Metric Clustering From Triplet Comparisons

Gokcan Tatli
Dept. of ECE
University of Wisconsin-Madison
Madison, WI, USA
gtatli@wisc.edu

Ramya Korlakai Vinayak
Dept. of ECE
University of Wisconsin-Madison
Madison, WI, USA
ramya@ece.wisc.edu

2024 60th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3315-4103-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ALLERTON63246.2024.10735326

***Abstract*—Using triplet comparison queries of the form "*Do you think item a is more similar to item b or item c?*" to learn a positive definite matrix to capture a distance metric in $\mathbb{R}^d$ has been a popular approach to capture how human perceive similarity and differences between various objects/concepts. Most of the existing works focus on learning a single metric using data from all people in the dataset. However, people can systematically differ in their notions of similarity over a set of objects due to their diverse backgrounds. Therefore, using a single metric for everyone has limited capacity in capturing the heterogeneity while modeling how people perceive objects in populations that have diverse subgroups. The subgroup structure is often salient and difficult to know a priori. We propose to learn the subgroup structure from the answers to triplet queries by clustering the user-triplet observation matrix. By modeling the problem of metric clustering as a low-rank matrix recovery problem, we leverage convex optimization based approach to perform clustering. We provide analysis for two cluster case that sheds light on when the approach succeeds and fails as function of distance between the metrics, size of the clusters, number of triplet queries answered per person and the noise level in the answers obtained. We validate our results through extensive simulations. Furthermore, we also provide analysis that shows how an outlier impacts the discovery of cluster structure.**

***Index Terms*—metric learning, distance learning, clustering, biclustering, low-rank matrix recovery**

I. Introduction

Understanding how people make preference decisions plays an important role in different areas ranging from recommendation systems [1], [2] to crowdsourced democracy [3], [4]. While comparing different options, what options are considered similar and dissimilar by the users can be helpful in modeling and predicting future preference decisions based on the past data. Metric learning [5] is a popular approach to learn such a mapping that represents similarities in the perception of a set of objects by people. Learning such a metric plays an important role in many human involved tasks [6] and helps to understand how humans perceive relations between different objects. A popular way to model human preference judgements is to use Mahalanobis distance $\mathrm{d}_{\mathbf{M}}(\boldsymbol{x},\boldsymbol{y}) := ||\boldsymbol{x}-\boldsymbol{y}||_{\mathbf{M}}$ as the distance function between the options or items with known representations $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d$, where $\mathbf{M}$ is a $d \times d$ positive semidefinite matrix and $\|\boldsymbol{x}\|_{\mathbf{M}} = \sqrt{\boldsymbol{x}^T\mathbf{M}\boldsymbol{x}}$. One can also view learning such a metric as learning an embedding in a d-dimensional Euclidean space such that the distances between the objects reflect the similarities perceived by humans. This is known as non-metric MDS setting which has been considered first in [7], [8]. Designing better queries for getting dissimilarities among items, query selection algorithms and more are studied [9].

Learning the distance metric from triplet comparisons in the form of "*Do you think item a is more similar to item b or item c?*" is a well studied problem [10]–[15], see [5], [16] for a survey. Specifically, learning a sparse/low-rank metric $\mathbf{M}$ from triplet comparison queries is also known as learning a linear metric in the literature. Vast majority of the literature focuses on learning single metric for the whole population. However, diverse groups of people have significant differences on how they perceive objects. Therefore, a single metric $\mathbf{M}$ is not often enough to model distance metrics of diverse populations. If we know the subgroups of people, we can apply metric learning to each of the subgroups separately. However, in practice, these groups emerge from complex interactions of various factors such as differences in demographics, language, culture, educational background and so on. Therefore, it is difficult to a priori decide the subgroups which pose a challenge to learning diverse set of metrics.

Our Contributions. In this paper, we propose to first find the latent subgroups in the population before proceeding to metric learning by using the answers provided by users to the triplet queries. We study the problem of clustering populations based on shared metric within subgroups by using answers to triplet comparison queries and make the following contributions: (1) *Modeling metric clustering as low-rank matrix recovery*: By representing the answers given by users to the triplet queries using a data matrix, where each row represents a specific user and each column represents answers to a specific triplet query, we formulate this problem as the problem of clustering rows of a binary matrix. Since it is impossible to expect every single user to answer all triplet queries, our target is to find cluster structure in the population using partial observations of this data matrix. Our formulation allows us view the problem of metric clustering from triplet comparisons as *recovering low-rank matrix* from a partially observed noisy binary matrix. Inspired from the rich literature on low-rank matrix recovery via nuclear norm regularization, we propose

This work was partially supported by NSF grants NCS-FO 2219903 and NSF CAREER Award CCF 2238876.

semidefinite program to recover the low-rank matrix that reflects the cluster structure. (2) Under assumptions of independent noise and partial observations for two cluster setting, we provide analysis of the proposed convex optimization approach that shows the success and failure conditions for recovery of the true user-triplet matrix. We also provide simulation results that confirm our success and failure bounds, and show transition from failure to success. (3) We also provide analysis of how this convex approach behaves for the metric clustering problem in the presence of an outlier.

