w_{ij} that we cannot directly observe. Instead, we have access to

$$\begin{aligned} \text{Triplets:} \quad \mathcal{T} &= \{(i, j, r) \in [n]^3 : w_{ij} > w_{ir}, i, j, r \text{ distinct}\}, & \text{or} \\ \text{Quadruplets:} \quad \mathcal{Q} &= \{(i, j, r, s) \in [n]^4 : w_{ij} > w_{rs}, i \neq j, r \neq s, (i, j) \neq (r, s)\}. \end{aligned} \quad (1)$$

There are $\mathcal{O}(n^4)$ possible quadruplets and $\mathcal{O}(n^3)$ possible triplets, and it is expensive to collect such a large number of comparisons via crowdsourcing. In practice, \mathcal{T} or \mathcal{Q} only contain a small fraction of all possible comparisons. We note that if a triple $i, j, r \in [n]$ is observed with i as reference item, then either $(i, j, r) \in \mathcal{T}$ or $(i, r, j) \in \mathcal{T}$ depending on whether i is more similar to j or to r . Similarly, when tuples (i, j) and (r, s) are compared, we have either $(i, j, r, s) \in \mathcal{Q}$ or $(r, s, i, j) \in \mathcal{Q}$.

Sampling and noise in comparisons. This paper focuses on passive observation of comparisons. To model this, we assume that the comparisons are obtained via uniform sampling, and every comparison is equally likely to be observed. Let $p \in (0, 1]$ denote a sampling rate that depends on n . We assume that every comparison (triplet or quadruplet) is independently observed with probability p . In expectation, $|\mathcal{Q}| = \mathcal{O}(pn^4)$ and $|\mathcal{T}| = \mathcal{O}(pn^3)$, and we can control the sampling rate p to study the effect of the number of observations, $|\mathcal{Q}|$ or $|\mathcal{T}|$, on the performance of an algorithm.

As noted in the introduction, the observed comparisons are typically noisy due to random flipping of answers by the crowd workers and inherent noise in the similarities. To model the external (crowd) noise we follow the work of Jain et al. (2016) and, given a parameter $\epsilon \in (0, 1]$, we assume that any observed comparison is correct with probability $\frac{1}{2}(1 + \epsilon)$ and flipped with probability $\frac{1}{2}(1 - \epsilon)$. To be precise, for observed triple $i, j, r \in [n]$ such that $w_{ij} > w_{ir}$,

$$\mathbf{P}((i, j, r) \in \mathcal{T} \mid w_{ij} > w_{ir}) = \frac{1 + \epsilon}{2}, \quad \text{whereas} \quad \mathbf{P}((i, r, j) \in \mathcal{T} \mid w_{ij} > w_{ir}) = \frac{1 - \epsilon}{2}. \quad (2)$$

The probabilities for flipping quadruplets can be similarly expressed. We model the inherent noise by assuming w_{ij} to be random, and present a model for the similarities under planted clustering.

Planted clustering model. We now present a theoretical model for the inherent noise in the similarities that reflects a clustered structure of the items. The following model is a variant of the popular stochastic block model, studied in the context of graph clustering (Abbe, 2017), and is related to the non-parametric weighted stochastic block model (Xu et al., 2020).

We assume that the item set $[n]$ is partitioned into k clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$ of sizes n_1, \dots, n_k , respectively, but **the number of clusters k as well as the clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$ are unknown to the algorithm.** Let F_{in} and F_{out} be two distributions defined on \mathbb{R} . We assume that the inherent (and unobserved) similarities $\{w_{ij} : i < j\}$ are random and mutually independent, and

$$w_{ij} \sim F_{in} \quad \text{if } i, j \in C_\ell \text{ for some } \ell, \quad \text{and} \quad w_{ij} \sim F_{out} \quad \text{otherwise.}$$

We further assume that w_{ii} is undefined, $w_{ji} = w_{ij}$, and that for w, w' independent,

$$\begin{aligned} \mathbf{P}_{w, w' \sim F_{in}}(w > w') &= \mathbf{P}_{w, w' \sim F_{out}}(w > w') = 1/2, \quad \text{and} \\ \mathbf{P}_{w \sim F_{in}, w' \sim F_{out}}(w > w') &= (1 + \delta)/2 \quad \text{for some } \delta \in (0, 1]. \end{aligned} \quad (3)$$

The first condition in (3) requires that F_{in}, F_{out} do not have point masses, and is assumed for analytical convenience. The second condition ensures that within cluster similarities are larger than inter-cluster similarities—a natural requirement. Ghoshdastidar et al. (2019) used a special case of the above model, where F_{in}, F_{out} are assumed to be Gaussian with identical variances σ^2 , and means satisfy $\mu_{in} > \mu_{out}$. In this case, $\delta = 2\Phi((\mu_{in} - \mu_{out})/\sqrt{2}\sigma) - 1$ where Φ is the cumulative distribution function of the standard normal distribution.

The goal of this paper is to obtain bounds on the number of passively obtained triplets/quadruplets that are sufficient to recover the aforementioned planted clusters with zero error. To this end, we propose two similarity functions respectively computed from triplet and quadruplet comparisons, and show that a similarity based clustering approach using semi-definite programming (SDP) can exactly recover clusters planted in the data using few passive comparisons.

3 A theoretical analysis of similarity based clustering

Before presenting our new comparison based similarity functions, we describe the SDP approach for clustering from similarity matrices that we use throughout the paper (Yan et al., 2018; Chen and Yang, 2020). In addition, we prove a generic theoretical guarantee for this approach that holds for any similarity matrix and, thus, that could be of interest even beyond the comparison based setting.

Similarity based clustering is widely used in machine learning, and there exist a range of popular approaches including spectral methods (von Luxburg, 2007), semi-definite relaxations (Yan and Sarkar, 2016), or linkage algorithms (Dasgupta, 2016) among others. We consider the following SDP for similarity based clustering. Let $S \in \mathbb{R}^{n \times n}$ be a symmetric similarity matrix among n items, and $Z \in \{0, 1\}^{n \times k}$ be the cluster assignment matrix that we wish to estimate. For unknown number of clusters k , it is difficult to directly determine Z , and hence, we estimate the *normalised clustering matrix* $X \in \mathbb{R}^{n \times n}$ such that $X_{ij} = \frac{1}{|C|}$ if i, j co-occur in estimated cluster C , and $X_{ij} = 0$ otherwise. Note that $\text{trace}(X) = k$. The following SDP was proposed and analysed by Yan et al. (2018) under the stochastic block model for graphs, and can also be applied in the more general context of data clustering (Chen and Yang, 2020). This SDP is agnostic to the number of clusters, but penalises large values of $\text{trace}(X)$ to restrict the number of estimated clusters:

$$\begin{aligned} \max_X \quad & \text{trace}(SX) - \lambda \text{trace}(X) \\ \text{s.t.} \quad & X \geq 0, \quad X \succeq 0, \quad X\mathbf{1} = \mathbf{1}. \end{aligned} \quad (\text{SDP-}\lambda)$$

Here, λ is a tuning parameter and $\mathbf{1}$ denotes the vector of all ones. The constraints $X \geq 0$ and $X \succeq 0$ restricts the optimisation to non-negative, positive semi-definite matrices.

We first present a general theoretical result for SDP- λ . Assume that the data has an implicit partition into k clusters C_1, \dots, C_k of sizes n_1, \dots, n_k and with cluster assignment matrix Z , and suppose that the similarity S is close to an *ideal similarity matrix* \tilde{S} that has a $k \times k$ block structure $\tilde{S} = Z\Sigma Z^T$. The matrix $\Sigma \in \mathbb{R}^{k \times k}$ is such that $\Sigma_{\ell\ell'}$ represents the ideal pairwise similarity between items from clusters C_ℓ and $C_{\ell'}$. Typically, under a random planted model, \tilde{S} is the same as $\mathbf{E}[S]$ up to possible differences in the diagonal terms. For $S = \tilde{S}$ and certain values of λ , the unique optimal solution of SDP- λ is a block diagonal matrix $X^* = ZN^{-1}Z^T$, where $N \in \mathbb{R}^{k \times k}$ is diagonal with entries n_1, \dots, n_k (see Appendix B). Thus, in the *ideal case*, solving the SDP provides the desired normalised clustering matrix from which one can recover the partition C_1, \dots, C_k . The following result shows that X^* is also the unique optimal solution of SDP- λ if S is sufficiently close to \tilde{S} .

Proposition 1 (Recovery of planted clusters using SDP- λ). *Let $Z \in \{0, 1\}^{n \times k}$ be the assignments for a planted k -way clustering, $\tilde{S} = Z\Sigma Z^T$, and $X^* = ZN^{-1}Z^T$ as defined above. Define*

$$\Delta_1 = \min_{\ell \neq \ell'} \left(\frac{\Sigma_{\ell\ell} + \Sigma_{\ell'\ell'}}{2} - \Sigma_{\ell\ell'} \right), \quad \text{and} \quad \Delta_2 = \max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{|C_\ell|} \sum_{j \in C_\ell} (S_{ij} - \tilde{S}_{ij}) \right|.$$

X^* is the unique optimal solution of $\text{SDP-}\lambda$ for any choice of λ in the interval

$$\|S - \tilde{S}\|_2 < \lambda < \min_{\ell} n_{\ell} \cdot \min \left\{ \frac{\Delta_1}{2}, \Delta_1 - 6\Delta_2 \right\}.$$

The proof of Proposition 1, given in Appendix B, is adapted from Yan et al. (2018) although uniqueness was not proved in this previous work. The term Δ_1 quantifies the separation between the ideal within and inter-cluster similarities, and is similar in spirit to the weak assortativity criterion for stochastic block models (Yan et al., 2018). On the other hand, the matrix spectral norm $\|S - \tilde{S}\|_2$ and the term Δ_2 both quantify the deviation of the similarities S from their ideal values \tilde{S} . Note that the number of clusters can be computed as $k = \text{trace}(X)$ and cluster assignment Z is obtained by clustering the rows of X^* using k -means or spectral clustering for example. In the experiments (Section 5), we present a data-dependent approach to tune λ and find k .

We conclude this section by noting that most of the previous analyses of SDP clustering either assume sub-Gaussian data (Yan and Sarkar, 2016) or consider similarity matrices with independence assumptions (Chen and Xu, 2014; Yan et al., 2018) that might not hold in general, and do not hold for our AddS-3 and AddS-4 similarities described in the next section. In contrast, the deterministic criteria stated in Proposition 1 make the result applicable in more general settings.

4 Similarities from passive comparisons

We present two new similarity functions computed from passive comparisons (AddS-3 and AddS-4) and guarantees for recovering planted clusters using $\text{SDP-}\lambda$ in conjunction with these similarities. Kleindessner and von Luxburg (2017) introduced pairwise similarities computed from triplets. A quadruplets variant was proposed by Ghoshdastidar et al. (2019). These similarities, detailed in Appendix A, are positive-definite kernels and have multiplicative forms. In contrast, we compute the similarity between items i, j by simply adding binary responses to comparisons involving i and j .

Similarity from quadruplets. We construct the additive similarity for quadruplets, referred to as AddS-4, in the following way. Recall the definition of \mathcal{Q} in Equation (1) and for every $i \neq j$, define

$$S_{ij} = \sum_{r \neq s} (\mathbb{I}_{\{(i,j,r,s) \in \mathcal{Q}\}} - \mathbb{I}_{\{(r,s,i,j) \in \mathcal{Q}\}}), \quad (\text{AddS-4})$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. The intuition is that if i, j are similar (w_{ij} is large), then for every observed tuple i, j, r, s , $w_{ij} > w_{rs}$ is more likely to be observed. Thus, (i, j, r, s) appears in \mathcal{Q} more often than (r, s, i, j) , and S_{ij} is a (possibly large) positive term. On the other hand, smaller w_{ij} leads to a negative value of S_{ij} . Under the aforementioned planted model with clusters of size n_1, \dots, n_k , one can verify that S_{ij} indeed reveals the planted clusters in expectation since if i, j belong to the same planted cluster, then $\mathbf{E}[S_{ij}] = p\epsilon\delta \sum_{\ell \in [k]} \frac{n_{\ell}(n - n_{\ell})}{2}$, and $\mathbf{E}[S_{ij}] = -p\epsilon\delta \sum_{\ell \in [k]} \binom{n_{\ell}}{2}$ otherwise.

Thus, in expectation, the within cluster similarity exceeds the inter-cluster similarity by $p\epsilon\delta \binom{n}{2}$.

Similarity from triplets. The additive similarity based on passive triplets AddS-3 is given by

$$S_{ij} = \sum_{r \neq i, j} (\mathbb{I}_{\{(i,j,r) \in \mathcal{T}\}} - \mathbb{I}_{\{(i,r,j) \in \mathcal{T}\}}) + (\mathbb{I}_{\{(j,i,r) \in \mathcal{T}\}} - \mathbb{I}_{\{(j,r,i) \in \mathcal{T}\}}) \quad (\text{AddS-3})$$

for every $i \neq j$. The AddS-3 similarity S_{ij} aggregates all the comparisons that involve both i and j , with either i or j as the reference item. Similar to the case of AddS-4, S_{ij} tends to be positive when w_{ij} is large, and negative for small w_{ij} . One can also verify that, under a planted model, the expected within cluster AddS-3 similarity exceeds the inter-cluster similarity by $p\epsilon\delta(n - 2)$.

A significant advantage of AddS-3 and AddS-4 over existing similarities is in terms of computational time for constructing S . Unlike existing kernels, both similarities can be computed from a single pass over \mathcal{T} or \mathcal{Q} . In addition, the following result shows that the proposed similarities can exactly recover planted clusters using only a few (near optimal) number of passive comparisons.

Theorem 1 (Cluster recovery using AddS-3 and AddS-4). *Let X^* denote the normalised clustering matrix corresponding to the true partition, and n_{\min} be the size of the smallest planted cluster.*

Given the triplet or the quadruplet setting, there exist absolute constants $c_1, c_2, c_3, c_4 > 0$ such that, with probability at least $1 - \frac{1}{n}$, X^* is the unique optimal solution of $\text{SDP-}\lambda$ if δ satisfies

$c_1 \frac{\sqrt{n \ln n}}{n_{\min}} < \delta \leq 1$, and one of the following two conditions hold:

- **(triplet setting)** S is given by AddS-3 , and the number of triplets $|\mathcal{T}|$ and the parameter λ satisfy

$$|\mathcal{T}| > c_2 \frac{n^3 (\ln n)^2}{\epsilon^2 \delta^2 n_{\min}^2} \quad \text{and} \quad c_3 \max \left\{ \sqrt{|\mathcal{T}| \frac{\ln n}{n}}, |\mathcal{T}| \epsilon \sqrt{\frac{\ln n}{n^3}}, (\ln n)^2 \right\} < \lambda < c_4 |\mathcal{T}| \frac{\epsilon \delta n_{\min}}{n^2};$$

- **(quadruplet setting)** S is given by AddS-4 , and the number of quadruplets $|\mathcal{Q}|$ and λ satisfy

$$|\mathcal{Q}| > c_2 \frac{n^3 (\ln n)^2}{\epsilon^2 \delta^2 n_{\min}^2} \quad \text{and} \quad c_3 \max \left\{ \sqrt{|\mathcal{Q}| \frac{\ln n}{n}}, |\mathcal{Q}| \epsilon \sqrt{\frac{\ln n}{n^3}}, (\ln n)^2 \right\} < \lambda < c_4 |\mathcal{Q}| \frac{\epsilon \delta n_{\min}}{n^2}.$$

The condition on δ and the number of comparisons ensure that the interval for λ is non-empty.

Theorem 1 is proved in Appendix C. This result shows that given a sufficient number of comparisons, one can exactly recover the planted clusters using $\text{SDP-}\lambda$ with an appropriate choice of λ . In particular, if there are k planted clusters of similar sizes and δ satisfies the stated condition, then recovery of the planted clusters with zero error is possible with only $\Omega\left(\frac{k^2}{\epsilon^2 \delta^2} n (\ln n)^2\right)$ passively obtained triplets or quadruplets. In this particular context, we make a few important remarks about the sufficient conditions.

Remark 1 (Comparison with existing results). For fixed k and fixed $\epsilon, \delta \in (0, 1]$, Theorem 1 states that $\Omega(n (\ln n)^2)$ passive comparisons (triplets or quadruplets) suffice to exactly recover the clusters. This significantly improves over the result of Ghoshdastidar et al. (2019) stating that $\Omega(n^{3.5} \ln n)$ passive quadruplets are sufficient in a planted setting, and the fact that $\Omega(n^3)$ triplets are necessary in the worst case (Emamjomeh-Zadeh and Kempe, 2018).

Remark 2 (Dependence of the number of comparisons on the noise levels ϵ, δ). When one can actively obtain comparisons, Emamjomeh-Zadeh and Kempe (2018) showed that it suffices to query $\Omega(n \ln(\frac{n}{\epsilon}))$ triplets. Compared to the $\ln(\frac{1}{\epsilon})$ dependence in the active setting, the sufficient number of passive comparisons in Theorem 1 has a stronger dependence of $\frac{1}{\epsilon^2}$ on the crowd noise level ϵ . While we do not know whether this dependence is optimal, the stronger criterion is intuitive since, unlike the active setting, the passive setting does not provide repeated observations of the same comparisons that can easily nullify the crowd noise. The number of comparisons also depends as $\frac{1}{\delta^2}$ on the inherent noise level, which is similar to the conditions in Ghoshdastidar et al. (2019).

Theorem 1 states that exact recovery primarily depends on two sufficient conditions, one on δ and the other on the number of passive comparisons ($|\mathcal{T}|$ or $|\mathcal{Q}|$). The following two remarks show that both conditions are necessary, up to possible differences in logarithmic factors.

Remark 3 (Necessity of the condition on δ). The condition on δ imposes the condition of $n_{\min} = \Omega(\sqrt{n \ln n})$. This requirement on n_{\min} appears naturally in planted problems. Indeed, assuming that all k clusters are of similar sizes, the above condition is equivalent to a requirement of $k = \mathcal{O}\left(\sqrt{\frac{n}{\ln n}}\right)$ and it is believed that polynomial time algorithms cannot recover $k \gg \sqrt{n}$ planted clusters (Chen and Xu, 2014, Conjecture 1).

Remark 4 (Near-optimal number of comparisons). To cluster n items, one needs to observe each example at least once. Hence, one trivially needs at least $\Omega(n)$ comparisons (active or passive). Similarly, existing works on actively obtained comparisons show that $\Omega(n \ln n)$ comparisons are sufficient for learning in supervised or unsupervised problems (Haghiri et al., 2017; Emamjomeh-Zadeh and Kempe, 2018; Ghoshdastidar et al., 2019). We observe that, in the setting of Remark 1, it suffices to have $\Omega(n (\ln n)^2)$ passive comparisons which matches the necessary conditions up to logarithmic factors. However, the sufficient condition on the number of comparisons becomes $\Omega(k^2 n (\ln n)^2)$ if k grows with n while ϵ and δ are fixed. It means that the worst case of $k = \mathcal{O}\left(\sqrt{\frac{n}{\ln n}}\right)$, stated in Remark 3, can only be tackled using at least $\Omega(n^2 \ln n)$ passive comparisons.

Remark 5 (No new information beyond $\Omega(n^2/\epsilon^2)$ comparisons). Theorem 1 shows that for large n and $\Omega(n^2/\epsilon^2)$ number of comparisons, the condition for exact recovery of the clusters is only

governed by the condition on δ as the interval for λ is always non empty. It means that, beyond a quadratic number of comparisons, no new information is gained by observing more comparisons. This explains why significantly fewer passive comparisons suffice in practice than the known worst-case requirements of $\Omega(n^3)$ passive triplets or $\Omega(n^4)$ passive quadruplets.

We conclude our theoretical discussion with a remark about recovering planted clusters when the pairwise similarities w_{ij} are observed. Our methods are near optimal even in this setting.

Remark 6 (Recovering planted clusters for non-parametric F_{in}, F_{out}). Theoretical studies in the classic setting of clustering with observed pairwise similarities $\{w_{ij} : i < j\}$ typically assume that the distributions F_{in} and F_{out} for the pairwise similarities are Bernoulli (in unweighted graphs), or take finitely many values (labelled graphs), or belong to exponential families (Chen and Xu, 2014; Aicher et al., 2015; Yun and Proutiere, 2016). Hence, the applicability of such results are restrictive. Recently, Xu et al. (2020) considered non-parametric distributions for F_{in}, F_{out} , and presented a near-optimal approach based on discretisation of the similarities into finitely many bins. Our work suggests an alternative approach: compute ordinal comparisons from the original similarities and use clustering on AddS-3 or AddS-4. Theorem 1 then guarantees, for any non-parametric and continuous F_{in} and F_{out} , exact recovery of the planted clusters under a near-optimal condition on δ .