The rest of the paper is structured as follows: In Section II, we provide details of our problem formulation and modeling. In Section III, we state our main results that shed light on when our proposed convex optimization approach is able to correctly recover the true user-triplet matrix and when it fails to do so. We also state results on analysis of the behavior of the optimization program under the presence of an outlier – that is a user who does not share the metric of a group.

II. Modeling Metric Clustering as Low-rank Matrix Recovery

In this section we describe our problem formulation, modeling and the optimization approach for metric clustering via triplet queries. We begin by describing the problem setting, notations and assumptions.

A. Problem Setting and Assumptions

Let $\mathcal{T}$ denote the set of triplets used for querying the users, and $|\mathcal{T}|$ represent the size of that set. Let $\{(\boldsymbol{x}_{a_j}, \boldsymbol{x}_{b_j}, \boldsymbol{x}_{c_j})\}_{j=1}^{|\mathcal{T}|}$, where $\boldsymbol{x} \in \mathbb{R}^d$, denote the representation of the items in the $|\mathcal{T}|$ triplets in the query set. Let m be the the total number of users under consideration. We assume that there are two disjoint clusters among the users who differ in how they perceive similarity across the objects. That is, there exist two different metrics $\mathbf{M}_1$ and $\mathbf{M}_2$ corresponding to the two clusters of users. Let m_1 and m_2 represent sizes of the two cluster of users and $m = m_1 + m_2$. Our goal is to cluster the users using their responses to triplet comparisons. Once the clusters are recovered, one can proceed to recover metrics $\mathbf{M}_1$ and $\mathbf{M}_2$ from these responses using approaches for metric learning via triplets [5], [11].

In the noiseless setting, the interaction protocol between the users and the queries is as follows: We consider a triplet query consisting of a triplet of objects whose representations are $(\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c)$. Let $i \in \{1, 2\}$ denote the cluster identification, query result y_i for triplet $t = \{a, b, c\}$ can be written as

$$y^i_{t=\{a,b,c\}} = \begin{cases} 1, & \text{if } \|\boldsymbol{x}_a - \boldsymbol{x}_b\|_{\mathbf{M}_i} > \|\boldsymbol{x}_a - \boldsymbol{x}_c\|_{\mathbf{M}_i} \\ -1, & \text{otherwise.} \end{cases}$$

As the data available to us is answers by the users to triplet comparisons of the type, “Do you think a is closer to b or c?”, how different the answers are to the set of triplet queries for the two metrics $\mathbf{M}_1$ and $\mathbf{M}_2$ plays an important role in determining how easy or difficult the clustering problem is. Therefore, the query triplet set $\mathcal{T}$ should have a subset of triplets such that responses to those triplets under the metrics $\mathbf{M}_1$ and $\mathbf{M}_2$ are different. It is important to note that the query set $\mathcal{T}$ should be diverse and large to be able to ensure ability to distinguish between the metrics which are unknown a priori. We assume that $\mathcal{T}$ is a *distinguishable* set for $\mathbf{M}_1$ and $\mathbf{M}_2$. Otherwise, it is impossible to distinguish them. Our analysis and results (Section III) will reveal the exact nature of how this *difference* between the metrics will play a role.

Noting that it is often difficult to have each user to answer every triplet query and that there can be noise in the answers, we make the following assumptions to capture these aspects:

A1. (*Partial observations*) Each person answers each query independently with probability $0 < r \leq 1$,

A2. (*Noisy answers*) Answer to each query is flipped independently with probability $q < 0.5$.

B. Optimization Approach

Consider the user-triplet matrix under noiseless setting and full observations, i.e., each user answers all triplets. Let $\mathbf{L}^* \in \mathbb{R}^{m \times |\mathcal{T}|}$ represent this true user-triplet matrix, where each row represents binary answers of a user and each column represents answers to a specific triplet query. Note that $\mathbf{L}^*$ is a low-rank matrix with entries from the set $\{-1, +1\}$. In the case of two clusters, this binary matrix will have rank equal to 2 as there are only two possible rows. With noisy and partial observations, what we have is a *perturbation* of this true rank-2 matrix $\mathbf{L}^*$ with only a random subset of the entries revealed. Let $\mathbf{A} \in \mathbb{R}^{m \times |\mathcal{T}|}$ denote the *observation matrix* where each row corresponds to a user and each column corresponds to a triplet. If a user i responds to a triplet j, then corresponding entry $\mathbf{A}_{ij}$ is either 1 or -1 depending on their answer, and if the user i does not respond to triplet j, then $\mathbf{A}_{ij} = 0$. That is, observed entries of $\mathbf{A}$ are either 1 or -1 and the unobserved entries are filled with 0.