5 Experiments

The goal of this section is three-fold: present a strategy to tune λ in SDP- λ ; empirically validate our theoretical findings; and demonstrate the performance of the proposed approaches on real datasets.

Choosing λ and estimating the number of clusters based on Theorem 1. Given a similarity matrix S , the main difficulty involved in using SDP- λ is tuning the parameter λ . Yan et al. (2018) proposed the algorithm SPUR to select the best λ as $\lambda^* = \arg \max_{0 \leq \lambda \leq \lambda_{\max}} \frac{\sum_{i \leq k_\lambda} \sigma_i(X_\lambda)}{\text{trace}(X_\lambda)}$ where X_λ is the solution of SDP- λ , k_λ is the closest integer to $\text{trace}(X_\lambda)$ and an estimate of the number of clusters, $\sigma_i(X_\lambda)$ is the i -th largest eigenvalue of X_λ , and λ_{\max} is a theoretically well-founded upper bound on λ . The maximum of the above objective is 1, achieved when X_λ has the same structure as X^* in Proposition 1. In our setting, Theorem 1 gives an upper bound on λ that depends on ϵ , δ and n_{\min} which are not known in practice. Furthermore, it is computationally beneficial to use the theoretical lower bound for λ instead of using $\lambda \geq 0$ as suggested in SPUR.

We propose to modify SPUR based on the fact that the estimated number of clusters k monotonically decreases with λ (details in Appendix D). Given Theorem 1, we choose $\lambda_{\min} = \sqrt{c(\ln n)/n}$ and $\lambda_{\max} = c/n$, where $c = |\mathcal{Q}|$ or $|\mathcal{T}|$. The trace of the SDP- λ solution then gives two estimates of the number of clusters, $k_{\lambda_{\min}}$ and $k_{\lambda_{\max}}$, and we search over $k \in [k_{\lambda_{\max}}, k_{\lambda_{\min}}]$ instead of searching over λ —in practice, it helps to search over the values $\max\{2, k_{\lambda_{\max}}\} \leq k \leq k_{\lambda_{\min}} + 2$. We select k that maximises the above SPUR objective, where X is computed using a simpler SDP (Yan et al., 2018):

$$\max_X \langle S, X \rangle \quad \text{s.t. } X \geq 0, \quad X \succeq 0, \quad X\mathbf{1} = \mathbf{1}, \quad \text{trace}(X) = k. \quad (\text{SDP-}k)$$

The overall approach is summarized in Algorithm 1.

Clustering with AddS-3 and AddS-4.³ For the proposed similarity matrices AddS-3 and AddS-4, the above strategy provides the optimal number of clusters k and a corresponding solution X_k of SDP- k . The partition is obtained by clustering the rows of X_k using k -means. Alternative approaches, such as spectral clustering, lead to similar performances (see Appendix E).

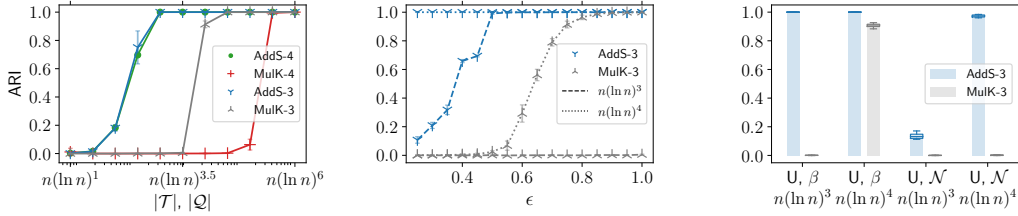
Evaluation function. We use the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) between the ground truth and the predictions. The ARI takes values in $[-1, 1]$ and measures the agreement between two partitions: 1 implies identical partitions, whereas 0 implies that the predicted clustering is random. In all the experiments, we report the mean and standard deviation over 10 repetitions.

Simulated data with planted clusters. We generate data using the planted model from Section 2 and verify that the learned clusters are similar to the planted ones. As default parameters we use $n = 1000$, $k = 4$, $\epsilon = 0.75$, $|\mathcal{T}| = |\mathcal{Q}| = n(\ln n)^4$ and $F_{in} = \mathcal{N}(\sqrt{2}\sigma\Phi^{-1}(\frac{1+\delta}{2}), \sigma^2)$, $F_{out} = \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$ and $\delta = 0.5$. In each experiment, we investigate the sensitivity of our method by varying one of the parameters while keeping the others fixed. We use SPUR to estimate the number of clusters.

³We provide a Python implementation on <https://github.com/mperron/AddS-Clustering>

Algorithm 1: Comparison-based SPUR

input : The number of examples n and the comparisons \mathcal{T} or \mathcal{Q} .**begin** Define $c = |\mathcal{T}|$ or $|\mathcal{Q}|$. Let S be obtained with AddS-3 or AddS-4. Define $\lambda_{\min} = \sqrt{\frac{c(\ln c)}{n}}$ and $\lambda_{\max} = \frac{c}{n}$. $X_{\lambda_{\min}}, X_{\lambda_{\max}} \leftarrow \text{SDP-}\lambda_{\min}, \text{SDP-}\lambda_{\max}$ on S . $k_{\lambda_{\min}}, k_{\lambda_{\max}} \leftarrow \lfloor \text{trace}(X_{\lambda_{\min}}) \rfloor, \lfloor \text{trace}(X_{\lambda_{\max}}) \rfloor$. **for** $k = \max\{2, k_{\lambda_{\max}}\}$ **to** $k_{\lambda_{\min}} + 2$ **do** | Solve SDP- k to obtain X_k . **end** Choose $\hat{k} = \underset{k}{\operatorname{argmax}} \frac{\sum_{i \leq k} \sigma_i(X_k)}{\text{trace}(X_k)}$, where $\sigma_i(X_k)$ denotes the i -th largest eigenvalue of X_k .**end****output** : Number of clusters \hat{k} , $X_{\hat{k}}$.

(a) Vary the number of comparisons (b) Vary the external noise level, ϵ (c) Vary the distributions F_{in}, F_{out} Figure 1: ARI of various methods on the planted model (higher is better). We vary: (1a) the number of comparisons $|\mathcal{T}|$ and $|\mathcal{Q}|$; (1b) the crowd noise level ϵ ; (1c) the distributions F_{in} and F_{out} .

As baselines, we use SDP- k (using the number of clusters estimated by our approaches) followed by k -means with two comparison based multiplicative kernels: MulK-3 for triplets (Kleindessner and von Luxburg, 2017) and MulK-4 for quadruplets (Ghoshdastidar et al., 2019).

We present some significant results in Figure 1 and defer the others to Appendix E. In Figure 1a, we vary the number of sampled comparisons. Unsurprisingly, our approaches are able to exactly recover the planted clusters using as few as $n(\ln n)^3$ comparisons—extra $\ln n$ factor compared to Theorem 1 accounts for ϵ, δ and constants. MulK-3 and MulK-4 respectively need $n(\ln n)^{4.5}$ and $n(\ln n)^{5.5}$ comparisons (both values exceed n^2 for $n = 1000$). In all our experiments, AddS-3 and AddS-4 have comparable performance while MulK-3 is significantly better than MulK-4. Thus we focus on triplets in the subsequent experiments for the sake of readability. In Figure 1b, we vary the external noise level ϵ . Given $n(\ln n)^4$ comparisons, AddS-3 exactly recovers the planted clusters for ϵ as small as 0.25 (high crowd noise) while, given the same number of comparisons, MulK-3 only recovers the planted clusters for $\epsilon > 0.9$. Figure 1c shows that AddS-3 outperforms MulK-3 even when different distributions for F_{in} and F_{out} are considered (Uniform + Beta or Uniform + Normal; details in Appendix E). It also shows that the distributions affect the performances, which is not evident from Theorem 1, indicating the possibility of a refined analysis under distributional assumptions.

MNIST clustering with comparisons. We consider two datasets which are subsets of the MNIST test dataset (LeCun and Cortes, 2010) that originally contains 10000 examples roughly equally distributed among the ten digits: (i) a subset of 2163 examples containing all the 1 and 7 (*MNIST 1vs.7*), two digits that are visually very similar, and (ii) a randomly selected subset of 2000 examples drawn without replacement and covering all 10 classes (*MNIST 10*). In both cases, to generate the comparisons, we use the Gaussian similarity (See Section F.3 in the supplementary) on a 2-dimensional embedding of the entire MNIST test data constructed with t-SNE (van der Maaten, 2014) and normalized so that each example lies in $[-1, 1]^2$. We focus on the triplet setting and we randomly and uniformly draw, without replacement, between $n(\ln n)^2$ and $n(\ln n)^4$ comparisons to be observed by the different approaches. We also consider two additional baselines. First, we

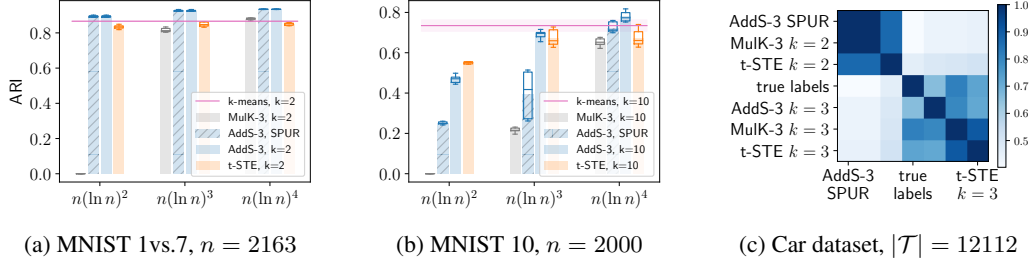


Figure 2: Experiments on real datasets. (2a)–(2b) ARI on MNIST; (2c) ARI similarity matrix comparing the clusters obtained by the different methods on car (darker means more agreement).

use t-STE (van der Maaten and Weinberger, 2012), an ordinal embedding approach, to embed the examples in 2 dimensions, and then cluster them using k -means on the embedded data. Second, we directly use k -means on the normalized data obtained with t-SNE. The latter is a baseline with access to Euclidean data instead of triplet comparisons.

For *MNIST 1vs.7* (Figure 2a), $|\mathcal{T}| = n(\ln n)^2$ is sufficient for AddS-3 to reach the performance of k -means and t-STE while MulK-3 requires $n(\ln n)^3$ triplets. Furthermore, note that AddS-3 with known number of clusters performs similarly to AddS-3 using SPUR, indicating that SPUR estimates the number of clusters correctly. If we consider *MNIST 10* (Figure 2b) and $|\mathcal{T}| = n(\ln n)^2$, AddS-3 with known k outperforms AddS-3 using SPUR, suggesting that the number of comparisons here is not sufficient to estimate the number of clusters accurately. Moreover, AddS-3 with known k outperforms MulK-3 while being close to the performance of t-STE. Finally for $n(\ln n)^4$ triplets, all ordinal methods converge to the baseline of k -means with access to original data. The ARI of AddS-3 SPUR improves when the number of comparisons increases due to better estimations of the number of clusters—estimated k increases from 3 for $|\mathcal{T}| = n(\ln n)^2$ up to 9 for $|\mathcal{T}| = n(\ln n)^4$.

Real comparison based data. First, we consider the Food dataset (Wilber et al., 2014) that contains 100 examples and 190376 triplet comparisons. Unfortunately, there is no ground truth and, thus, quantitatively assessing the quality of the obtained partitions is difficult. Thus, in Appendix F, we simply compare the different methods with respect to one another and present the partition obtained by AddS-3 for visual inspection. Second, we consider the Car dataset (Kleindessner and von Luxburg, 2016). It contains 60 examples grouped into 3 classes (SUV, city cars, sport cars) with 4 outliers, and exhibits 12112 triplet comparisons. For this dataset, AddS-3 SPUR estimates $k = 2$ instead of the correct 3 clusters. Figure 2c considers all ordinal methods with $k = 2$ and $k = 3$, and shows the pairwise agreement (ARI) between different methods and also with the true labels. While MulK-3 with $k = 3$ agrees the most with the true labels, all the clustering methods agree well for $k = 2$ (top-left 3×3 block). Hence, the data may have another natural clustering with two clusters, suggesting possible discrepancies in how different people judge the similarities between cars (for instance, color or brand instead of the specified classes).

6 Conclusion

It is generally believed that a large number of passive comparisons is necessary in comparison based learning. Existing results on clustering require at least $\Omega(n^3)$ passive comparisons in the worst-case or under a planted framework. We show that, in fact, $\Omega(n(\ln n)^2)$ passive comparisons suffice for accurately recovering planted clusters. This number of comparisons is near-optimal, and almost matches the number of active comparisons typically needed for learning. Our theoretical findings are based on two simple approaches for constructing pairwise similarity matrices from passive comparisons (AddS-3 and AddS-4). The present analysis is in a restricted framework, where all within (or inter) cluster similarities are assumed to be identically distributed. Based on existing work on robustness of SDPs (Moitra et al., 2016), we believe that our theoretical result holds in a more general semi-random setting. Lastly, while we studied the merits of AddS-3 and AddS-4 in the context of clustering, they could be used for other problems such as semi-supervised learning, data embedding, or classification.

Broader Impact

This work primarily has applications in the fields of psychophysics and crowdsourcing, and more generally, in learning from human responses. Such data and learning problems could be affected by implicit biases in human responses. However, this latter issue is beyond the scope of this work and, thus, was not formally analysed.

Acknowledgments and Disclosure of Funding

The work of DG is partly supported by the Baden-Württemberg Stiftung through the BW Eliteprogramm for postdocs. The work of MP has been supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

References

- E. Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.
- S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie. Generalized non-metric multidimensional scaling. In *International Conference on Artificial Intelligence and Statistics*, pages 11–18, 2007.
- C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2015.
- E. Amid and A. Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*, pages 1472–1480, 2015.
- E. Arias-Castro. Some theory for ordinal embedding. *Bernoulli*, 23(3):1663–1693, 2017.
- D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, page 1027–1035, 2007.
- I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.
- X. Chen and Y. Yang. Diffusion k-means clustering on manifolds: provable exact recovery via semidefinite relaxations. *Applied and Computational Harmonic Analysis*, 2020.
- Y. Chen and J. Xu. Statistical-computational phase transitions in planted models: The high-dimensional setting. In *International Conference on Machine Learning*, pages 244–252, 2014.
- S. Dasgupta. A cost function for similarity-based hierarchical clustering. In *Symposium on Theory of Computing*, pages 118–127, 2016.
- E. Emamjomeh-Zadeh and D. Kempe. Adaptive hierarchical clustering using ordinal queries. In *Symposium on Discrete Algorithms*, pages 415–429, 2018.
- Alyson K Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, 2009.
- D. Ghoshdastidar, M. Perrot, and U. von Luxburg. Foundations of comparison-based hierarchical clustering. In *Advances in Neural Information Processing Systems*, pages 7454–7464, 2019.
- S. Haghir, D. Ghoshdastidar, and U. von Luxburg. Comparison-based nearest neighbor search. In *International Conference on Artificial Intelligence and Statistics*, pages 851–859, 2017.
- S. Haghir, D. Garreau, and U. von Luxburg. Comparison-based random forests. In *International Conference on Machine Learning*, pages 1866–1875, 2018.
- L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.
- L. Jain, K. G. Jamieson, and R. Nowak. Finite sample prediction and recovery bounds for ordinal embedding. In *Advances in Neural Information Processing Systems*, pages 2711–2719, 2016.

- K. G. Jamieson and R. D. Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1077–1084, 2011.
- S. Janson and A. Ruciński. The infamous upper tail. *Random Structures & Algorithms*, 20(3): 317–342, 2002.
- M. Kleindessner and U. von Luxburg. Dimensionality estimation without distances. In *International Conference on Artificial Intelligence and Statistics*, pages 471–479, 2015.
- M. Kleindessner and U. von Luxburg. Lens depth function and k-relative neighborhood graph: versatile tools for ordinal data analysis, 2016.
- M. Kleindessner and U. von Luxburg. Kernel functions based on triplet similarity comparisons. In *Advances in Neural Information Processing Systems*, pages 6807–6817, 2017.
- Y. LeCun and C. Cortes. MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>, 2010.
- B. Mason, L. Jain, and R. Nowak. Learning low-dimensional metrics. In *Advances in neural information processing systems*, pages 4139–4147, 2017.
- Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841, 2016.
- R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.
- J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.
- N. Stewart, G. D. A. Brown, and N. Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.
- O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. In *International Conference on Machine Learning*, pages 673–680, 2011.
- Y. Terada and U. von Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.
- J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- A. Ukkonen. Crowdsourced correlation clustering with relative distance comparisons. *arXiv preprint arXiv:1709.08459*, 2017.
- A. Ukkonen, B. Derakhshan, and H. Heikinheimo. Crowdsourced nonparametric density estimation using relative distances. In *AAAI Conference on Human Computation and Crowdsourcing*, 2015.
- L. van der Maaten and K. Weinberger. Stochastic triplet embedding. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.
- Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.
- S. Vikram and S. Dasgupta. Interactive bayesian hierarchical clustering. In *International Conference on Machine Learning*, pages 2081–2090, 2016.
- U. von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- M. Wilber, S. Kwak, and S. Belongie. Cost-effective hits for relative similarity comparisons. In *Human Computation and Crowdsourcing (HCOMP)*, Pittsburgh, 2014.
- M. Xu, V. Jog, and P.-L. Loh. Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics*, 48(1):183–204, 2020.

- B. Yan and P. Sarkar. On robustness of kernel clustering. In *Advances in Neural Information Processing Systems*, pages 3098–3106, 2016.
- B. Yan, P. Sarkar, and X. Cheng. Provable estimation of the number of blocks in block models. In *International Conference on Artificial Intelligence and Statistics*, pages 1185–1194, 2018.
- F. W. Young. *Multidimensional scaling: History, theory, and applications*. Lawrence Erlbaum Associates, 1987.
- S.-Y. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.
- L. Zhang, S. Maji, and R. Tomioka. Jointly learning multiple measures of similarities from triplet comparisons. *arXiv preprint arXiv:1503.01521*, 2015.

A Existing comparison based similarities / kernel functions

The literature on ordinal embedding from triplet comparisons is extensive (Jamieson and Nowak, 2011; Arias-Castro, 2017). In contrast, the idea of directly constructing similarity or kernel matrices from the comparisons, without embedding the data in an Euclidean space, is rather new. Such an approach is known to be significantly faster than embedding methods, and provides similar or sometimes better performances in certain learning tasks. To the best of our knowledge, there are only two works that learn kernel functions from comparisons (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019), while the works of Jain et al. (2016) and Mason et al. (2017) estimate a Gram (or kernel) matrix from the triplets, which is then further used for data embedding. In this section, we describe the aforementioned approaches for constructing pairwise similarities from comparisons. Through this discussion, we illustrate the fundamental difference between the proposed additive similarities, AddS-3 and AddS-4, and the existing kernels that are of multiplicative nature (Kleindessner and von Luxburg, 2017; Ghoshdastidar et al., 2019).

Kernels from ordinal data were introduced by Kleindessner and von Luxburg (2017), who proposed two kernel functions (named k_1 and k_2) based on observed triplets. The kernels originated from the notion of Kendall’s τ correlation between two rankings, and k_1 was empirically observed to perform slightly better. We mention this kernel function, which we refer to as a multiplicative triplet kernel (MulK-3). For any distinct $i, j \in [n]$, the MulK-3 similarity is computed as

$$S_{ij} = \frac{\sum_{r < s} (\mathbb{I}_{\{(i,r,s) \in \mathcal{T}\}} - \mathbb{I}_{\{(i,s,r) \in \mathcal{T}\}}) (\mathbb{I}_{\{(j,r,s) \in \mathcal{T}\}} - \mathbb{I}_{\{(j,s,r) \in \mathcal{T}\}})}{\sqrt{|\{(\ell, r, s) \in \mathcal{T} : \ell = i\}|} \sqrt{|\{(\ell, r, s) \in \mathcal{T} : \ell = j\}|}} \quad (\text{MulK-3})$$

where \mathcal{T} is the set of observed triplets. Note that this kernel does not consider comparisons involving w_{ij} but, instead, uses multiplicative terms indicating how i and j behave with respect to every pair r, s . For uniform sampling with rate $p \gg \frac{\ln n}{n^2}$, the denominators in MulK-3 are approximately $p \binom{n}{2}$ for every $i \neq j$. Hence, it suffices to focus only on the numerator. Ghoshdastidar et al. (2019) proposed a kernel similar to MulK-3 for the case of quadruplets, which is referred to as multiplicative quadruplet kernel (MulK-4). For $i \neq j$, it is given by

$$S_{ij} = \sum_{\ell \neq i, j} \sum_{r < s} (\mathbb{I}_{\{(i,\ell,r,s) \in \mathcal{Q}\}} - \mathbb{I}_{\{(r,s,i,\ell) \in \mathcal{Q}\}}) (\mathbb{I}_{\{(j,\ell,r,s) \in \mathcal{Q}\}} - \mathbb{I}_{\{(r,s,j,\ell) \in \mathcal{Q}\}}). \quad (\text{MulK-4})$$

Ghoshdastidar et al. (2019) studied MulK-4 in the context of hierarchical clustering, and showed that it requires $\mathcal{O}(n^{3.5} \ln n)$ passive quadruplet comparisons to exactly recover a planted hierarchical structure in the data. Combining their concentration results with Proposition 1 shows that the same number of passive quadruplets suffices to recover the planted clusters considered in this work. Note that both MulK-3 and MulK-4 kernel functions have a multiplicative nature since each entry is an aggregate of products. This is essential for their positive semi-definite property. In contrast, the proposed AddS-3 and AddS-4 similarities simply aggregate comparisons involving the pairwise similarity w_{ij} , and hence, are not positive semi-definite kernels.