Our modeling thus allows us to view the problem of clustering metrics via triplet queries as low-rank matrix recovery from noisy and partially observed matrix $\mathbf{A}$. There is a rich literature on clustering using low-rank matrix recovery approach [17]–[26]. Taking inspiration from this line of literature, we aim to estimate clusters in the population by recovering the user-triplet low-rank matrix $\mathbf{L}^* \in \mathbb{R}^{m \times |\mathcal{T}|}$ using following convex optimization approach:

$$\begin{aligned} \underset{\mathbf{L}}{\text{minimize}} \quad & \lambda \|\mathbf{L}\|_* - \langle \mathbf{A}, \mathbf{L} \rangle \\ \text{subject to} \quad & 1 \geq \mathbf{L}_{i,j} \geq -1 \text{ for all } i, j, \end{aligned} \tag{1}$$

where $||\mathbf{L}||_*$ denotes the nuclear norm of matrix $\mathbf{L}$ which is the sum of singular values of the matrix. We consider this approach as *successful* if the argument of this optimization yields a unique solution that is equal to $\mathbf{L}^\star$ and deem it a failure otherwise. In Section III, we provide details of our analysis that sheds light on when this approach succeeds and when it fails.

Remark 1. *We note that while our theoretical analysis primarily focuses on two cluster setting, our modeling approach and the optimization problem described above apply to the settings with multiple disjoint clusters.*

C. Key Quantities and Additional Notation

As noted earlier, to be able to distinguish the two clusters, it is important that the set of triplet queries used have a subset of queries where the answer to the triplet queries is different under the two metrics. Let $\mathcal{T}_m \subset \mathcal{T}$ denote the subset of triplets on which both the subgroups of people give the same answer when there is no noise, i.e., matched queries, and $\mathcal{T}_{mm} \subset \mathcal{T}$ be the subset of queries on which answers from different subgroups of people differ, i.e., mismatched queries. Cardinality of the set of matched $(\mathbf{y}_t^1 = \mathbf{y}_t^2)$ and mismatched $(\mathbf{y}_t^1 \neq \mathbf{y}_t^2)$ queries can be expressed as follows,

$$|\mathcal{T}_{mm}| = \frac{1}{2}\sum_{t\in\mathcal{T}} |y_t^1 - y_t^2|, \qquad |\mathcal{T}_m| = \frac{1}{2}\sum_{t\in\mathcal{T}} |y_t^1 + y_t^2|$$

Let $m_0 = \min\{m_1, m_2\}$ denote the size of the smallest cluster and $\mathcal{M}_0 = \max\{m_1, m_2\}$ denote the size of the largest cluster. Let $\tau_0 = \min\{|\mathcal{T}_m|, |\mathcal{T}_{mm}|\}$ and $\mathcal{T}_0 = \max\{|\mathcal{T}_m|, |\mathcal{T}_{mm}|\}$, denoting the smallest and the largest number of columns of $\mathbf{L}^\star$ that are either all similar or all different. Let $n = m + |\mathcal{T}|$ denote the total number of users and triplets. Let $\omega = \sqrt{m|\mathcal{T}| + 4\sqrt{m_1 m_2 \mathcal{T}_m \mathcal{T}_{mm}}}$. We note that ω can be thought as an indicator of inequality among cluster sizes in the submatrix level. As the sizes get close to each other, ω increases for constant m and $|\mathcal{T}|$, since the term $m_1 m_2 \mathcal{T}_m \mathcal{T}_{mm}$ gets larger. We observe a similar trend for Δ, where it gets larger, as the cluster sizes get closer.

Our analysis shows that the following key quantities emerge as important in determining when the optimization approach proposed in (1) succeeds or fails in recovering the true $\mathbf{L}^\star$:

- Maximum value that defines a lower bound on λ without which the optimization problem 1 would be a failure:
$$\Sigma_{\text{fail}} = \sqrt{rq\frac{\mathcal{M}_0 \mathcal{T}_0}{\min\{m, |\mathcal{T}|\} + 1}}.$$
- The strength of signal in the observation about the cluster structure:
$$\Delta = r(1-2q)\omega\frac{\sqrt{m_0\tau_0}}{\sqrt{m_0\tau_0} + \sqrt{\mathcal{T}_0\mathcal{M}_0}}.$$
- Minimum value λ for success:
$$\Sigma_{\text{succ}} = 2\left(\sqrt{nr(1 - r + 4rq(1-q))}\right).$$

We use $\mathbf{H}_\alpha$ to represent the entries of $\mathbf{H}$ corresponding to position pairs (i,j)'s in the set α for a given matrix $\mathbf{H}$.