We also mention the work on fast ordinal triplet embedding (FORTE) (Mason et al., 2017), which learns a metric from the given triplet comparisons. One can easily adapt the formulation to that of

learning a kernel matrix $K \in \mathbb{R}^{n \times n}$ from triplets. Consider the squared distance in the corresponding reproducing kernel Hilbert space (RKHS), $d_K^2(i, j) = K_{ii} - 2K_{ij} + K_{jj}$. Assuming that the triplets adhere to the distance relation in the RKHS, it is easy to see that when a comparison of $t = \{i, r, s\}$ with i as pivot is available, then

$$\begin{aligned} y_t &:= \mathbb{I}_{\{(i,r,s) \in \mathcal{T}\}} - \mathbb{I}_{\{(i,s,r) \in \mathcal{T}\}} = \text{sign}(d_K^2(i, r) - d_K^2(i, s)) \\ &= \text{sign}(K_{rr} - 2K_{ir} - K_{ss} + 2K_{is}), \end{aligned}$$

which is the sign of a linear map of K , which we can denote as $\text{sign}(\langle M_t, K \rangle)$ for some $M_t \in \mathbb{R}^{n \times n}$. One can learn the optimal kernel matrix, that satisfies most triplet comparisons, by minimising the empirical loss $\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \ell(y_t \langle M_t, K \rangle)$ with positive definiteness constraints for K , where ℓ is a loss function (log loss is suggested by Jain et al. (2016)).

B Proof of Proposition 1

In this section, we first provide a proof of Proposition 1 which is split into two parts: the proof of optimality of X^* , and the proof of uniqueness of the optimal solution. In addition, we provide a derivation for the claim that X^* is the unique optimal solution for SDP- λ when $S = \tilde{S}$ and $0 < \lambda < n_{\min} \Delta_1$. The derivation, given at the end of the section, follows from simplifying some of the computations in the proof of Proposition 1.

B.1 Optimality of X^* when S is close to \tilde{S}

The proof is adapted from Yan et al. (2018). We first state the Karush-Kahn-Tucker (KKT) conditions for SDP- λ . Let $\Gamma, \Lambda \in \mathbb{R}^{n \times n}$ be the Lagrange parameters for the non-negativity constraint ($X \geq 0$) and the positive semi-definiteness constraint ($X \succeq 0$), respectively. Let $\alpha \in \mathbb{R}^n$ be the Lagrange parameter for the row sum constraints. The tuple $(X, \Lambda, \Gamma, \alpha)$ is a primal-dual optimal solution for SDP- λ if and only if it satisfies the following KKT conditions:

$$\begin{aligned} \text{Stationarity :} & & S - \lambda I + \Lambda + \Gamma - \mathbf{1}\alpha^T - \alpha\mathbf{1}^T &= 0 \\ \text{Primal feasibility :} & & X \geq 0 \quad ; \quad X \succeq 0 \quad ; \quad X\mathbf{1} = \mathbf{1} \\ \text{Dual feasibility :} & & \Lambda \succeq 0 \quad ; \quad \Gamma \geq 0 \\ \text{Complementary slackness :} & & \langle \Lambda, X \rangle = 0 \quad ; \quad \Gamma_{ij} X_{ij} = 0 \quad \forall i, j \end{aligned}$$

where we use $\langle A, B \rangle$ to denote trace (AB) for symmetric matrices A, B . The above derivation is straightforward. The term $\mathbf{1}\alpha^T + \alpha\mathbf{1}^T$ in the stationarity condition arises due to the symmetry of X , that is, since row-sum and column-sum are identical. We construct a primal-dual witness to show that X^* is the optimal solution of SDP- λ under the stated conditions on λ . We use the following notations. For any vector $u \in \mathbb{R}^n$ and $\mathcal{C} \subset [n]$, we let $u_{\mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|}$ be the projection of u on the indices contained in \mathcal{C} . Similarly, for a matrix $A \in \mathbb{R}^{n \times n}$, $A_{\mathcal{C}\mathcal{C}'}$ is the sub-matrix corresponding to row indices in \mathcal{C} and column indices in \mathcal{C}' . We also define $\mathbf{1}_m$ and I_m the constant vector of ones and the identity matrix of size m , respectively. We use $\mathcal{C}_1, \dots, \mathcal{C}_k$ to denote the planted clusters of size n_1, \dots, n_k . We consider the following primal-dual construction that is similar to Yan et al. (2018), where $X = X^*$. For every $j, \ell \in \{1, \dots, k\}$ and $\ell \neq j$, we define

$$\alpha : \quad \alpha_{\mathcal{C}_j} = \frac{1}{n_j} S_{\mathcal{C}_j \mathcal{C}_j} \mathbf{1}_{n_j} - \left(\frac{\lambda}{2n_j} + \frac{1}{2n_j^2} \mathbf{1}_{n_j}^T S_{\mathcal{C}_j \mathcal{C}_j} \mathbf{1}_{n_j} \right) \mathbf{1}_{n_j} \quad (4)$$

$$\Lambda : \quad \begin{cases} \Lambda_{\mathcal{C}_j \mathcal{C}_j} = -S_{\mathcal{C}_j \mathcal{C}_j} + \alpha_{\mathcal{C}_j} \mathbf{1}_{n_j}^T + \mathbf{1}_{n_j} \alpha_{\mathcal{C}_j}^T + \lambda I_{n_j} \\ \Lambda_{\mathcal{C}_j \mathcal{C}_\ell} = - \left(I_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T \right) S_{\mathcal{C}_j \mathcal{C}_\ell} \left(I_{n_\ell} - \frac{1}{n_\ell} \mathbf{1}_{n_\ell} \mathbf{1}_{n_\ell}^T \right) \end{cases} \quad (5)$$

$$\Gamma : \quad \begin{cases} \Gamma_{\mathcal{C}_j \mathcal{C}_j} = 0 \\ \Gamma_{\mathcal{C}_j \mathcal{C}_\ell} = -S_{\mathcal{C}_j \mathcal{C}_\ell} - \Lambda_{\mathcal{C}_j \mathcal{C}_\ell} + \alpha_{\mathcal{C}_j} \mathbf{1}_{n_\ell}^T + \mathbf{1}_{n_j} \alpha_{\mathcal{C}_\ell}^T. \end{cases} \quad (6)$$

The proof of Proposition 1 is based on verifying the KKT conditions for the above Λ, Γ, α and $X = X^*$. To this end, note that $X_{\mathcal{C}_j \mathcal{C}_j}^* = \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T$ and $X_{\mathcal{C}_j \mathcal{C}_\ell}^* = 0$ for $\ell \neq j$. Primal feasibility

is obviously satisfied by X^* , and it is easy to see that the choice of $\Lambda_{\mathcal{C}_j\mathcal{C}_j}$ and $\Gamma_{\mathcal{C}_j\mathcal{C}_\ell}$ ensures that stationarity holds. Hence, we only need to verify dual feasibility and complementary slackness.

The complementary slackness condition for Γ holds since $\Gamma_{\mathcal{C}_j\mathcal{C}_j} = 0$ and $X_{\mathcal{C}_j\mathcal{C}_\ell}^* = 0$ for $j \neq \ell$. To verify $\langle \Lambda, X^* \rangle = 0$, observe that

$$\begin{aligned} \langle \Lambda, X^* \rangle &= \sum_{j,\ell} \langle \Lambda_{\mathcal{C}_j\mathcal{C}_\ell}, X_{\mathcal{C}_j\mathcal{C}_\ell}^* \rangle = \sum_j \langle \Lambda_{\mathcal{C}_j\mathcal{C}_j}, X_{\mathcal{C}_j\mathcal{C}_j}^* \rangle = \sum_j \frac{1}{n_j} \mathbf{1}_{n_j}^T \Lambda_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j} \\ &= \sum_j -\frac{1}{n_j} \mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j} + 2\mathbf{1}_{n_j}^T \alpha_{\mathcal{C}_j} + \lambda, \end{aligned}$$

where the last step follows by substituting $\Lambda_{\mathcal{C}_j\mathcal{C}_j}$ from (5) and noting that $\mathbf{1}_{n_j}^T \mathbf{1}_{n_j} = n_j$. Substituting the value of $\alpha_{\mathcal{C}_j}$ above shows that each term in the sum is zero, and hence, $\langle \Lambda, X^* \rangle = 0$.

We now verify the dual feasibility and first prove that $\Gamma \geq 0$, in particular, $\Gamma_{\mathcal{C}_j\mathcal{C}_\ell} \geq 0$ for $j \neq \ell$. We substitute $\Lambda_{\mathcal{C}_j\mathcal{C}_\ell}$ and $\alpha_{\mathcal{C}_j}$ in (6) to obtain

$$\begin{aligned} \Gamma_{\mathcal{C}_j\mathcal{C}_\ell} &= -S_{\mathcal{C}_j\mathcal{C}_\ell} + \left(I_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T \right) S_{\mathcal{C}_j\mathcal{C}_\ell} \left(I_{n_\ell} - \frac{1}{n_\ell} \mathbf{1}_{n_\ell} \mathbf{1}_{n_\ell}^T \right) \\ &\quad + \frac{1}{n_j} S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T - \left(\frac{\lambda}{2n_j} + \frac{1}{2n_j^2} \mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j} \right) \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T \\ &\quad + \frac{1}{n_\ell} \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell\mathcal{C}_\ell} - \left(\frac{\lambda}{2n_\ell} + \frac{1}{2n_\ell^2} \mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell\mathcal{C}_\ell} \mathbf{1}_{n_\ell} \right) \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T \\ &= -\frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_\ell} - \frac{1}{n_\ell} S_{\mathcal{C}_j\mathcal{C}_\ell} \mathbf{1}_{n_\ell} \mathbf{1}_{n_\ell}^T + \frac{1}{n_j} S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T + \frac{1}{n_\ell} \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell\mathcal{C}_\ell} \\ &\quad + \left(\frac{\mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_\ell}}{n_j n_\ell} - \frac{\lambda}{2n_j} - \frac{\mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j}}{2n_j^2} - \frac{\lambda}{2n_\ell} - \frac{\mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell\mathcal{C}_\ell} \mathbf{1}_{n_\ell}}{2n_\ell^2} \right) \mathbf{1}_{n_j} \mathbf{1}_{n_\ell}^T. \end{aligned}$$

Consider $i \in \mathcal{C}_j$ and $r \in \mathcal{C}_\ell$. From above, we can compute Γ_{ir} as

$$\begin{aligned} \Gamma_{ir} &= -\frac{1}{n_j} \mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_\ell} - \frac{1}{n_\ell} S_{i\mathcal{C}_\ell} \mathbf{1}_{n_\ell} + \frac{1}{n_j} S_{i\mathcal{C}_j} \mathbf{1}_{n_j} + \frac{1}{n_\ell} \mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell\mathcal{C}_\ell} \\ &\quad + \frac{\mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_\ell} \mathbf{1}_{n_\ell}}{n_j n_\ell} - \frac{\mathbf{1}_{n_j}^T S_{\mathcal{C}_j\mathcal{C}_j} \mathbf{1}_{n_j}}{2n_j^2} - \frac{\mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell\mathcal{C}_\ell} \mathbf{1}_{n_\ell}}{2n_\ell^2} - \frac{\lambda}{2n_j} - \frac{\lambda}{2n_\ell} \\ &= -\frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} S_{i'r} - \frac{1}{n_\ell} \sum_{r' \in \mathcal{C}_\ell} S_{ir'} + \frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} S_{ii'} + \frac{1}{n_\ell} \sum_{r' \in \mathcal{C}_\ell} S_{rr'} \\ &\quad + \frac{1}{n_j n_\ell} \sum_{i' \in \mathcal{C}_j, r' \in \mathcal{C}_\ell} S_{i'r'} - \frac{1}{2n_j^2} \sum_{i, i' \in \mathcal{C}_j} S_{ii'} - \frac{1}{2n_\ell^2} \sum_{r, r' \in \mathcal{C}_\ell} S_{rr'} - \frac{\lambda}{2n_j} - \frac{\lambda}{2n_\ell}. \end{aligned}$$

Our goal is to derive a lower bound for Γ_{ir} and show that, for suitable values of λ , $\Gamma_{ir} \geq 0$ for all $i \in \mathcal{C}_j, r \in \mathcal{C}_\ell$. We bound each of the terms from below. For the last two terms involving λ , we note that both terms are at least $-\frac{\lambda}{2n_{\min}}$, where $n_{\min} = \min_\ell n_\ell$. For each of the other terms, we rewrite the summations in terms of the ideal similarity matrix \tilde{S} and bound the deviation in terms of

$$\Delta_2 = \max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{r \in \mathcal{C}_\ell} \left(S_{ir} - \tilde{S}_{ir} \right) \right|. \text{ For the first term, we have}$$

$$\begin{aligned} -\frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} S_{i'r} &= -\frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} \tilde{S}_{i'r} - \frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} \left(S_{i'r} - \tilde{S}_{i'r} \right) \\ &= -\Sigma_{j\ell} - \frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} \left(S_{i'r} - \tilde{S}_{i'r} \right) \\ &\geq -\Sigma_{j\ell} - \Delta_2. \end{aligned}$$

For the second inequality, we use the structure of \tilde{S} to note that $\tilde{S}_{ir} = \Sigma_{j\ell}$ for every $i \in \mathcal{C}_i, r \in \mathcal{C}_\ell$, and finally the deviation term is bounded by Δ_2 . Similarly, one can bound the second, third and fourth terms from below by $(-\Sigma_{j\ell} - \Delta_2)$, $(\Sigma_{jj} - \Delta_2)$ and $(\Sigma_{\ell\ell} - \Delta_2)$, respectively. For the fifth term, we write

$$\begin{aligned} \frac{1}{n_j n_\ell} \sum_{i' \in \mathcal{C}_j, r' \in \mathcal{C}_\ell} S_{i'r'} &= \frac{1}{n_j n_\ell} \sum_{i' \in \mathcal{C}_j, r' \in \mathcal{C}_\ell} \tilde{S}_{i'r'} + \frac{1}{n_j n_\ell} \sum_{i' \in \mathcal{C}_j, r' \in \mathcal{C}_\ell} (S_{i'r'} - \tilde{S}_{i'r'}) \\ &= \Sigma_{j\ell} + \frac{1}{n_j} \sum_{i' \in \mathcal{C}_j} \left(\frac{1}{n_\ell} \sum_{r' \in \mathcal{C}_\ell} (S_{i'r'} - \tilde{S}_{i'r'}) \right) \\ &\geq \Sigma_{j\ell} - \Delta_2, \end{aligned}$$

since each term in the outer summation is at least $-\Delta_2$. Similarly, one can bound the sixth and seventh terms from below by $\frac{1}{2}(\Sigma_{jj} - \Delta_2)$ and $\frac{1}{2}(\Sigma_{\ell\ell} - \Delta_2)$, respectively. Combining the above lower bounds, we have

$$\Gamma_{ir} \geq \frac{1}{2}\Sigma_{jj} + \frac{1}{2}\Sigma_{\ell\ell} - \Sigma_{j\ell} - 6\Delta_2 - \frac{\lambda}{n_{\min}} \geq (\Delta_1 - 6\Delta_2) - \frac{\lambda}{n_{\min}},$$

where we recall that $\Delta_1 = \min_{\ell \neq \ell'} \left(\frac{\Sigma_{\ell\ell} + \Sigma_{\ell'\ell'}}{2} - \Sigma_{\ell\ell'} \right)$. Hence, for $\lambda \leq n_{\min}(\Delta_1 - 6\Delta_2)$, as stated in Proposition 1, $\Gamma_{ir} \geq 0$, and more generally, Γ is non-negative.

We finally derive the positive semi-definiteness of Λ . Define the vectors $u_1, \dots, u_k \in \mathbb{R}^n$ such that $(u_\ell)_i = 1$ if $i \in \mathcal{C}_\ell$ and 0 otherwise. We first claim that u_1, \dots, u_k lie in the null space of Λ . To verify this, we compute the \mathcal{C}_j -th block of Λu_ℓ . For $j \neq \ell$,

$$(\Lambda u_\ell)_{\mathcal{C}_j} = \Lambda_{\mathcal{C}_j \mathcal{C}_\ell} \mathbf{1}_{n_\ell} = - \left(I_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T \right) S_{\mathcal{C}_j \mathcal{C}_\ell} \left(I_{n_\ell} - \frac{1}{n_\ell} \mathbf{1}_{n_\ell} \mathbf{1}_{n_\ell}^T \right) \mathbf{1}_{n_\ell} = 0,$$

whereas for $j = \ell$, we have from (4) and (5),

$$\begin{aligned} (\Lambda u_\ell)_{\mathcal{C}_\ell} &= \Lambda_{\mathcal{C}_\ell \mathcal{C}_\ell} \mathbf{1}_{n_\ell} \\ &= -S_{\mathcal{C}_\ell \mathcal{C}_\ell} \mathbf{1}_{n_\ell} + n_\ell \alpha_{\mathcal{C}_\ell} + \mathbf{1}_{n_\ell} \alpha_{\mathcal{C}_\ell}^T \mathbf{1}_{n_\ell} + \lambda \mathbf{1}_{n_\ell} \\ &= -S_{\mathcal{C}_\ell \mathcal{C}_\ell} \mathbf{1}_{n_\ell} + S_{\mathcal{C}_\ell \mathcal{C}_\ell} \mathbf{1}_{n_\ell} - 2 \left(\frac{\lambda}{2} + \frac{\mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell \mathcal{C}_\ell} \mathbf{1}_{n_\ell}}{2n_\ell} \right) \mathbf{1}_{n_\ell} + \mathbf{1}_{n_\ell} \frac{\mathbf{1}_{n_\ell}^T S_{\mathcal{C}_\ell \mathcal{C}_\ell} \mathbf{1}_{n_\ell}}{n_\ell} + \lambda \mathbf{1}_{n_\ell} \\ &= 0. \end{aligned}$$

Thus $\Lambda u_\ell = 0$ for $\ell = 1, \dots, k$, and to prove that $\Lambda \succeq 0$, it suffices to show that $u^T \Lambda u \geq 0$ for all $u \in \mathbb{R}^n$ that are orthogonal to u_1, \dots, u_k . In other words, we consider only u such that $u_{\mathcal{C}_\ell}^T \mathbf{1}_{n_\ell} = 0$ for every ℓ . For such a vector u , we have

$$\begin{aligned} u^T \Lambda u &= \sum_{j,\ell=1}^k u_{\mathcal{C}_j}^T \Lambda_{\mathcal{C}_j \mathcal{C}_\ell} u_{\mathcal{C}_\ell} = \sum_{j=1}^k u_{\mathcal{C}_j}^T \Lambda_{\mathcal{C}_j \mathcal{C}_j} u_{\mathcal{C}_j} + \sum_{j \neq \ell} u_{\mathcal{C}_j}^T \Lambda_{\mathcal{C}_j \mathcal{C}_\ell} u_{\mathcal{C}_\ell} \\ &= \sum_{j=1}^k u_{\mathcal{C}_j}^T (-S_{\mathcal{C}_j \mathcal{C}_j} + \lambda I_{n_j}) u_{\mathcal{C}_j} - \sum_{j \neq \ell} u_{\mathcal{C}_j}^T S_{\mathcal{C}_j \mathcal{C}_\ell} u_{\mathcal{C}_\ell} \\ &= \sum_{j=1}^k \lambda u_{\mathcal{C}_j}^T u_{\mathcal{C}_j} - \sum_{j,\ell} u_{\mathcal{C}_j}^T S_{\mathcal{C}_j \mathcal{C}_\ell} u_{\mathcal{C}_\ell} \\ &= \lambda \|u\|^2 - u^T S u, \end{aligned}$$

where $\|u\|$ is the Euclidean norm. The third equality follows from (5) and $u_{\mathcal{C}_\ell}^T \mathbf{1}_{n_\ell} = 0$ for every ℓ . In addition to above, recall that $\tilde{S} = Z \Sigma Z^T$, where $Z = [u_1 \dots u_k]$. Hence, for u orthogonal to u_1, \dots, u_k , we have $u^T \tilde{S} u = 0$, which, combined with above, gives

$$u^T \Lambda u = \lambda \|u\|^2 - u^T S u$$

$$\begin{aligned}
&= \lambda \|u\|^2 - u^T (S - \tilde{S}) u \\
&\geq \left(\lambda - \|S - \tilde{S}\|_2 \right) \|u\|^2 \\
&> 0
\end{aligned}$$

for all u if $\lambda > \|S - \tilde{S}\|_2$, which is the condition stated in Proposition 1. Thus, for the specified range of λ , the KKT conditions are satisfied and X^* is the optimal solution for SDP- λ .

B.2 Uniqueness of the optimal solution X^*

The uniqueness of the solution can be shown by proving that any other optimal solution X' for SDP- λ must satisfy $X' = X^*$. This is shown in two steps. First, we show that any optimal solution X' must have the same block structure as X^* and $X^* - X' \succeq 0$. We use this fact to show that the objective value for X^* is strictly greater than that for any such X' .

Note that the previously constructed Lagrange parameters in (4)–(6) need not correspond to the optimal solution associated with X' . However, for the previously defined α, Λ, Γ , we can still use the condition for stationarity to write

$$\begin{aligned}
\langle \Lambda + \Gamma, X' \rangle &= \langle -S + \mathbf{1}\alpha^T + \alpha\mathbf{1}^T + \lambda I, X' \rangle \\
&= -\langle S, X' \rangle + \sum_{i,j=1}^n (\alpha_i + \alpha_j) X'_{ij} + \lambda \text{trace}(X) \\
&= -\text{trace}(SX') + \lambda \text{trace}(X') + 2 \sum_{i=1}^n \alpha_i
\end{aligned}$$

where the simplification happens noting that X' is primal feasible and hence $\sum_j X'_{ij} = 1$. Due to optimality of X' and X^* , we have $\text{trace}(SX') - \lambda \text{trace}(X') = \text{trace}(SX^*) - \lambda \text{trace}(X^*)$, and so,

$$\begin{aligned}
\langle \Lambda + \Gamma, X' \rangle &= -\text{trace}(SX^*) + \lambda \text{trace}(X^*) + 2\mathbf{1}^T \alpha \\
&= \lambda + \sum_{j=1}^k \left(-\frac{\mathbf{1}_{n_j}^T S c_j c_j \mathbf{1}_{n_j}}{n_j} + 2\mathbf{1}_{n_j}^T \alpha c_j \right) = 0,
\end{aligned}$$

where the final step follows by substituting αc_j from (4). From above, we argue that both $\langle \Lambda, X' \rangle$ and $\langle \Gamma, X' \rangle$ are zero. To verify this, note that Γ and X' are both non-negative and hence, $\langle \Gamma, X' \rangle \geq 0$. On the other hand, from the definition of Frobenius (or Hilbert-Schmidt) norm, we have $\langle \Lambda, X' \rangle = \|\Lambda^{1/2} X'^{1/2}\|_F^2 \geq 0$, where the matrices square roots exist since Λ, X' are both positive semi-definite. Since both inner products, $\langle \Lambda, X' \rangle$ and $\langle \Gamma, X' \rangle$, are non-negative and yet their sum is zero, we can conclude that each of them equals zero.

Note that $\langle \Lambda, X' \rangle = \|\Lambda^{1/2} X'^{1/2}\|_F^2 = 0$ implies $\Lambda X' = 0$, or the range space of X' lies in the null space of Λ . Recall, from the proof of positive semi-definiteness of Λ , that, for $\lambda > \|S - \tilde{S}\|_2$, the null space of Λ is exactly spanned by $Z = [u_1 \dots u_k]$. Thus, the range space of X' is spanned by the columns of Z , or in other words $X' = ZAZ^T$ for some $A \in \mathbb{R}^{k \times k}$ that is symmetric, non-negative, and positive semi-definite (to ensure that X' is primal feasible), and $\sum_j A_{ij} n_j = 1$ (to satisfy the row sum constraint). Recall that $X^* = ZN^{-1}Z^T$ where $N = \text{diag}(n_1, \dots, n_k)$. Thus, X' has the same block structure as X^* . However, this result does not imply that we can recover k planted clusters from X' since it is possible that A has less than k distinct rows.

We now argue that $X^* - X'$ must be positive semi-definite, a property that we use later. To see this, note that

$$X^* - X' = ZN^{-1/2} \left(I_k - N^{1/2} A N^{1/2} \right) N^{-1/2} Z^T,$$

where $ZN^{-1/2}$ is a matrix with orthonormal columns. Hence, to prove that $X^* - X' \succeq 0$, it suffices to show that $I_k - N^{1/2} A N^{1/2} \succeq 0$ or, equivalently, that the largest eigenvalue of $N^{1/2} A N^{1/2}$ is smaller than 1. This can be verified as

$$\left\| N^{1/2} A N^{1/2} \right\|_2 = \max_{u: \|u\|=1} u^T N^{1/2} A N^{1/2} u = \max_{u: \|u\|=1} \sum_{i,j=1}^k A_{ij} \sqrt{n_i n_j} u_i u_j.$$

From the AM-GM inequality, we have $\sqrt{n_i n_j} u_i u_j \leq \frac{1}{2} (n_i u_j^2 + n_j u_i^2)$. Hence,

$$\left\| N^{1/2} A N^{1/2} \right\|_2 \leq \max_{u: \|u\|=1} \frac{1}{2} \sum_{i,j=1}^k A_{ij} (n_i u_j^2 + n_j u_i^2) = \sum_{i=1}^k u_i^2 = 1,$$

where we use the fact that $\sum_j A_{ij} n_j = \sum_i A_{ij} n_i = 1$. From this discussion, we have $X^* - X' \succeq 0$. We now claim that

$$\left| \text{trace} \left((S - \tilde{S})(X^* - X') \right) \right| \leq \left\| S - \tilde{S} \right\|_2 \text{trace} (X^* - X'), \quad (7)$$

which follows from von Neumann's trace inequality and the fact that $X^* - X'$ is positive semi-definite.

We now prove that for any $X' = ZAZ^T \neq X^*$, with A satisfying the above mentioned conditions, and for $\|S - \tilde{S}\|_2 < \lambda < \frac{1}{2} \Delta_1 n_{\min}$,

$$\text{trace} (SX^*) - \lambda \text{trace} (X^*) > \text{trace} (SX') - \lambda \text{trace} (X'), \quad (8)$$

which shows that X^* is the unique optimal solution. We compute

$$\begin{aligned} & \text{trace} (SX^*) - \lambda \text{trace} (X^*) - \text{trace} (SX') + \lambda \text{trace} (X') \\ &= \text{trace} (S(X^* - X')) - \lambda \text{trace} (X^* - X') \\ &= \text{trace} \left(\tilde{S}(X^* - X') \right) + \text{trace} \left((S - \tilde{S})(X^* - X') \right) - \lambda \text{trace} (X^* - X') \\ &> \text{trace} \left(\tilde{S}(X^* - X') \right) - \left\| S - \tilde{S} \right\|_2 \text{trace} (X^* - X') - \lambda \text{trace} (X^* - X') \\ &> \text{trace} \left(\tilde{S}(X^* - X') \right) - n_{\min} \Delta_1 \text{trace} (X^* - X'). \end{aligned}$$

In the last step, we use $\left\| S - \tilde{S} \right\|_2 + \lambda < 2\lambda < n_{\min} \Delta_1$. We later prove that

$$\text{trace} \left(\tilde{S}(X^* - X') \right) \geq \sum_{\ell=1}^k n_\ell (1 - A_{\ell\ell} n_\ell) \Delta_1 \geq n_{\min} \Delta_1 \text{trace} (X^* - X'). \quad (9)$$

Using (9) in the previous derivation proves (8) or the fact that X^* is the unique optimal solution, provided that $\text{trace} (X^* - X') > 0$ for all $X' \neq X^*$. Hence, we need to verify the strict positivity of the trace. Assume that $\text{trace} (X^* - X') = 0$. Due to the row sum constraint for X' , we have $\sum_j A_{\ell j} n_j = 1$, which implies $A_{\ell\ell} n_\ell \leq 1$. On the other hand $\text{trace} (X') = \text{trace} (X^*) = k$ holds if $\sum_\ell A_{\ell\ell} n_\ell = k$, which is only possible if $A_{\ell\ell} n_\ell = 1$ for every ℓ , and hence $A_{\ell j} = 0$ for $j \neq \ell$. Thus, $\text{trace} (X^* - X') = 0$ if and only if $X' = X^*$. For every $X' \neq X^*$, we have $\text{trace} (X^* - X') = \sum_\ell (1 - A_{\ell\ell} n_\ell) > 0$. We conclude the proof by proving (9). We compute

$$\begin{aligned} \text{trace} \left(\tilde{S}(X^* - X') \right) &= \text{trace} (Z \Sigma Z^T Z (N^{-1} - A) Z^T) \\ &= \text{trace} (\Sigma Z^T Z (N^{-1} - A) Z^T Z) \\ &= \text{trace} (\Sigma N (N^{-1} - A) N) \\ &= \sum_{\ell=1}^k \Sigma_{\ell\ell} n_\ell (1 - A_{\ell\ell} n_\ell) - \sum_{\ell=1}^k \sum_{j \neq \ell} A_{\ell j} n_j n_\ell \Sigma_{\ell j}, \end{aligned}$$

where the third equality follows from the fact $Z^T Z = N$. Recall from the definition of Δ_1 that $\Sigma_{\ell j} \leq \frac{1}{2} (\Sigma_{jj} + \Sigma_{\ell\ell}) - \Delta_1$. Using this, we can write

$$\begin{aligned} & \text{trace} \left(\tilde{S}(X^* - X') \right) \\ & \geq \sum_{\ell=1}^k \Sigma_{\ell\ell} n_\ell (1 - A_{\ell\ell} n_\ell) + \sum_{\ell=1}^k \sum_{j \neq \ell} A_{\ell j} n_j n_\ell \Delta_1 - \frac{1}{2} \sum_{\ell=1}^k \sum_{j \neq \ell} A_{\ell j} n_j n_\ell (\Sigma_{\ell\ell} + \Sigma_{jj}) \\ & = \sum_{\ell=1}^k \Sigma_{\ell\ell} n_\ell (1 - A_{\ell\ell} n_\ell) + \sum_{\ell=1}^k \sum_{j \neq \ell} A_{\ell j} n_j n_\ell \Delta_1 - \sum_{\ell=1}^k \sum_{j \neq \ell} A_{\ell j} n_j n_\ell \Sigma_{\ell\ell} \end{aligned}$$

$$= \sum_{\ell=1}^k \Sigma_{\ell\ell} n_{\ell} (1 - A_{\ell\ell} n_{\ell}) + \sum_{\ell=1}^k n_{\ell} \Delta_1 (1 - A_{\ell\ell} n_{\ell}) - \sum_{\ell=1}^k n_{\ell} \Sigma_{\ell\ell} (1 - A_{\ell\ell} n_{\ell}).$$

In the first equality, we exploit the symmetry of the third summation, while the second equality uses the row sum constraint to write $\sum_{j \neq \ell} A_{\ell j} n_j = 1 - A_{\ell\ell} n_{\ell}$. Cancelling first and third terms, we get

$$\begin{aligned} \text{trace} \left(\tilde{S}(X^* - X') \right) &\geq \Delta_1 \sum_{\ell=1}^k n_{\ell} (1 - A_{\ell\ell} n_{\ell}) \\ &\geq \Delta_1 n_{\min} \sum_{\ell=1}^k (1 - A_{\ell\ell} n_{\ell}) = n_{\min} \Delta_1 \text{trace} (X^* - X'), \end{aligned}$$

which proves (9), and completes the proof.

B.3 Unique optimality of X^* when $S = \tilde{S}$

We now prove that X^* is the unique optimal solution when $S = \tilde{S} = Z\Sigma Z^T$ and $0 < \lambda < n_{\min} \Delta_1$. This claim does not immediately follow from Proposition 1, but can be derived from the proof.

We first prove the optimality of X^* in this case. Recall, from the proof of Proposition 1, that the claim hinges on showing that $\Gamma \geq 0$ and $\Lambda \succeq 0$. From the previous proof, it suffices to show that $\Gamma_{C_j C_{\ell}} \geq 0$ and $u^T \Lambda u \geq 0$ for any u that is orthogonal to the columns of Z . To show that the latter holds, recall that

$$u^T \Lambda u = \lambda \|u\|^2 - u^T S u.$$

Since $S = \tilde{S} = Z\Sigma Z^T$ and $Z^T u = 0$, we get $u^T \Lambda u = \lambda \|u\|^2 \geq 0$, which in turn shows that $\Lambda \succeq 0$ for all $\lambda > 0$. To verify the non-negativity of $\Gamma_{C_j C_{\ell}}$, we observe that, in this case, it can be computed as

$$\begin{aligned} \Gamma_{C_j C_{\ell}} &= -\frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T \tilde{S}_{C_j C_{\ell}} - \frac{1}{n_{\ell}} \tilde{S}_{C_j C_{\ell}} \mathbf{1}_{n_{\ell}} \mathbf{1}_{n_{\ell}}^T + \frac{1}{n_j} \tilde{S}_{C_j C_j} \mathbf{1}_{n_j} \mathbf{1}_{n_{\ell}}^T + \frac{1}{n_{\ell}} \mathbf{1}_{n_j} \mathbf{1}_{n_{\ell}}^T \tilde{S}_{C_{\ell} C_{\ell}} \\ &\quad + \left(\frac{\mathbf{1}_{n_j}^T \tilde{S}_{C_j C_{\ell}} \mathbf{1}_{n_{\ell}}}{n_j n_{\ell}} - \frac{\lambda}{2n_j} - \frac{\mathbf{1}_{n_j}^T \tilde{S}_{C_j C_j} \mathbf{1}_{n_j}}{2n_j^2} - \frac{\lambda}{2n_{\ell}} - \frac{\mathbf{1}_{n_{\ell}}^T \tilde{S}_{C_{\ell} C_{\ell}} \mathbf{1}_{n_{\ell}}}{2n_{\ell}^2} \right) \mathbf{1}_{n_j} \mathbf{1}_{n_{\ell}}^T \\ &= \left(-2\Sigma_{j\ell} + \Sigma_{jj} + \Sigma_{\ell\ell} + \Sigma_{j\ell} - \frac{\lambda}{2} \left(\frac{1}{n_j} + \frac{1}{n_{\ell}} \right) - \frac{\Sigma_{jj} + \Sigma_{\ell\ell}}{2} \right) \mathbf{1}_{n_j} \mathbf{1}_{n_{\ell}}^T \\ &\geq \left(\Delta_1 - \frac{\lambda}{n_{\min}} \right) \mathbf{1}_{n_j} \mathbf{1}_{n_{\ell}}^T \end{aligned}$$

So for $\lambda \leq n_{\min} \Delta_1$, $\Gamma_{C_j C_{\ell}} \geq 0$, and hence Γ is non-negative. Combining this with the previous proof of optimality, we derive that X^* is an optimal solution in this case for $0 < \lambda \leq n_{\min} \Delta_1$.

The proof of uniqueness is similar to the more general case in Proposition 1. We use the previously derived claim that any optimal solution X' must be of the form $X' = ZAZ^T$ for some $A \in \mathbb{R}^{k \times k}$, and $X^* - X \succeq 0$. We have also previously shown that $\text{trace} \left(\tilde{S}(X^* - X') \right) \geq n_{\min} \Delta_1 \text{trace} (X^* - X')$. Hence, we have

$$\text{trace} \left(\tilde{S} X^* \right) - \lambda \text{trace} (X^*) - \left(\text{trace} \left(\tilde{S} X' \right) - \lambda \text{trace} (X') \right) \geq (n_{\min} \Delta_1 - \lambda) \text{trace} (X^* - X'),$$

which is strictly positive for $\lambda < n_{\min} \Delta_1$, and hence, X^* is the unique optimal solution in this case.

C Proof of Theorem 1

We prove the result for triplets and quadruplets in separate sections. While the proof structure is the same in both cases, the computations are quite different. Before presenting the proofs, we list the key steps.

We first compute the expectation of the similarity matrix S computed using AddS-3 or AddS-4, and derive appropriate ideal matrices \tilde{S} in each case. In our proofs, $\tilde{S} = \mathbf{E}[S]$, except differences in the diagonal entries since $S_{ii} = 0$ for all i . From the block structure of \tilde{S} , we can compute Δ_1 .

Subsequently, concentration inequalities are used to derive upper bounds on $\|S - \tilde{S}\|_2$ and Δ_2 in terms of the model parameters. In this context, note that though the pairwise similarities $\{w_{ij} : i < j\}$ are independent, the entries of the matrix S are highly dependent since each w_{ij} appears in multiple entries of S . Hence, to decouple such dependencies, we use a technique by Janson and Ruciński (2002), which considers the dependency graphs of the random variables and finds an equitable colouring to find independent sets of comparable sizes. To the best of our knowledge, the present work is the first study which uses the equitable colouring approach of Janson and Ruciński (2002) to derive spectral norm bounds. Ghoshdastidar et al. (2019) use this technique only to bound matrix entries.

Finally, we use concentration to show that for a sampling rate p large enough, the number of comparisons ($|\mathcal{Q}|$ or $|\mathcal{T}|$) is close to its expected value. Hence, we can replace the sampling rate p in the previously derived bounds by the number of comparisons, leading to differences in constants only.

Notation. For the sake of simplicity, we will ignore absolute constants in the inequalities stated below, and use the notations \lesssim and \gtrsim to write inequalities that hold up to some multiplicative absolute constant.

C.1 Quadruplet setting

We first present the proof for the quadruplet setting using the aforementioned steps.

Computation of Δ_1 . We first derive the expectation of the AddS-4 similarity matrix S , where for $i \neq j$,

$$\mathbf{E}[S_{ij}] = \sum_{r \neq s} \mathbf{P}((i, j, r, s) \in \mathcal{Q}) - \mathbf{P}((r, s, i, j) \in \mathcal{Q}). \quad (10)$$

Note that the summation in AddS-4 is a sum over all distinct pairs r, s , noting that we do not count both (s, r) and (r, s) since they refer to the same comparison. To compute the expectation of each term in the summation, recall that the items belong to the planted clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$ and, for each item i , use $\psi_i \in [k]$ to denote the cluster index in which i belongs, that is, $i \in \mathcal{C}_{\psi_i}$. The expected values of the terms are given in Table 1.

Table 1: Value of each term in the summation in (10), assuming $i \neq j, r \neq s, (i, j) \neq (r, s)$.

Case	$\mathbf{P}((i, j, r, s) \in \mathcal{Q})$	$\mathbf{P}((r, s, i, j) \in \mathcal{Q})$	Difference
$\psi_i = \psi_j; \psi_r = \psi_s$	$p/2$	$p/2$	0
$\psi_i = \psi_j; \psi_r \neq \psi_s$	$p(1 + \epsilon\delta)/2$	$p(1 - \epsilon\delta)/2$	$p\epsilon\delta$
$\psi_i \neq \psi_j; \psi_r = \psi_s$	$p(1 - \epsilon\delta)/2$	$p(1 + \epsilon\delta)/2$	$-p\epsilon\delta$
$\psi_i \neq \psi_j; \psi_r \neq \psi_s$	$p/2$	$p/2$	0

We only explain the derivation of $\mathbf{P}((i, j, r, s) \in \mathcal{Q})$ for the case $\psi_i = \psi_j$ and $\psi_r \neq \psi_s$ as the other values are computed similarly. In this case,

$$\begin{aligned} \mathbf{P}((i, j, r, s) \in \mathcal{Q}) &= \mathbf{P}((i, j, r, s) \in \mathcal{Q} | (i, j), (r, s) \text{ compared}) \mathbf{P}((i, j), (r, s) \text{ compared}) \\ &= p \mathbf{P}((i, j, r, s) \in \mathcal{Q} | (i, j), (r, s) \text{ compared}) \\ &= p \left[\mathbf{P}((i, j, r, s) \in \mathcal{Q} | w_{ij} > w_{rs}; (i, j), (r, s) \text{ compared}) \mathbf{P}(w_{ij} > w_{rs}) \right. \\ &\quad \left. + \mathbf{P}((i, j, r, s) \in \mathcal{Q} | w_{ij} < w_{rs}; (i, j), (r, s) \text{ compared}) \mathbf{P}(w_{ij} < w_{rs}) \right] \\ &= p \left[\frac{(1 + \epsilon)(1 + \delta)}{2} + \frac{(1 - \epsilon)(1 - \delta)}{2} \right] = p \frac{(1 + \epsilon\delta)}{2}, \end{aligned}$$

where in each product, the term $\mathbf{P}(w_{ij} > w_{rs})$ is computed from (3), and the other term, denoting flipped answers, follows from the quadruplet variant of (2). Based on Table 1, we have for i, j such that $\psi_i = \psi_j$

$$\mathbf{E}[S_{ij}] = \sum_{(r, s): \psi_r \neq \psi_s} p\epsilon\delta = p\epsilon\delta \sum_{\ell=1}^k \frac{n_\ell(n - n_\ell)}{2}$$

if $i \neq j$. For $i = j$, obviously $\mathbf{E}[S_{ij}] = 0$. For i, j such that $\psi_i \neq \psi_j$, we have

$$\mathbf{E}[S_{ij}] = \sum_{(r,s):\psi_r=\psi_s} -p\epsilon\delta = -p\epsilon\delta \sum_{\ell=1}^k \binom{n_\ell}{2}.$$

Hence, we define the ideal similarity matrix as $\tilde{S}_{ij} = \mathbf{E}[S_{ij}]$ for $i \neq j$, and $\tilde{S}_{ii} = p\epsilon\delta \sum_{\ell=1}^k \frac{n_\ell(n-n_\ell)}{2}$.

Observe that $\tilde{S} = Z\Sigma Z^T$, where $Z \in \{0, 1\}^{n \times k}$ is the assignment matrix for the planted clusters, and $\Sigma \in \mathbb{R}^{k \times k}$ such that $\Sigma_{\ell\ell} = p\epsilon\delta \sum_{\ell} \frac{n_\ell(n-n_\ell)}{2}$ and $\Sigma_{\ell\ell'} = -p\epsilon\delta \sum_{\ell} \binom{n_\ell}{2}$ for $\ell \neq \ell'$. Hence, in this case, we have

$$\Delta_1 = p\epsilon\delta \binom{n}{2}. \quad (11)$$

Preliminary computations and definitions for concentration. As noted earlier, \tilde{S} and $\mathbf{E}[S]$ are identical, except in the diagonal entries. Hence, we mainly have to obtain concentration of $f(S - \mathbf{E}[S])$, where f is a non-negative scalar function. In the case of Δ_2 , f denotes the maximum partial row sum, whereas f is the spectral norm in the bound for $\|S - \tilde{S}\|_2$. Define $\mathcal{W} = \{w_{ij} : i < j\}$ as the collection of random pairwise similarities. We write

$$S - \mathbf{E}[S] = (S - \mathbf{E}[S|\mathcal{W}]) + (\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]),$$

where the first difference accounts for randomness in sampling and crowd noise, while the second difference accounts for the inherent noise in \mathcal{W} . This helps in separately concentrating both terms, which have different dependence structures. Formally, we perform the concentration of $f(S - \mathbf{E}[S])$ in the following way, assuming f satisfies triangle inequality (which holds in the cases that we later consider).

$$\begin{aligned} \mathbf{P}(f(S - \mathbf{E}[S]) > t) &\leq \mathbf{P}(f(S - \mathbf{E}[S|\mathcal{W}]) + f(\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]) > t) \\ &\leq \mathbf{P}(f(S - \mathbf{E}[S|\mathcal{W}]) > t/2) + \mathbf{P}(f(\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]) > t/2) \\ &\leq \mathbf{E}_{\mathcal{W}} [\mathbf{P}_{\cdot|\mathcal{W}}(f(S - \mathbf{E}[S|\mathcal{W}]) > t/2)] + \mathbf{P}(f(\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]) > t/2), \end{aligned}$$

where $\mathbf{P}_{\cdot|\mathcal{W}}$ denotes the probability over sampling and crowd noise, but conditioned on \mathcal{W} . In fact, we derive an uniform upper bound on the conditional probability, irrespective of \mathcal{W} , and hence the expectation is trivially bounded. To separately deal with the randomness in \mathcal{W} and the randomness due to sampling and crowd noise, we write

$$S_{ij} = \sum_{r \neq s} (\mathbb{I}_{\{(i,j,r,s) \in \mathcal{Q}\}} - \mathbb{I}_{\{(r,s,i,j) \in \mathcal{Q}\}}) = \sum_{r < s} \xi_{ijrs} (\mathbb{I}_{\{w_{ij} > w_{rs}\}} - \mathbb{I}_{\{w_{ij} < w_{rs}\}}) \quad (12)$$

where $\xi_{ijrs} \in \{-1, 0, +1\}$ denotes whether the comparison between (i, j) and (r, s) is observed ($\xi_{ijrs} = 0$ if not observed), and whether the crowd response was correct ($\xi_{ijrs} = +1$) or flipped ($\xi_{ijrs} = -1$). Under our sampling and noise model,

$$\mathbf{P}(\xi_{ijrs} = 0) = 1 - p, \quad \mathbf{P}(\xi_{ijrs} = 1) = \frac{p(1 + \epsilon)}{2}, \quad \mathbf{P}(\xi_{ijrs} = -1) = \frac{p(1 - \epsilon)}{2}$$

and so, $\mathbf{E}[\xi_{ijrs}] = p\epsilon$ and $\text{Var}(\xi_{ijrs}) \leq p$. Note that the set $\Xi = \{\xi_{ijrs} : i < j, r < s, (i, j) < (r, s)\}$ is a collection of independent random variables. Here, $(i, j) < (r, s)$ denotes a lexicographic ordering of tuples since we do not care about the ordering between (i, j) and (r, s) .

In addition, recall that F_{in}, F_{out} are continuous, and hence, with probability 1, any two pairwise similarities are distinct. Hence, we can write $\mathbb{I}_{\{w_{ij} > w_{rs}\}} - \mathbb{I}_{\{w_{ij} < w_{rs}\}} = 2\mathbb{I}_{\{w_{ij} > w_{rs}\}} - 1$. It is noted that ξ_{ijrs} is independent of $(2\mathbb{I}_{\{w_{ij} > w_{rs}\}} - 1)$, and furthermore, the latter variable is deterministic conditioned on \mathcal{W} . Based on this and using the notation of ξ_{ijrs} , we write

$$\begin{aligned} S_{ij} - \mathbf{E}[S_{ij}|\mathcal{W}] &= \sum_{r < s} B_{ijrs}, \quad \text{where} \quad B_{ijrs} = (\xi_{ijrs} - p\epsilon)(2\mathbb{I}_{\{w_{ij} > w_{rs}\}} - 1) \\ \mathbf{E}[S_{ij}|\mathcal{W}] - \mathbf{E}[S_{ij}] &= \sum_{r < s} B'_{ijrs}, \quad \text{where} \quad B'_{ijrs} = 2p\epsilon(\mathbb{I}_{\{w_{ij} > w_{rs}\}} - \mathbf{P}(w_{ij} > w_{rs})). \end{aligned} \quad (13)$$

We make the following observations about the collection of random variables B_{ijrs}, B'_{ijrs} , which are crucial to the subsequent concentration results. It is easy to see that $|B_{ijrs}| \leq 2, |B'_{ijrs}| \leq 2p\epsilon$ with probability 1, and $\mathbf{E}[B_{ijrs}] = \mathbf{E}[B'_{ijrs}] = 0, \text{Var}(B_{ijrs}) \leq p$ and $\text{Var}(B'_{ijrs}) \leq 4p^2\epsilon^2$. Define the sets

$$\begin{aligned} \mathcal{B} &= \{B_{ijrs} : i < j, r < s, (i, j) \neq (r, s)\}, \\ \mathcal{B}' &= \{B'_{ijrs} : i < j, r < s, (i, j) \neq (r, s)\}, \\ \mathcal{B}_{i\ell} &= \{B_{ijrs} : j \in \mathcal{C}_\ell, j \neq i, r < s, (i, j) \neq (r, s)\} \quad \text{for every } i \in [n], \ell \in [k], \\ \text{and } \mathcal{B}'_{i\ell} &= \{B'_{ijrs} : j \in \mathcal{C}_\ell, j \neq i, r < s, (i, j) \neq (r, s)\} \quad \text{for every } i \in [n], \ell \in [k]. \end{aligned} \quad (14)$$

Each of \mathcal{B} and \mathcal{B}' have $\binom{n}{2} \left(\binom{n}{2} - 1 \right)$ random variables. $B_{ijrs} = -B_{rsij}$, but conditioned on \mathcal{W} , B_{ijrs} is independent of all other variables in \mathcal{B} . Thus, a dependency graph on \mathcal{B} , conditioned on \mathcal{W} , has a maximum degree of 1. On the other hand, B'_{ijrs} depends on all the random variables of the form $B'_{ijr's'}, B'_{i'j'r_s}, B'_{r's'ij}$ and $B'_{rsi'j'}$, and so, the dependence graph for \mathcal{B}' has degree smaller than $4\binom{n}{2} - 7$. Similarly, $\mathcal{B}_{i\ell}, \mathcal{B}'_{i\ell}$ have at most $n_\ell \left(\binom{n}{2} - 1 \right)$ random variables. While $\mathcal{B}_{i\ell}$ has a dependency graph with degree at most 1, the dependency graph of $\mathcal{B}'_{i\ell}$ has degree at most $n_\ell + \binom{n}{2} - 3$. We now use the above discussion to derive upper bounds on Δ_2 and $\|S - \tilde{S}\|_2$.

Upper bound for Δ_2 . To derive a bound on Δ_2 , we first note that

$$\Delta_2 \leq \max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij}] \right| + \frac{a_0}{n_{\min}}$$

where $a_0 = \tilde{S}_{ii}$ takes into account the fact that \tilde{S} and $\mathbf{E}[S]$ differ only in diagonal terms. In the subsequent steps, we bound the first term. For any $t > 0$, the union bound leads to

$$\begin{aligned} & \mathbf{P} \left(\max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij}] \right| > t \right) \\ & \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij}|\mathcal{W}] + \mathbf{E}[S_{ij}|\mathcal{W}] - \mathbf{E}[S_{ij}] \right| > n_\ell t \right) \\ & \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{E}_{\mathcal{W}} \left[\mathbf{P}_{\cdot|\mathcal{W}} \left(\left| \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij}|\mathcal{W}] \right| > \frac{n_\ell t}{2} \right) \right] + \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} \mathbf{E}[S_{ij}|\mathcal{W}] - \mathbf{E}[S_{ij}] \right| > n_\ell t \right) \\ & \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{E}_{\mathcal{W}} \left[\mathbf{P}_{\cdot|\mathcal{W}} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r < s} B_{ijrs} \right| > \frac{n_\ell t}{2} \right) \right] + \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r < s} B'_{ijrs} \right| > \frac{n_\ell t}{2} \right) \end{aligned} \quad (15)$$

For the probability conditioned on \mathcal{W} , the summation involves terms in the set $\mathcal{B}_{i\ell}$ in (14). The discussion on $\mathcal{B}_{i\ell}$ shows that, conditioned on \mathcal{W} , the summation is a sum of independent random variables B_{ijrs} whose properties are stated after (13). Hence, we can apply Bernstein's inequality to bound the conditional probability as

$$\begin{aligned} \mathbf{P}_{\cdot|\mathcal{W}} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r < s} B_{ijrs} \right| > \frac{n_\ell t}{2} \right) & \leq 2 \exp \left(- \frac{\left(\frac{n_\ell t}{2} \right)^2}{2pn_\ell \binom{n}{2} + \frac{2}{3} 2 \frac{n_\ell t}{2}} \right) \\ & \lesssim 2 \exp \left(- \min \left\{ \frac{n_\ell t^2}{pn^2}, n_\ell t \right\} \right) \lesssim \frac{1}{n^3} \end{aligned}$$

for $t \gtrsim \max \left\{ \sqrt{\frac{pn^2 \ln n}{n_{\min}}}, \frac{\ln n}{n_{\min}} \right\}$. Since the $\mathcal{O} \left(\frac{1}{n^3} \right)$ bound on the probability holds uniformly for all \mathcal{W} , it also bounds the first term in (15).

For the second probability in (15), note that the B'_{ijrs} in the summation are not independent, and we cannot directly apply Bernstein inequality. Hence, we apply the technique in Janson and Ruciński

(2002, Theorem 5) which bounds the probability by partitioning the random variables in $\mathcal{B}'_{i\ell}$ into independent sets. Since the dependency graph on $\mathcal{B}'_{i\ell}$ has maximum degree $d = n_\ell + \binom{n}{2} - 3$, we can obtain an equitable $(d+1)$ -colouring, with each independent set of size $\lfloor |\mathcal{B}'_{i\ell}|/(d+1) \rfloor$ or $\lceil |\mathcal{B}'_{i\ell}|/(d+1) \rceil$, which are both smaller than n_ℓ . Denote the independent sets by $\mathcal{B}'_{i\ell,(1)}, \dots, \mathcal{B}'_{i\ell,(d+1)}$, and we can apply Bernstein's inequality to bound the summation over each independent set. Hence, we bound the second probability in (15), for every i, ℓ , as

$$\begin{aligned}
& \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r < s} B'_{ijrs} \right| > \frac{n_\ell t}{2} \right) \\
& \leq \mathbf{P} \left(\max_{r \in \{1, \dots, d+1\}} \left| \sum_{B' \in \mathcal{B}'_{i\ell,(r)}} B' \right| > \frac{n_\ell t}{2(d+1)} \right) \\
& \leq \sum_{r=1}^{d+1} \mathbf{P} \left(\left| \sum_{B' \in \mathcal{B}'_{i\ell,(r)}} B' \right| > \frac{n_\ell t}{2(d+1)} \right) \quad (\text{union bound}) \\
& \leq 2(d+1) \exp \left(- \frac{\left(\frac{n_\ell t}{2(d+1)} \right)^2}{2 \sum_{B' \in \mathcal{B}'_{i\ell,(r)}} \text{Var}(B') + \frac{2}{3} 2p\epsilon \frac{n_\ell t}{2(d+1)}} \right) \quad (\text{Bernstein bound}) \\
& \leq 2(d+1) \exp \left(- \frac{\left(\frac{n_\ell t}{d+1} \right)^2}{8p^2\epsilon^2 n_\ell + \frac{2}{3} p\epsilon \frac{n_\ell t}{d+1}} \right) \\
& \lesssim n^2 \exp \left(- \min \left\{ \frac{n_\ell t^2}{p^2 \epsilon^2 n^4}, \frac{n_\ell t}{p\epsilon n^2} \right\} \right),
\end{aligned}$$

which is $\mathcal{O}(\frac{1}{n^3})$ for $t \gtrsim p\epsilon n^2 \cdot \max \left\{ \sqrt{\frac{\ln n}{n_{\min}}}, \frac{\ln n}{n_{\min}} \right\}$. The first term dominates since, under the condition on δ , we have $n_{\min} \gtrsim \ln n$. Thus, we conclude that, with probability $1 - \frac{1}{4n}$,

$$\begin{aligned}
\Delta_2 & \leq \max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij}] \right| + \frac{a_0}{n_{\min}} \\
& \lesssim \max \left\{ \sqrt{\frac{pn^2 \ln n}{n_{\min}}}, \frac{\ln n}{n_{\min}}, p\epsilon n^2 \sqrt{\frac{\ln n}{n_{\min}}}, \frac{p\epsilon \delta n^2}{n_{\min}} \right\}, \quad (16)
\end{aligned}$$

where the last term is obviously dominated by the third term.

Upper bound for $\|S - \tilde{S}\|_2$. Similar to the case of Δ_2 , we bound the spectral norm as

$$\|S - \tilde{S}\|_2 \leq \|S - \mathbf{E}[S|\mathcal{W}]\|_2 + \|\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]\|_2 + \|\mathbf{E}[S] - \tilde{S}\|_2,$$

where the last term equals $a_0 = \tilde{S}_{ii}$ since $\mathbf{E}[S] - \tilde{S}$ is a diagonal matrix. For the first term, we derive a bound conditioned on \mathcal{W} . Recall from (13)–(14) that, conditioned on \mathcal{W} , the matrix $S - \mathbf{E}[S|\mathcal{W}]$ comprises of variables in \mathcal{B} , which has a dependence graph with degree 1. We partition \mathcal{B} into two independent sets via equitable colouring, and write $S - \mathbf{E}[S|\mathcal{W}] = A + A'$, where A and A' are the symmetric matrices corresponding to each of the independent sets. We derive a spectral norm for each of A and A' . For this, we first claim that, conditioned on \mathcal{W} , the event $\mathcal{E} = \left\{ \max_{i,j} \{|A_{ij}|, |A'_{ij}|\} \lesssim \max \left\{ \sqrt{pn^2 \ln n}, \ln n \right\} \right\}$ occurs with probability $1 - \mathcal{O}(\frac{1}{n})$. To see this, observe that A_{ij} (or A'_{ij}) is a sum of at most $\binom{n}{2}$ independent random variables B_{ijrs} . By Bernstein inequality,

$$\mathbf{P}_{\cdot|\mathcal{W}}(|A_{ij}| > \tau) \leq 2 \exp \left(- \frac{\tau^2}{2p \binom{n}{2} + \frac{4}{3}\tau} \right) \lesssim \exp \left(- \min \left\{ \frac{\tau^2}{pn^2}, \tau \right\} \right)$$

which is $\mathcal{O}\left(\frac{1}{n^3}\right)$ for $\tau \gtrsim \max\left\{\sqrt{pn^2 \ln n}, \ln n\right\}$. Applying the union bound gives $\mathbf{P}(\mathcal{E}^c) = \mathcal{O}\left(\frac{1}{n}\right)$.

Conditioned on \mathcal{W} and \mathcal{E} , the matrices A, A' have independent zero mean entries, with each entry bounded by $\mathcal{O}\left(\max\left\{\sqrt{pn^2 \ln n}, \ln n\right\}\right)$. Furthermore, from the variance of B_{ijrs} , we have $\max_i \sum_j \text{Var}(A_{ij}) < pn^3$, and same for A' . Hence, by matrix Bernstein inequality (Tropp, 2012),

$$\begin{aligned} \mathbf{P}_{\cdot|\mathcal{W},\mathcal{E}}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t) &\leq \mathbf{P}_{\cdot|\mathcal{W},\mathcal{E}}(\|A\|_2 > t/2) + \mathbf{P}_{\cdot|\mathcal{W},\mathcal{E}}(\|A'\|_2 > t/2) \\ &\leq 2n \exp\left(-\frac{t^2/4}{pn^3 + \frac{1}{3}t \cdot \max\left\{\sqrt{pn^2 \ln n}, \ln n\right\}}\right) \\ &\lesssim n \exp\left(-\min\left\{\frac{t^2}{pn^3}, \frac{t}{\sqrt{pn^2 \ln n}}, \frac{t}{\ln n}\right\}\right) \lesssim \frac{1}{n} \end{aligned}$$

for $t \gtrsim \left\{\sqrt{pn^3 \ln n}, \sqrt{pn^2(\ln n)^3}, (\ln n)^2\right\}$, where the second term is smaller than the first for n large enough. Denote the complement of \mathcal{E} by \mathcal{E}^c . For t satisfying the stated condition,

$$\begin{aligned} &\mathbf{P}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t) \\ &= \mathbf{E}_{\mathcal{W}}[\mathbf{P}_{\cdot|\mathcal{W}}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t)] \\ &= \mathbf{E}_{\mathcal{W}}[\mathbf{P}_{\cdot|\mathcal{W},\mathcal{E}}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t) \mathbf{P}_{\cdot|\mathcal{W}}(\mathcal{E}) + \mathbf{P}_{\cdot|\mathcal{W},\mathcal{E}^c}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t) \mathbf{P}_{\cdot|\mathcal{W}}(\mathcal{E}^c)] \\ &\lesssim \mathbf{E}_{\mathcal{W}}[\mathbf{P}_{\cdot|\mathcal{W},\mathcal{E}}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t) + \mathbf{P}_{\cdot|\mathcal{W}}(\mathcal{E}^c)] \\ &\lesssim \frac{1}{n} \end{aligned}$$

as each term in the expectation is $\mathcal{O}\left(\frac{1}{n}\right)$. Thus, we have $\|S - \mathbf{E}[S|\mathcal{W}]\|_2 \lesssim \left\{\sqrt{pn^3 \ln n}, (\ln n)^2\right\}$.

To bound $\|\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]\|_2$, we note that the entries of the matrix comprises of mutually dependent variables in the set \mathcal{B}' defined in (14). We need to partition the entries into independent sets. Since the dependency graph for \mathcal{B}' has maximum degree $d = 4\binom{n}{2} - 1$, we can partition \mathcal{B}' into $d + 1$ independent sets of nearly identical sizes (equitable colouring). Let $\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S] = A^{(1)} + \dots + A^{(d+1)}$ denote the corresponding partition of the matrix, where $A^{(\ell)} \in \mathbb{R}^{n \times n}$ is a symmetric matrix, consisting of the variables in the ℓ -th independent set. Due to independence of variables, we have $A_{ij}^{(\ell)} = B'_{ijrs}$ for some r, s , and hence, we can conclude that each $A^{(\ell)}$ is a symmetric matrix with independent zero-mean entries, bounded by $2p\epsilon$ and variance at most $4p^2\epsilon^2$ (follows from properties of B'_{ijrs}). Thus, by matrix Bernstein inequality (Tropp, 2012), we have

$$\mathbf{P}\left(\|A^{(\ell)}\|_2 > \tau\right) \leq n \exp\left(-\frac{t^2}{8p^2\epsilon^2n + \frac{2}{3}p\epsilon t}\right),$$

and combining with the union bound,

$$\begin{aligned} \mathbf{P}(\|\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]\|_2 > t) &\leq \mathbf{P}\left(\max_{\ell \in [d+1]} \|A^{(\ell)}\|_2 > \frac{t}{d+1}\right) \\ &\leq n(d+1) \exp\left(-\frac{\left(\frac{t}{d+1}\right)^2}{8p^2\epsilon^2n + \frac{2}{3}p\epsilon \frac{t}{d+1}}\right) \\ &\lesssim n^3 \exp\left(-\min\left\{\frac{t^2}{p^2\epsilon^2n^5}, \frac{t}{pen^2}\right\}\right), \end{aligned}$$

which is $\mathcal{O}\left(\frac{1}{n}\right)$ for $t \gtrsim pen^2 \cdot \max\left\{\sqrt{n \ln n}, \ln n\right\}$, where the first term obviously dominates.

Combining the above derivations, we have with probability $1 - \frac{1}{4n}$,

$$\|S - \tilde{S}\|_2 \lesssim \max\left\{\sqrt{pn^3 \ln n}, (\ln n)^2, pen^2\sqrt{n \ln n}, p\epsilon\delta n^2\right\} \quad (17)$$

where the last term (arising due to a_0) is dominated by the third.

Deriving interval for λ in terms of $|\mathcal{Q}|$. We now use (11), (16) and (17) to complete the proof for the quadruplet setting. To this end, our main objective is to verify the conditions in Proposition 1:

$$\frac{\Delta_1}{2} < \Delta_1 - 6\Delta_2, \text{ that is, } \Delta_2 < \frac{\Delta_1}{12} \quad \text{and} \quad \|S - \tilde{S}\|_2 < \frac{n_{\min}\Delta_1}{2}.$$

Using (11) and the bound in (16), we observe that $\Delta_2 \lesssim \Delta_1$ if

$$p \gtrsim \max \left\{ \frac{\ln n}{\epsilon^2 \delta^2 n^2 n_{\min}}, \frac{\ln n}{\epsilon \delta n^2 n_{\min}} \right\} \quad \text{and} \quad \delta \gtrsim \sqrt{\frac{\ln n}{n_{\min}}},$$

where the condition on δ arises due to the third term in the bound in (16). Similarly, comparing the bound in (17) to (11), we get that $\|S - \tilde{S}\|_2 \lesssim n_{\min}\Delta_1$ if

$$p \gtrsim \max \left\{ \frac{\ln n}{\epsilon^2 \delta^2 n n_{\min}^2}, \frac{(\ln n)^2}{\epsilon \delta n^2 n_{\min}} \right\} \quad \text{and} \quad \delta \gtrsim \frac{\sqrt{n \ln n}}{n_{\min}},$$

where the condition on δ arises from the third bound in (17). Combining the above cases, we conclude that if

$$\delta \gtrsim \frac{\sqrt{n \ln n}}{n_{\min}} \quad \text{and} \quad p \gtrsim \frac{(\ln n)^2}{\epsilon^2 \delta^2 n n_{\min}^2}, \quad (18)$$

then the criteria for Δ_2 and $\|S - \tilde{S}\|_2$ are satisfied, and by Proposition 1, X^* is the unique optimal solution for SDP- λ with the range of λ given by

$$\|S - \tilde{S}\|_2 \lesssim \max \left\{ \sqrt{pn^3 \ln n}, p\epsilon \sqrt{n^5 \ln n}, (\ln n)^2 \right\} \lesssim \lambda < \frac{p\epsilon \delta n_{\min}}{2} \binom{n}{2} = \frac{\Delta_1}{2}. \quad (19)$$

We finally show that the condition on p holds under the stated condition of $|\mathcal{Q}| \gtrsim \frac{n^3(\ln n)^2}{\epsilon^2 \delta^2 n_{\min}^2}$, and state the above interval for λ in terms of $|\mathcal{Q}|$. Under the assumption that each quadruplet is observed independently with probability p , we have that $\mathbf{E}[|\mathcal{Q}|] = p \binom{n}{2} = O(pn^4)$. By Bernstein inequality, it is easy to verify that for $p \gtrsim \frac{\ln n}{n^4}$ or equivalently $|\mathcal{Q}| \gtrsim \ln n$, we have $|\mathcal{Q}| \in (\frac{1}{2}\mathbf{E}[|\mathcal{Q}|], \frac{3}{2}\mathbf{E}[|\mathcal{Q}|])$ with probability $1 - \mathcal{O}(\frac{1}{n})$. Hence, we can replace p by $\frac{|\mathcal{Q}|}{n^4}$ in (18)–(19) up to difference in constants, which leads to the statement of Theorem 1 in the quadruplet setting.

C.2 Triplet setting

The proof structure in the triplet case is similar to that of the quadruplet setting. We derive an appropriate ideal matrix \tilde{S} , where $\tilde{S} = \mathbf{E}[S]$, except for some differences in the diagonal entries since $S_{ii} = 0$ for all i . From the block structure of \tilde{S} , we can compute Δ_1 . Subsequently, concentration inequalities are used to derive upper bounds on $\|S - \tilde{S}\|_2$ and Δ_2 in terms of the model parameters. As done in the analysis of AddS-4, we let \mathcal{W} denote the collection of random pairwise similarities, and decompose

$$S - \tilde{S} = (S - \mathbf{E}[S|\mathcal{W}]) + (\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]) + (\mathbf{E}[S] - \tilde{S}).$$

The last term is easy to tackle, and we use separate concentration for the first two terms both in the context of Δ_2 and the spectral norm. Bounds on these terms, combined with Proposition 1, provide sufficient conditions on δ and sampling rate p such that exact recovery occurs. Finally, we show that for p large enough, the number of triplets $|\mathcal{T}|$ is close to its expected value $pn \binom{n-1}{2}$, and state the conditions in terms of $|\mathcal{T}|$.

Computation of Δ_1 . The expectation of the AddS-3 similarity S_{ij} , for $i \neq j$, is given by

$$\mathbf{E}[S_{ij}] = \sum_{r \neq i, j} \mathbf{P}((i, j, r) \in \mathcal{T}) - \mathbf{P}((i, r, j) \in \mathcal{T}) + \mathbf{P}((j, i, r) \in \mathcal{T}) - \mathbf{P}((j, r, i) \in \mathcal{T}). \quad (20)$$

We now compute each term in the summation using the notation $\psi_i \in [k]$ to indicate $i \in \mathcal{C}_{\psi_i}$. The expected values of the terms are given in Table 2, where the last column represents the overall term

Table 2: Value of each term in the summation in (20), assuming i, j, r are distinct.

Case	$\mathbf{P}((i, j, r) \in \mathcal{T})$	$\mathbf{P}((i, r, j) \in \mathcal{T})$	$\mathbf{P}((j, i, r) \in \mathcal{T})$	$\mathbf{P}((j, r, i) \in \mathcal{T})$	Aggregate
$\psi_i = \psi_j = \psi_r$	$p/2$	$p/2$	$p/2$	$p/2$	0
$\psi_i = \psi_j \neq \psi_r$	$p(1 + \epsilon\delta)/2$	$p(1 - \epsilon\delta)/2$	$p(1 + \epsilon\delta)/2$	$p(1 - \epsilon\delta)/2$	$2p\epsilon\delta$
$\psi_i \neq \psi_j = \psi_r$	$p/2$	$p/2$	$p(1 - \epsilon\delta)/2$	$p(1 + \epsilon\delta)/2$	$-p\epsilon\delta$
$\psi_i = \psi_r \neq \psi_j$	$p(1 - \epsilon\delta)/2$	$p(1 + \epsilon\delta)/2$	$p/2$	$p/2$	$-p\epsilon\delta$
$\psi_i \neq \psi_j \neq \psi_r$	$p/2$	$p/2$	$p/2$	$p/2$	0

for each $r \neq i, j$ in (20). The derivation for these values is identical to the one in the quadruplet setting.

Based on Table 2, we can infer that for $i \neq j$ such that $\psi_i = \psi_j$,

$$\mathbf{E}[S_{ij}] = \sum_{r \in \mathcal{C}_{\psi_i}} 2p\epsilon\delta = 2p\epsilon\delta(n - n_{\psi_i})$$

For i, j such that $\psi_i \neq \psi_j$, we have

$$\mathbf{E}[S_{ij}] = \sum_{r \in \mathcal{C}_{\psi_i}, r \neq i} (-p\epsilon\delta) + \sum_{r \in \mathcal{C}_{\psi_j}, r \neq j} (-p\epsilon\delta) = -p\epsilon\delta(n_{\psi_i} + n_{\psi_j} - 2).$$

Hence, we define the ideal similarity matrix as $\tilde{S}_{ij} = Z\Sigma Z^T$, where $\Sigma_{\ell\ell} = 2p\epsilon\delta(n - n_\ell)$ and $\Sigma_{\ell\ell'} = -p\epsilon\delta(n_\ell + n_{\ell'} - 2)$ for $\ell \neq \ell'$, and we can compute

$$\Delta_1 = p\epsilon\delta(n - 2). \quad (21)$$

Preliminary computations and definitions for concentration. We define $\mathcal{W} = \{w_{ij} : i < j\}$ as the collection of random pairwise similarities, and split the concentration of Δ_2 and $\|S - \tilde{S}\|_2$ into terms involving $S - \mathbf{E}[S|\mathcal{W}]$ and $\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]$. The basic idea is discussed in the corresponding part of the quadruplet setting, and here, we introduce the key random variables. We first write

$$\begin{aligned} S_{ij} &= \sum_{r \neq i, j} (\mathbb{I}_{\{(i, j, r) \in \mathcal{T}\}} - \mathbb{I}_{\{(i, r, j) \in \mathcal{T}\}}) + (\mathbb{I}_{\{(j, i, r) \in \mathcal{T}\}} - \mathbb{I}_{\{(j, r, i) \in \mathcal{T}\}}) \\ &= \sum_{r \neq i, j} \xi_{ijr} (\mathbb{I}_{\{w_{ij} > w_{ir}\}} - \mathbb{I}_{\{w_{ij} < w_{ir}\}}) + \xi_{jir} (\mathbb{I}_{\{w_{ji} > w_{jr}\}} - \mathbb{I}_{\{w_{ji} < w_{jr}\}}) \end{aligned} \quad (22)$$

where $\xi_{ijr} \in \{-1, 0, +1\}$ denotes whether the comparison between (i, j) and (i, r) is observed ($\xi_{ijr} = 0$ if not observed), and whether the crowd response was correct ($\xi_{ijr} = +1$) or flipped ($\xi_{ijr} = -1$). Under our sampling and noise model,

$$\mathbf{P}(\xi_{ijr} = 0) = 1 - p, \quad \mathbf{P}(\xi_{ijr} = 1) = \frac{p(1 + \epsilon)}{2}, \quad \mathbf{P}(\xi_{ijr} = -1) = \frac{p(1 - \epsilon)}{2}$$

and so, $\mathbf{E}[\xi_{ijr}] = p\epsilon$ and $\text{Var}(\xi_{ijr}) \leq p$. The set $\Xi = \{\xi_{ijr} : j < r, i \neq j, r\}$ denotes the collection of such random variables, where we abuse notation by using ξ_{ijr} and ξ_{irj} to refer to the same variable. We note that the variables in Ξ are mutually independent.

We use the continuous nature of F_{in}, F_{out} to write $\mathbb{I}_{\{w_{ij} > w_{ir}\}} - \mathbb{I}_{\{w_{ij} < w_{ir}\}} = 2\mathbb{I}_{\{w_{ij} > w_{ir}\}} - 1$, and further define

$$\begin{aligned} S_{ij} - \mathbf{E}[S_{ij}|\mathcal{W}] &= \sum_{r \neq i, j} B_{ijr} + B_{jir} \quad \text{with} \quad B_{ijr} = (\xi_{ijr} - p\epsilon)(2\mathbb{I}_{\{w_{ij} > w_{ir}\}} - 1), \\ \mathbf{E}[S_{ij}|\mathcal{W}] - \mathbf{E}[S_{ij}] &= \sum_{r \neq i, j} B'_{ijr} + B'_{jir} \quad \text{with} \quad B'_{ijr} = 2p\epsilon(\mathbb{I}_{\{w_{ij} > w_{ir}\}} - \mathbf{P}(w_{ij} > w_{ir})). \end{aligned} \quad (23)$$

The random variables B_{ijr}, B'_{ijr} have the following properties: $|B_{ijr}| \leq 2$, $|B'_{ijr}| \leq 2p\epsilon$ with probability 1, $\mathbf{E}[B_{ijr}] = \mathbf{E}[B'_{ijr}] = 0$, $\text{Var}(B_{ijr}) \leq p$ and $\text{Var}(B'_{ijr}) \leq 4p^2\epsilon^2$. We define the sets

$$\begin{aligned} \mathcal{B} &= \{B_{ijr} : j \neq i, r \neq i, j\}, \\ \mathcal{B}' &= \{B'_{ijr} : j \neq i, r \neq i, j\}, \\ \mathcal{B}_{i\ell} &= \{B_{ijr}, B_{jir} : j \in \mathcal{C}_\ell, r \neq i, j\} \quad \text{for every } i \in [n], \ell \in [k], \\ \text{and } \mathcal{B}'_{i\ell} &= \{B'_{ijr}, B'_{jir} : j \in \mathcal{C}_\ell, r \neq i, j\} \quad \text{for every } i \in [n], \ell \in [k]. \end{aligned} \quad (24)$$

Each of \mathcal{B} and \mathcal{B}' have $n(n-1)(n-2)$ random variables, whereas $\mathcal{B}_{i\ell}, \mathcal{B}'_{i\ell}$ have at most $2n_\ell(n-2)$ random variables. Note that $B_{ijr} = -B_{irj}$, but conditioned on \mathcal{W} , B_{ijr} is independent of all other variables in \mathcal{B} . Thus, a dependency graph on \mathcal{B} , conditioned on \mathcal{W} , has a maximum degree of 1. The same is also true for $\mathcal{B}'_{i\ell}$. On the other hand, B'_{ijr} depends on the random variables that involve either w_{ij} or w_{ir} , that is $B'_{irj}, B'_{jir}, B'_{jri}, B'_{rij}, B'_{rji}$, as well as all six variants of variables $B'_{ijr'}$ and $B'_{ij'r}, j', r' \notin \{i, j, r\}$. Thus each B'_{ijr} depends on at most $5 + 12(n-3) = \mathcal{O}(n)$ variables in \mathcal{B}' . The same holds when we restrict the set to $\mathcal{B}'_{i\ell}$. Thus, the dependency graph for \mathcal{B}' and $\mathcal{B}'_{i\ell}$ have dependency graph with $\mathcal{O}(n)$ maximum degree. We now use the above defined random variables and their properties to derive upper bounds on Δ_2 and $\|S - \tilde{S}\|_2$.

Upper bound for Δ_2 . To derive a bound on Δ_2 , we first note that

$$\Delta_2 \leq \max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij} | \mathcal{W}] \right| + \max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} \mathbf{E}[S_{ij} | \mathcal{W}] - \mathbf{E}[S_{ij}] \right| + \max_{i \in [n]} \frac{\tilde{S}_{ii}}{n_{\min}}.$$

In the subsequent steps, we bound the first term. For any $t > 0$, the union bound leads to

$$\begin{aligned} & \mathbf{P} \left(\max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij} | \mathcal{W}] \right| > t \right) \\ & \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij} | \mathcal{W}] \right| > n_\ell t \right) \\ & = \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{E}_{\mathcal{W}} \left[\mathbf{P}_{\cdot | \mathcal{W}} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r \neq i, j} B_{ijr} + B_{jir} \right| > n_\ell t \right) \right] \end{aligned}$$

The summation inside the conditional probability involves terms in $\mathcal{B}_{i\ell}$ defined in (24), and the previous discussions show that the dependency graph of $\mathcal{B}_{i\ell}$ has maximum degree of 1. Hence, we can split the $2n_\ell(n-2)$ variables in $\mathcal{B}_{i\ell}$ into two independent sets, say $\mathcal{B}_{i\ell, (1)}$ and $\mathcal{B}_{i\ell, (2)}$, and derive concentration for each of them separately using Bernstein inequality in the following way.

$$\begin{aligned} & \mathbf{P} \left(\max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} S_{ij} - \mathbf{E}[S_{ij} | \mathcal{W}] \right| > t \right) \\ & \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{E}_{\mathcal{W}} \left[\mathbf{P}_{\cdot | \mathcal{W}} \left(\left| \sum_{B \in \mathcal{B}_{i\ell, (1)}} B \right| > \frac{n_\ell t}{2} \right) + \mathbf{P}_{\cdot | \mathcal{W}} \left(\left| \sum_{B \in \mathcal{B}_{i\ell, (2)}} B \right| > \frac{n_\ell t}{2} \right) \right] \\ & \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{E}_{\mathcal{W}} \left[2 \exp \left(-\frac{(n_\ell t/2)^2}{2p|\mathcal{B}_{i\ell, (1)}| + 2n_\ell t/3} \right) + 2 \exp \left(-\frac{(n_\ell t/2)^2}{2p|\mathcal{B}_{i\ell, (2)}| + 2n_\ell t/3} \right) \right] \\ & \lesssim n^2 \exp \left(-\min \left\{ \frac{n_{\min} t^2}{pn}, n_{\min} t \right\} \right), \end{aligned}$$

where the last step follows by noting that each of the two independent sets have $\Omega(nn_\ell)$ variables, and the bounds are independent of \mathcal{W} . The above probability is $\mathcal{O}(\frac{1}{n})$ for $t \gtrsim \max \left\{ \sqrt{\frac{pn \ln n}{n_{\min}}}, \frac{\ln n}{n_{\min}} \right\}$.

For the second term in the upper bound for Δ_2 , we have

$$\mathbf{P} \left(\max_{i \in [n]} \max_{\ell \in [k]} \left| \frac{1}{n_\ell} \sum_{j \in \mathcal{C}_\ell} \mathbf{E}[S_{ij} | \mathcal{W}] - \mathbf{E}[S_{ij}] \right| > t \right) \leq \sum_{i \in [n]} \sum_{\ell \in [k]} \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r \neq i, j} B'_{ijr} + B'_{jir} \right| > n_\ell t \right)$$

where the tail bound is for the sum of all random variables in $\mathcal{B}'_{i\ell}$. Since the dependency graph on $\mathcal{B}'_{i\ell}$ has maximum degree $d = \mathcal{O}(n)$, we can obtain an equitable $(d+1)$ -colouring with each independent set of size $\lfloor |\mathcal{B}'_{i\ell}|/(d+1) \rfloor$ or $\lceil |\mathcal{B}'_{i\ell}|/(d+1) \rceil$, which are smaller than n_ℓ . We denote the

independent sets by $\mathcal{B}'_{i\ell,(1)}, \dots, \mathcal{B}'_{i\ell,(d+1)}$, and use Bernstein inequality to bound the summation over each independent set. Hence, we bound the probability for every i, ℓ , as

$$\begin{aligned} \mathbf{P} \left(\left| \sum_{j \in \mathcal{C}_\ell} \sum_{r \neq i, j} B'_{ijr} + B'_{jir} \right| > n_\ell t \right) &\leq \mathbf{P} \left(\max_{r \in \{1, \dots, d+1\}} \left| \sum_{B' \in \mathcal{B}'_{i\ell,(r)}} B' \right| > \frac{n_\ell t}{(d+1)} \right) \\ &\leq \sum_{r=1}^{d+1} \mathbf{P} \left(\left| \sum_{B' \in \mathcal{B}'_{i\ell,(r)}} B' \right| > \frac{n_\ell t}{(d+1)} \right) \\ &\leq 2(d+1) \exp \left(- \frac{\left(\frac{n_\ell t}{(d+1)} \right)^2}{8p^2 \epsilon^2 |\mathcal{B}'_{i\ell,(r)}| + \frac{4}{3} p \epsilon \frac{n_\ell t}{(d+1)}} \right) \\ &\lesssim n \exp \left(- \min \left\{ \frac{n_\ell t^2}{p^2 \epsilon^2 n^2}, \frac{n_\ell t}{p \epsilon n} \right\} \right), \end{aligned}$$

which is $\mathcal{O}(\frac{1}{n^3})$ for $t \gtrsim p \epsilon n \cdot \max \left\{ \sqrt{\frac{\ln n}{n_{\min}}}, \frac{\ln n}{n_{\min}} \right\}$. The first term dominates for $n_{\min} \gtrsim \ln n$, which arises due to the condition on δ . Combining the above discussions we claim that, with probability $1 - \frac{1}{4n}$,

$$\Delta_2 \lesssim \max \left\{ \sqrt{\frac{pn \ln n}{n_{\min}}}, \frac{\ln n}{n_{\min}}, p \epsilon n \sqrt{\frac{\ln n}{n_{\min}}}, \frac{p \epsilon \delta n}{n_{\min}} \right\}, \quad (25)$$

where the last term is obviously dominated by the third term.

Upper bound for $\|S - \tilde{S}\|_2$. Similar to the case of Δ_2 , we bound

$$\|S - \tilde{S}\|_2 \leq \|S - \mathbf{E}[S|\mathcal{W}]\|_2 + \|\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]\|_2 + \|\mathbf{E}[S] - \tilde{S}\|_2,$$

where the last term equals $\max_i \tilde{S}_{ii}$. For the first term, we derive a bound conditioned on \mathcal{W} . Recall from (23)–(24) that, conditioned on \mathcal{W} , the matrix $S - \mathbf{E}[S|\mathcal{W}]$ comprises of variables in \mathcal{B} , which has a dependence graph with degree 1. We partition \mathcal{B} into two independent sets via equitable colouring, and write $S - \mathbf{E}[S|\mathcal{W}] = A + A'$, where A and A' are the matrices corresponding to each of the independent sets. We derive a spectral norm for each of A and A' . For this, we first claim that, conditioned on \mathcal{W} , the event $\mathcal{E} = \{\max_{i,j} \{|A_{ij}|, |A'_{ij}|\} \lesssim \max\{\sqrt{pn \ln n}, \ln n\}\}$ occurs with probability $1 - \mathcal{O}(\frac{1}{n})$. To see this, observe that A_{ij} (or A'_{ij}) is a sum of $2(n-2)$ independent random variables B_{ijr}, B_{jir} . By Bernstein inequality,

$$\mathbf{P}_{\cdot|\mathcal{W}}(|A_{ij}| > \tau) \leq 2 \exp \left(- \frac{\tau^2}{4p(n-2) + \frac{4}{3}\tau} \right) \lesssim \exp \left(- \min \left\{ \frac{\tau^2}{pn}, \tau \right\} \right)$$

which is $\mathcal{O}(\frac{1}{n^3})$ for $\tau \gtrsim \max\{\sqrt{pn \ln n}, \ln n\}$. Applying the union bound gives $\mathbf{P}(\mathcal{E}^c) = \mathcal{O}(\frac{1}{n})$.

Conditioned on \mathcal{W} and \mathcal{E} , the matrices A, A' have independent zero mean entries, with each entry bounded by $\mathcal{O}(\max\{\sqrt{pn \ln n}, \ln n\})$. Furthermore, from the variance of B_{ijr} , we have $\max_i \sum_j \text{Var}(A_{ij}) < 2pn^2$, and the same holds for A' . Hence, by matrix Bernstein inequality (Tropp, 2012),

$$\begin{aligned} \mathbf{P}_{\cdot|\mathcal{W}, \mathcal{E}}(\|S - \mathbf{E}[S|\mathcal{W}]\|_2 > t) &\leq \mathbf{P}_{\cdot|\mathcal{W}, \mathcal{E}}(\|A\|_2 > t/2) + \mathbf{P}_{\cdot|\mathcal{W}, \mathcal{E}}(\|A'\|_2 > t/2) \\ &\leq 2n \exp \left(- \frac{t^2/4}{pn^2 + \frac{1}{3}t \cdot \max\{\sqrt{pn \ln n}, \ln n\}} \right) \\ &\lesssim n \exp \left(- \min \left\{ \frac{t^2}{pn^2}, \frac{t}{\sqrt{pn \ln n}}, \frac{t}{\ln n} \right\} \right) \lesssim \frac{1}{n} \end{aligned}$$

for $t \gtrsim \left\{ \sqrt{pn^2 \ln n}, \sqrt{pn(\ln n)^3}, (\ln n)^2 \right\}$, where the second term is smaller than the first for n large enough. As in the quadruplet setting, we add the probability $\mathbf{P}(\mathcal{E}^c)$ and take expectation over \mathcal{W} to obtain $\|S - \mathbf{E}[S|\mathcal{W}]\|_2 \lesssim \left\{ \sqrt{pn^2 \ln n}, (\ln n)^2 \right\}$ with probability $1 - \mathcal{O}(\frac{1}{n})$.

To bound $\|\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]\|_2$, we note that the entries of the matrix comprises of mutually dependent variables in the set \mathcal{B}' defined in (24). Since the dependency graph for \mathcal{B}' has maximum degree $d = \mathcal{O}(n)$, we partition \mathcal{B}' into $d + 1$ independent sets of nearly identical sizes (equitable colouring). Let $\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S] = A^{(1)} + \dots + A^{(d+1)}$ denote the corresponding partition of the matrix, where $A^{(\ell)} \in \mathbb{R}^{n \times n}$ is a symmetric matrix consisting of the variables in the ℓ -th independent set. Due to the independence of the variables, we have $A_{ij}^{(\ell)} = A_{ji}^{(\ell)} = B'_{ijr}$ or B'_{jir} for some $r \neq i, j$. Hence, each $A^{(\ell)}$ is a symmetric matrix with independent zero-mean entries, bounded by $2p\epsilon$ and variance at most $4p^2\epsilon^2$ (follows from properties of B'_{ijr}). Thus, by matrix Bernstein inequality (Tropp, 2012), we have

$$\mathbf{P}\left(\|A^{(\ell)}\|_2 > \tau\right) \leq n \exp\left(-\frac{t^2}{8p^2\epsilon^2n + \frac{2}{3}p\epsilon t}\right),$$

and combining with the union bound,

$$\begin{aligned} \mathbf{P}\left(\|\mathbf{E}[S|\mathcal{W}] - \mathbf{E}[S]\|_2 > t\right) &\leq \mathbf{P}\left(\max_{\ell \in [d+1]} \|A^{(\ell)}\|_2 > \frac{t}{d+1}\right) \\ &\leq n(d+1) \exp\left(-\frac{\left(\frac{t}{d+1}\right)^2}{8p^2\epsilon^2n + \frac{2}{3}p\epsilon \frac{t}{d+1}}\right) \\ &\lesssim n^2 \exp\left(-\min\left\{\frac{t^2}{p^2\epsilon^2n^3}, \frac{t}{pen}\right\}\right), \end{aligned}$$

which is $\mathcal{O}\left(\frac{1}{n}\right)$ for $t \gtrsim pen \cdot \max\left\{\sqrt{n \ln n}, \ln n\right\}$, where the first term obviously dominates. Combining the above derivations, we have with probability $1 - \frac{1}{4n}$,

$$\|S - \tilde{S}\|_2 \lesssim \max\left\{\sqrt{pn^2 \ln n}, (\ln n)^2, pen\sqrt{n \ln n}, p\epsilon\delta n\right\} \quad (26)$$

where the last term (arising due to $\max_i \tilde{S}_{ii}$) is dominated by the third.

Deriving interval for λ in terms of $|\mathcal{T}|$. We now use (21), (25) and (26) to complete the proof for the triplet setting. We verify the conditions in Proposition 1 by deriving conditions under which $\Delta_2 < \frac{1}{12}\Delta_1$ and $\|S - \tilde{S}\|_2 < \frac{1}{2}n_{\min}\Delta_1$. Similar to the proof for the quadruplet setting, we compare the upper bounds in (16) and (17) with Δ_1 and $n_{\min}\Delta_1$, respectively. As in the previous setting, the first two bounds in (16)–(17) lead to conditions on p , while the third term leads to a condition on δ . Combining the different cases, it follows that if

$$\delta \gtrsim \frac{\sqrt{n \ln n}}{n_{\min}} \quad \text{and} \quad p \gtrsim \frac{(\ln n)^2}{\epsilon^2\delta^2n_{\min}^2}, \quad (27)$$

then the criteria for Δ_2 and $\|S - \tilde{S}\|_2$ are satisfied, and by Proposition 1, X^* is the unique optimal solution for SDP- λ with the range of λ given by

$$\|S - \tilde{S}\|_2 \lesssim \max\left\{\sqrt{pn^2 \ln n}, p\epsilon\sqrt{n^3 \ln n}, (\ln n)^2\right\} \lesssim \lambda < \frac{p\epsilon\delta n_{\min}(n-2)}{2} = \frac{\Delta_1}{2}. \quad (28)$$

We finally show that the condition on p holds under the stated condition of $|\mathcal{T}| \gtrsim \frac{n^3(\ln n)^2}{\epsilon^2\delta^2n_{\min}^2}$, and state the above interval for λ in terms of $|\mathcal{T}|$. Under the assumption that each triplet is observed independently with probability p , we $\mathbf{E}[|\mathcal{T}|] = pn\binom{n-1}{2} = \mathcal{O}(pn^3)$. By Bernstein inequality, it is easy to verify that for $p \gtrsim \frac{\ln n}{n^3}$ or equivalently $|\mathcal{T}| \gtrsim \ln n$, we have $|\mathcal{T}| \in \left(\frac{1}{2}\mathbf{E}[|\mathcal{T}|], \frac{3}{2}\mathbf{E}[|\mathcal{T}|]\right)$ with probability $1 - \mathcal{O}\left(\frac{1}{n}\right)$. Hence, we can replace p by $\frac{|\mathcal{T}|}{n^3}$ in (27)–(28) up to differences in constants, which leads to the statement of Theorem 1 in the triplet setting.

D Algorithmic details

In this section, we provide details on the modified SPUR algorithm that we use to tune the parameter λ , and to select the number of clusters.

SPUR, acronym for Semidefinite Program with Unknown r (r denoting the number of clusters), was proposed by Yan et al. (2018) to tune the parameter λ of SDP- λ in the context of graph clustering (see Algorithm 2). The underlying idea of this approach is to search for the optimal λ using a grid search over the range $0 < \lambda < \lambda_{\max}$, where λ_{\max} is derived from an exact recovery result under stochastic block model.

Algorithm 2: Semidefinite Program with Unknown k (SPUR).

input : graph A , number of candidates T .

begin

for $t = 1$ to T **do**

$\lambda_t = \exp\left(\frac{t}{T} \ln(1 + \lambda_{\max})\right) - 1$. (Yan et al. (2018) set $\lambda_{\max} = \|A\|_{op}$)

 Solve SDP- λ with $\lambda = \lambda_t$ to obtain X_t .

 Estimate $k_t = \text{integer approximation of trace}(X_t)$.

end

 Choose $\hat{t} = \arg \max_t \frac{\sum_{i \leq k_t} \sigma_i(X_t)}{\text{trace}(X_t)}$, where $\sigma_i(X_t)$ denotes i -th largest eigenvalue of X_t .

end

output : Number of clusters $k_{\hat{t}}, X_{\hat{t}}$.

In the present setting, Theorem 1 shows that the planted clusters can be exactly recovered given a sufficient number of comparisons and an appropriate choice of λ . From Theorem 1, a candidate for λ_{\max} can be chosen as $\frac{\lfloor T \rfloor}{n}$ (for triplets) or $\frac{\lfloor Q \rfloor}{n}$ (for quadruplets), which is a loose upper bound for the theoretical interval for λ , obtained by noting that $\epsilon \delta n_{\min} \leq n$. Thus, following Yan et al. (2018), we could use Algorithm 2 with our choice of λ_{\max} .

Unfortunately, this approach has two main drawbacks. First, it ignores the lower bound in Theorem 1 and, second, setting T , the number of λ values that should be considered in Algorithm 2, is difficult. To address the former issue, we propose to consider Theorem 1 once more and to use $\lambda_{\min} = \sqrt{c(\ln n)/n}$ as a lower bound for λ instead of 0, as used in Yan et al. (2018). To address the latter issue, we use the fact that the estimated number of clusters k monotonically decreases with λ as shown in the next Lemma.

Lemma 1 (The estimated number of clusters decreases monotonically with increasing λ). *For any $\lambda > 0$, let X_λ denote the solution of SDP- λ and $k_\lambda = \lfloor \text{trace}(X_\lambda) \rfloor$ be the integer approximation of $\text{trace}(X_\lambda)$, which is an estimate of the number of clusters. Then, k_λ is a non-increasing function of λ , that is*

$$\lambda' \geq \lambda \Rightarrow k_{\lambda'} \leq k_\lambda.$$

Proof. We start this proof by noting that since k_λ is the integer approximation of $\text{trace}(X_\lambda)$, it suffices to show that $\text{trace}(X_\lambda)$ is a non-increasing function of λ . Then, consider distinct λ', λ and let $X_{\lambda'}, X_\lambda$ be the solutions of SDP- λ with parameters λ', λ , respectively. We have

$$\text{trace}(SX_\lambda) - \lambda \text{trace}(X_\lambda) \geq \text{trace}(SX_{\lambda'}) - \lambda \text{trace}(X_{\lambda'}) ,$$

$$\text{trace}(SX_\lambda) - \lambda' \text{trace}(X_\lambda) \leq \text{trace}(SX_{\lambda'}) - \lambda' \text{trace}(X_{\lambda'}) .$$

Subtracting the second inequality from the first inequality implies

$$\begin{aligned} \text{trace}(SX_\lambda) - \lambda \text{trace}(X_\lambda) - (\text{trace}(SX_\lambda) - \lambda' \text{trace}(X_\lambda)) \\ \geq \text{trace}(SX_{\lambda'}) - \lambda \text{trace}(X_{\lambda'}) - \text{trace}(SX_{\lambda'}) + \lambda' \text{trace}(X_{\lambda'}) \end{aligned}$$

which implies

$$(\lambda' - \lambda) \text{trace}(X_\lambda) \geq (\lambda' - \lambda) \text{trace}(X_{\lambda'})$$

or equivalently, $(\lambda' - \lambda)(\text{trace}(X_{\lambda'}) - \text{trace}(X_\lambda)) \leq 0$. Thus, for $\lambda' > \lambda$, we can conclude that $\text{trace}(X_{\lambda'}) \leq \text{trace}(X_\lambda)$, which shows that $\text{trace}(X_\lambda)$ and k_λ are non-increasing functions of λ . \square

Following this, using λ_{\min} and λ_{\max} , we get two estimates of the number of clusters, $k_{\lambda_{\min}}$ and $k_{\lambda_{\max}}$. Then, we search over $k \in [k_{\lambda_{\max}}, k_{\lambda_{\min}}]$ instead of searching over λ —in practice, it helps to search over the values $\max\{2, k_{\lambda_{\max}}\} \leq k \leq k_{\lambda_{\min}} + 2$. We select k that maximises the above SPUR objective, where X_k is computed using the simpler SDP- k (Yan et al., 2018). This approach is summarized in Algorithm 1 in the main paper.

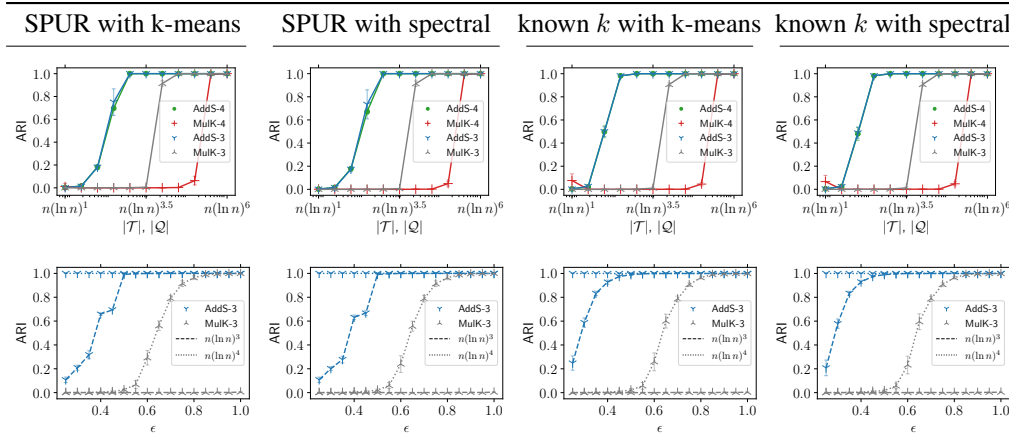


Figure 3: Comparing clustering algorithms to partition X in the last step. Using k-means or spectral clustering does not affect the output of our approach.

E Additional results for the planted model

In this section, we provide additional experiments on our planted model. We show that changing the clustering method used in the last step of our approach to cluster the matrix X learned by SDP- λ or SDP- k does not affect the results. We demonstrate that, given a sufficient number of comparisons, SPUR correctly estimates the number of clusters. We give details on the distributions used in Figure 1c. Finally, we consider several additional experiments where we vary the planted model parameters that were ignored in Section 5 in the main paper.

E.1 Clustering method in the last step

In the last step of our approach, we use k -means to cluster the learned matrix X_k . We experimentally demonstrate here that the partition obtained is, in fact, independent of the clustering algorithm used in this step. Hence, in Figure 3, we compare spectral clustering with k-means. As in the main paper, we here consider varying the number of observations, $|\mathcal{T}|, |\mathcal{Q}|$ and varying the crowd noise ϵ for both the setting where k is estimated by SPUR and where we consider k to be known. There is no differences between the ARI obtained when using k-means or spectral clustering.

E.2 Compare SPUR with known k

An important question is how good is SPUR at estimating the true number of clusters. We illustrate this in Figure 4. We start by comparing the first two columns, showing how the ARI changes for various parameters of the planted model. In the setting of $|\mathcal{Q}|, |\mathcal{T}| = n(\ln n)^3$ we see that using a known number of clusters outperforms SPUR, especially in parameter ranges that are harder to cluster (e.g. small δ, ϵ or for a larger number of clusters). If we consider $|\mathcal{Q}|, |\mathcal{T}| = n(\ln n)^4$, SPUR correctly estimates the number of clusters and thus we omit the plots with known k .

E.3 Experimental details for changing F_{in}, F_{out} in the planted model

In this section, we give implementation details on the different distributions considered in Figure 1c. In the following let ϕ be the normal pdf and Φ the normal cdf. Recall that, in all the experiments, we fix $\delta = 0.5$ as the default.

Parameters for F_{in} and F_{out} normal distributions. Let $F_{in} = \mathcal{N}(\mu_{in}, \sigma)$ and $F_{out} = \mathcal{N}(\mu_{out}, \sigma)$. We fix $\sigma = 0.1$ and $\mu_{out} = 0$. Using δ we can compute μ_{in} . Indeed, in this case, the cumulative distribution function is known and, thus, by setting it equal to $\mathbf{P}_{w \sim F_{in}, w' \sim F_{out}}(w > w') = \frac{1+\delta}{2}$ for some $\delta \in (0, 1]$ (as given in Equation (3)) we directly get the δ defined in Section 2: $\delta = 2\Phi((\mu_{in} - \mu_{out})/(\sqrt{2}\sigma)) - 1$. Then, assuming that $\mu_{out} = 0$, we get $\mu_{in} = \sqrt{2}\sigma\Phi^{-1}(\frac{1+\delta}{2})$.

Parameters for F_{in} and F_{out} Beta distributions. Let $F_{in} = \text{Beta}(\alpha, \beta)$, $F_{out} = \text{Beta}(\alpha', \beta')$. We set $\alpha' = \beta' = 1$ such that $F_{out} = \text{Beta}(1, 1) = \text{Unif}(0, 1)$. We can then compute

$$\begin{aligned} \mathbf{P}_{w \sim \text{Beta}(\alpha, \beta), w' \sim \text{Beta}(1, 1)}(w > w') &= \mathbb{E}_w \left[\int_0^w dw' \right] \\ &= \mathbb{E}_w [w] \\ &= \frac{\alpha}{\alpha + \beta} \end{aligned}$$

where the last line follows from the mean of the Beta distribution. Setting this equal to $\frac{1+\delta}{2}$ and solving for α gives: $\alpha = \beta \left(\frac{1+\delta}{1-\delta} \right)$. In our experiments, we fix $\beta = 2$.

Parameters for F_{in} Normal and F_{out} Uniform. Let $F_{in} = \mathcal{N}(\mu, 0)$, $F_{out} = \text{Unif}(0, 1)$. To set μ , we compute:

$$\begin{aligned} \mathbf{P}_{w \sim \mathcal{N}(\mu, 0), w' \sim \text{Unif}(0, 1)}(w > w') &= \int_0^\infty \phi(w - \mu) dw \left[\int_0^{\min(w, 1)} dw' \right] \\ &= \int_0^1 w \phi(w - \mu) dw + \int_1^\infty \phi(w - \mu) dw + \mu (\Phi(1 - \mu) - \Phi(-\mu)) \\ &= 1 + \phi(-\mu) - \phi(1 - \mu) + (\mu - 1) \Phi(1 - \mu) - \mu \Phi(-\mu) \end{aligned}$$

Solving numerically for μ gives $\mu = \frac{1+\delta}{2}$.

E.4 Influence of different planted model parameters

In this section we present additional experiments where we vary various parameters of the planted model. Recall that we consider the following parameters as default: $n = 1000$, $k = 4$, $\epsilon = 0.75$, $|\mathcal{T}| = |\mathcal{Q}| = n(\ln n)^4$ and $F_{in} = \mathcal{N}(\sqrt{2}\sigma\Phi^{-1}(\frac{1+\delta}{2}), \sigma^2)$, $F_{out} = \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$ and $\delta = 0.5$.

Number of samples n , first row in Figure 4. We can first note that for $|\mathcal{Q}|, |\mathcal{T}| = n(\ln n)^3$ there is no difference in the behaviour between SPUR and known k . Both AddS-3 and AddS-4 achieve full recovery while MulK-3 and MulK-4 predictions are random. To learn somewhat meaningful partitions with MulK-3, one needs to increase the number of observations to $n(\ln n)^4$. However, even with this many comparisons, MulK-4 still learns random clusters.

Intrinsic noise δ , second row in Figure 4. Using $|\mathcal{Q}|, |\mathcal{T}| = n(\ln n)^3$, we see that, for both SPUR and known k , AddS-3 and AddS-4 exactly recover the clusters even when the intrinsic noise is high, that is $\delta = 0.4$. MulK-3 and MulK-4 can only make random predictions in this case. When the number of observations increases to $n(\ln n)^4$, AddS-3 and AddS-4 exactly recover the clusters even for values of δ that are as small as 0.25. In this case, MulK-4 still predicts random clusters, while MulK-3 is able to recover the clusters when the intrinsic noise is sufficiently small, that is $\delta \geq 0.6$.

Crowd noise ϵ , third row in Figure 4. This parameter was already analyzed in the main paper. The plots are recalled here for the sake of completeness.

Number of clusters k , fourth row in Figure 4. Finally, we vary the number of planted clusters. Here, we observe the most noticeable difference between SPUR and known k . For $|\mathcal{Q}|, |\mathcal{T}| = n(\ln n)^3$, AddS-3 and AddS-4 with SPUR achieve perfect recovery for up to five clusters. While we notice a similar behaviour for AddS-3 and AddS-4 with known k , the drop in ARI only starts for $k > 7$ and is far less important than with SPUR. For $n(\ln n)^4$ observations AddS-3 and AddS-4 consistently recover all the clusters. On the other hand, MulK-3 only recovers clusters up to $k = 3$ (here, MulK-3 uses the number of clusters estimated by AddS-3 with SPUR, that is $k = 3$). Once again, MulK-4 can only make random predictions.

F Further results for experiments on real comparison based data

In this final section we present supporting results for the real data experiments presented in Section 5.

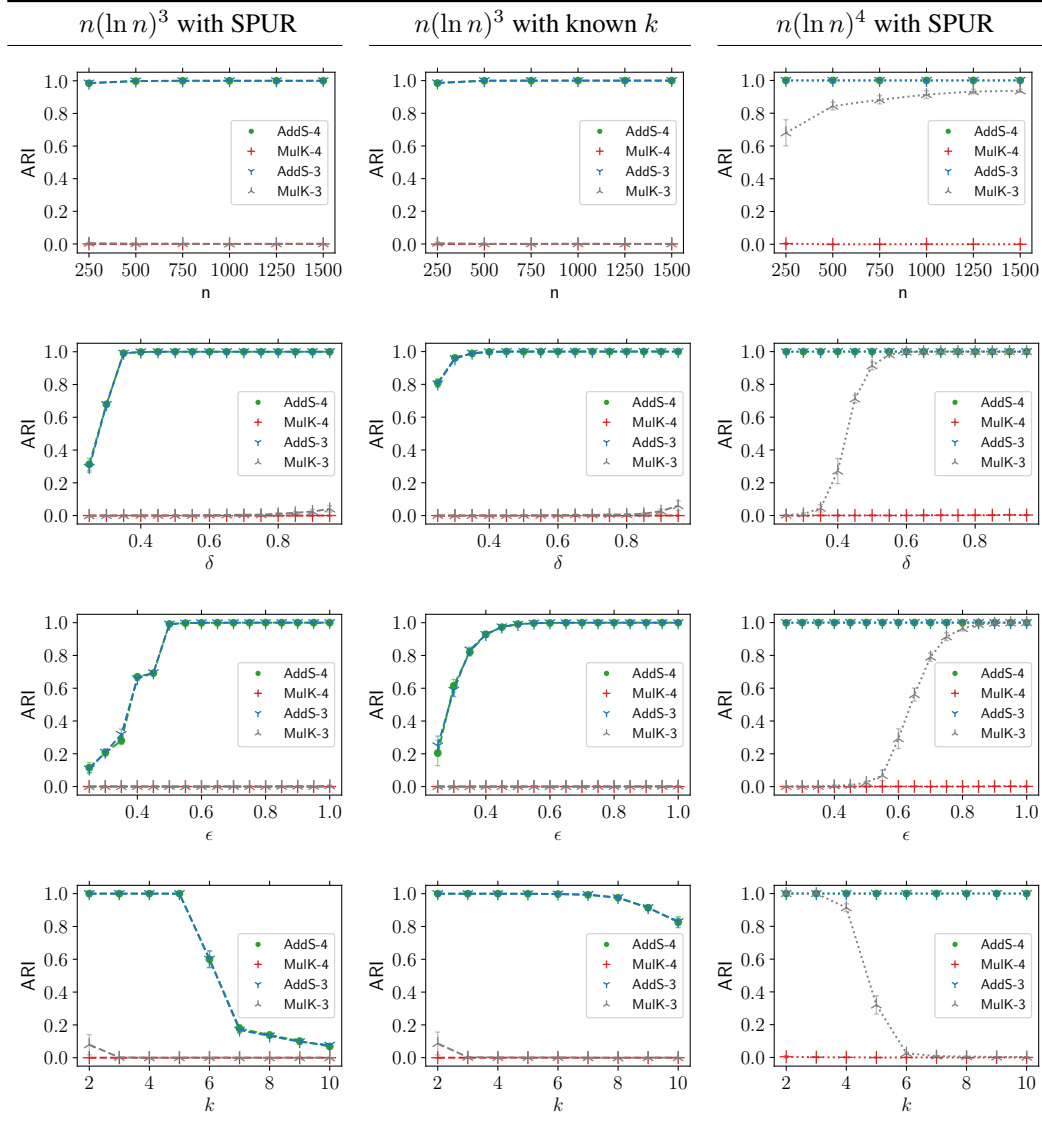


Figure 4: Further experiments on the planted model. On the one hand, SPUR needs sufficiently many comparisons to correctly estimate the number of underlying clusters. On the other hand, our approaches are not overly sensitive to changes in the planted model parameters and are able to exactly recover the planted clusters with $n(\ln n)^3$ comparisons even in fairly difficult cases (small δ , high k , ...). Furthermore, given $n(\ln n)^4$ comparisons, our approaches are able to exactly recover the planted clusters in all the considered cases.

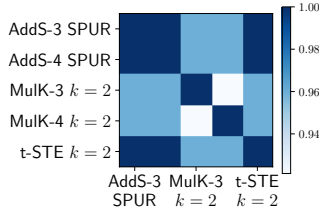


Figure 5: ARI between the clustering obtained by the different baselines. AddS-3 and AddS-4 with SPUR both estimate that the number of cluster is $k = 2$. There is a high degree of agreement between the different approaches.

F.1 Details on the Car dataset

The Car dataset (Kleindessner and von Luxburg, 2016) is a comparison based dataset that contains 60 examples grouped into 3 classes (SUV, city cars, sport cars) with 4 outliers. This dataset originally comes with a set of 6056 comparisons of the form “ x_i is most central in the triple x_i, x_j, x_k .” Each of these comparisons corresponds to two triplets: “ x_j is more similar to x_i than to x_k ” and “ x_k is more similar to x_i than to x_j .” Hence, we have access to 12112 triplet comparisons.

F.2 Food Dataset

In addition to the Car dataset we now look at a second comparison based dataset called Food (Wilber et al., 2014). It contains 100 food images and comes with 190376 triplet comparisons. Since there are no ground truth labels for the food dataset, we use the number of clusters estimated by SPUR for all methods and plot, in Figure 5, the similarity matrix between the different clustering approaches considered. Here, there is a high degree of agreement between all the clustering methods. Thus, most approaches predict the same clusters up to minor differences for a few data points. In Figure 7, we plot the clusters obtained by AddS-3 with SPUR (estimated k is 2). The two clusters seem to separate *Sweet foods* from *Savoury foods*. Intuitively, it seems indeed natural for humans to judge that two sweet foods are more similar to each other than to a third savoury food.

F.3 MNIST

In this section, we consider additional experiments on the MNIST dataset. First, we consider a second similarity measure to generate the triplets. Then, we illustrate the partitions obtained with AddS-3 with known k and SPUR respectively.

Gaussian similarity. In the main paper, we use the Gaussian similarity to generate the comparisons. More precisely, we compute the similarity between two examples x_i and x_j as

$$w_{ij} = \exp\left(\frac{\|x_i - x_j\|_2^2}{\gamma^2}\right) \text{ with } \gamma = 1.$$

Cosine similarity. Instead of the Gaussian similarity, we could consider alternatives to generate the comparisons. For example, the Cosine similarity:

$$w_{ij} = \frac{\langle x_i, x_j \rangle}{\|x_i\|_2 \|x_j\|_2}.$$

In Figure 6, we show that using this alternative similarity affects the absolute results of the considered approaches. However, it does not change the overall trend, that is, as the number of comparisons increases, AddS-3 converges to the baseline of k -means with access to the original similarities.

Clustering using known k . Figure 8a shows the t-SNE embedding of 2000 MNIST samples of all ten classes, where we see a clear separation between some classes (for example, 0 and 1) and very close embedding between others (for example, 1 and 9). Note that the classes obtained by AddS-3 are shown up to permutations and may not reflect the majority label in the different clusters. Further note that the data presented here corresponds to a single repetition out of the 10 repetitions used to

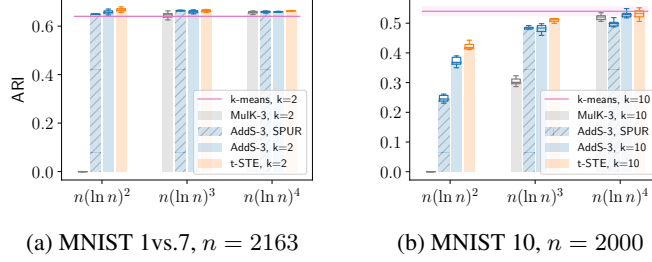
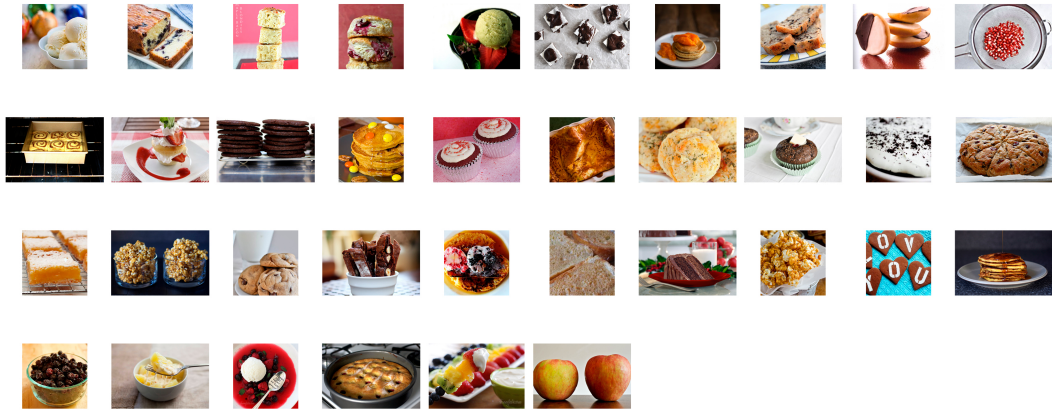


Figure 6: Experiments on MNIST using the cosine similarity. The absolute ARI performances are different from the Gaussian similarity. However, the overall trend is preserved and, given sufficiently many comparisons, all the ordinal baselines reach the performance of k -means on the original data.

compute the mean ARI (with standard deviation) in the main paper and this appendix. In Figure 8d, we see that, for $|\mathcal{T}| = n(\ln n)^2$, the learned partition is not very representative of the original labels. Figure 8c shows that, when the number of comparisons increases to $|\mathcal{T}| = n(\ln n)^3$, the recovery ability of AddS-3 is greatly improved. However, the obtained partitions are not entirely satisfactory. Finally, Figure 8b shows that, when the number of comparisons further increases to $|\mathcal{T}| = n(\ln n)^4$, the clustering obtained is close to the true labeling and most clusters are correctly identified.

Clustering using SPUR. In this second set of experiments, we extend our observations from the previous paragraph to the labeling obtained by AddS-3 using SPUR. One can note that SPUR always underestimates the number of clusters. Hence, in Figure 9a, with $|\mathcal{T}| = n(\ln n)^3$, the number of predicted clusters is $k = 6$ while, in Figure 9b, with $|\mathcal{T}| = n(\ln n)^4$, the number of predicted clusters is $k = 8$. This explain the slightly worse behaviour of SPUR compared to known k in Figure 2b in the main paper. Nevertheless, the difference in average ARI is not so significant when $|\mathcal{T}| = n(\ln n)^4$, suggesting that 8 clusters is, in fact, a good estimate of the number of clusters that can reliably be distinguished by the different methods.

Class A with 36 images (*Sweet?*)



Class B with 64 images (*Savoury?*)

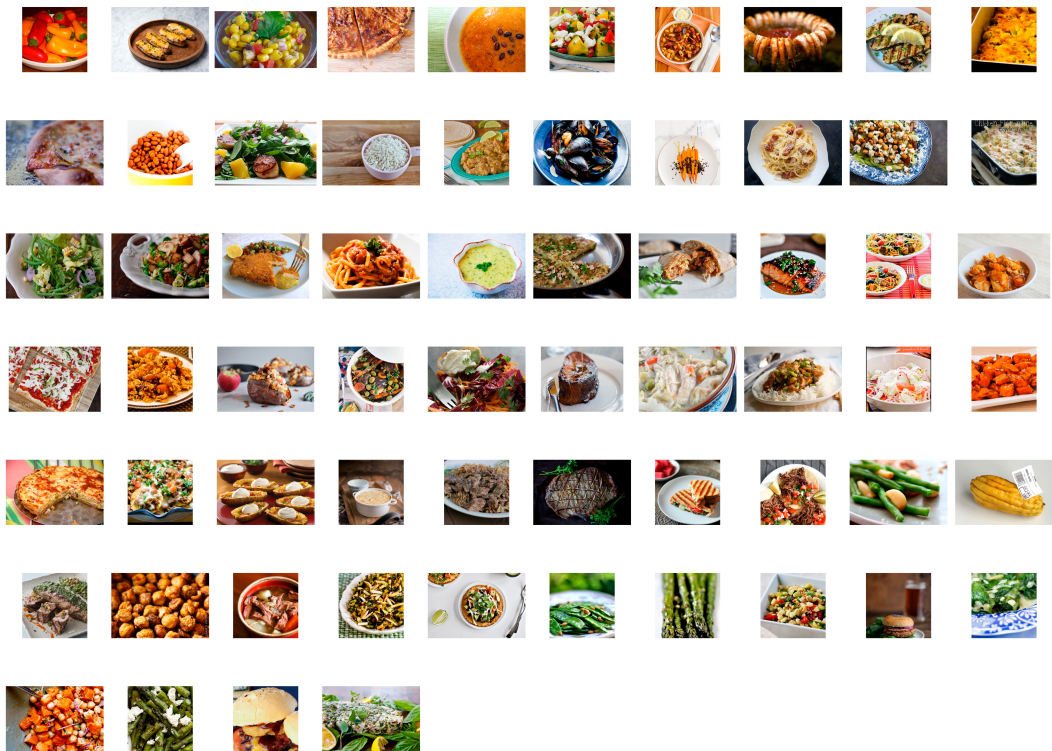


Figure 7: Clusters obtained by AddS-3 on the food dataset. It seems that the Sweet foods are separated from the Savoury ones.

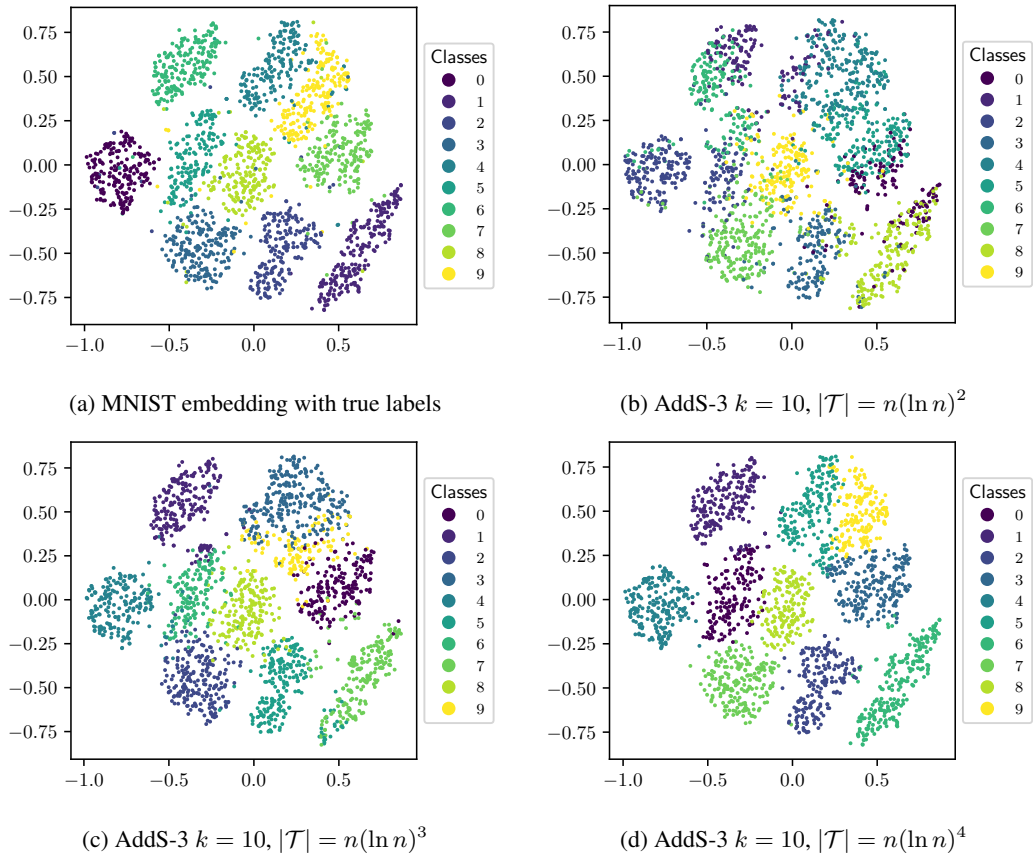


Figure 8: t-SNE embedding of 2000 MNIST samples with (8a) true labeling and (8d)–(8b) clusters obtained by AddS-3 with known $k = 10$ and varying number of observations. The classes are given up to permutations and may not reflect the majority label in each cluster.

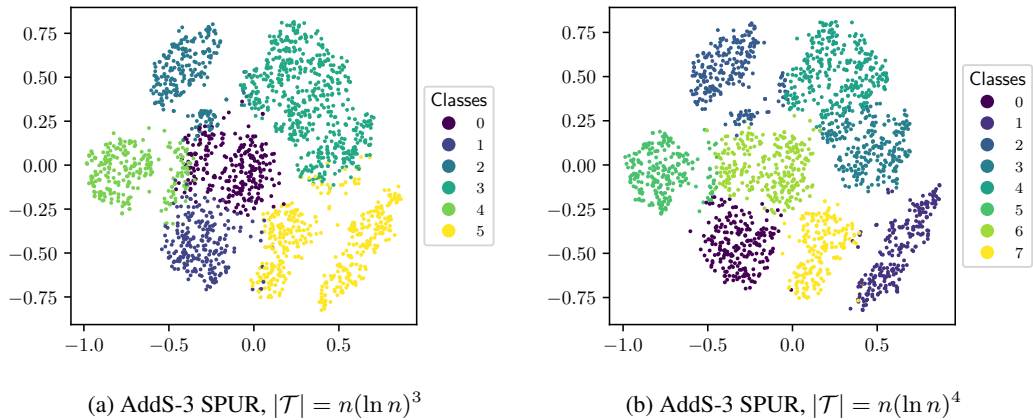


Figure 9: t-SNE embedding of 2000 MNIST samples with the clusters predicted by AddS-3 using SPUR and varying number of comparisons. The classes are given up to permutations and may not reflect the majority label in each cluster.