III. Main Results

In this section, we describe the main results of our analysis of optimization approach in (1) that sheds light on when it succeeds in exactly recovering the true user-triplet matrix $\mathbf{L}^\star$ and when it fails to do so.

A. Conditions for success and failure without outliers

Let $\mathbf{A}$ be the observation matrix generated by the responses to triplet queries in $\mathcal{T}$ according to the model, under the assumptions **A1** and **A2**, defined in Section II-A. Then we have the following result that captures when the optimization approach in (1) succeeds,

Theorem III.1 (**Condition for success**). *Given $\epsilon > 0$, there exist positive constants c_1 and c_2 such that, with probability at least $1 - c_1 m|\mathcal{T}|\exp(-c_2\min\{m_0, \tau_0\})$, the optimization in* (1) *recovers $\mathbf{L}^*$ when the following condition holds,*

$$\Delta(1-\epsilon) \geq \lambda \geq (1+\epsilon)\Sigma_{succ}, \tag{2}$$

From the above result, we also note that the *signal* has to be larger than a certain *noise* floor, that is, $\Delta > \Sigma_{\text{succ}}$, for success condition in Theorem III.1 to hold. Reflecting on this condition further, we can obtain the following sufficient condition on the size of the smallest cluster for the optimization in (1) to have a regularizer that leads to successful recovery of $\mathbf{L}^\star$.

Proposition 1. *(Minimum Cluster Size) For the condition $\Delta > \Sigma_{succ}$ to hold, the minimum cluster size has to satisfy the following,*

$$m_0 > \frac{4n}{\tau_0(1-2q)^2}\left(\frac{1}{r} - 1 + 4q(1-q)\right).$$

Remark 2. *The above proposition can also be interpreted in terms of the minimum number of triplets (matched or mismatched) in a cluster that is sufficient for recovery:*

$$\tau_0 > \frac{4n}{m_0(1-2q)^2}\left(\frac{1}{r} - 1 + 4q(1-q)\right)$$

The following theorem captures conditions under which the optimization problem in (1) fails with high probability to recover the true $\mathbf{L}^\star$.

Theorem III.2 (**Conditions for failure**). *Given $\epsilon > 0$, there exist positive constants c_1, c_2 such that,*

1) *If $\lambda \leq (1+\epsilon)\Sigma_{fail}$, then the optimization in* (1) *fails to recover $\mathbf{L}^*$ with probability $1 - c_1\exp(-c_2 m_0\tau_0)$.*
2) *If $\lambda \geq (1+\epsilon)\Sigma_{succ}$ and $\lambda \geq \Delta(1+\epsilon)$, then the optimization in* (1) *fails to recover $\mathbf{L}^*$ with probability $1 - c_1\exp(-c_2 m_0\tau_0)$.*

This theorem provides insights on how the value of regularizer is important for the recovery. If it is set too small, or too large, then the optimization problem will fail to recover $\mathbf{L}^\star$ with high probability.

B. Analysis in the presence of outliers

In previous sections, we focused on understanding the conditions for success and failure for recovering the true user-triplet matrix $\mathbf{L}^\star$ for the case with two clusters and no outliers. Besides having noisy answers, population may also have outliers, that is, people whose preferences cannot be modelled reasonably well by common metrics for subgroups. In this section, we focus on understanding the impact of outliers on the approach using the optimization problem in (1) in recovering $\mathbf{L}^\star$. If the outliers can be separated or if they merge with some of the subgroups *without* altering the corresponding entries of $\mathbf{L}^\star$ for the subgroups to which the outlier(s) merge into, then it would still be a successful scenario. This is because the rows of $\mathbf{L}^\star$ corresponding to different clusters provide the noiseless and complete answers to all the triplet queries under the metric corresponding to that cluster. Therefore, if they are not altered, then the downstream task of learning the metrics for different subgroups is not altered. So, we aim to study conditions that make these scenarios possible. For simplicity, we focus first on the scenario where there is just one cluster, i.e., just one metric, and one outlier.

Let m represent the number of people in the cluster. Consider $\mathbf{M}_{in}$ as the true metric of the cluster and $\mathbf{M}_{out}$ as the metric of the outlier. $\mathcal{Y}_{in}$ and $\mathcal{Y}_{out}$ are the corresponding binary vectors with true answers to triplets $\mathcal{T}$. Define distance d as $0.5\min\{|\mathcal{Y}_{in} - \mathcal{Y}_{out}|, |\mathcal{Y}_{in} + \mathcal{Y}_{out}|\}/|\mathcal{T}|$. We wish to recover non-outlier entries of $\mathbf{L}^*$ correctly.

Suppose $\mathbf{L}^c \in \{-1, +1\}^{(m+1)\times|\mathcal{T}|}$ is the matrix with all rows equal to $\mathcal{Y}_{in}$. The row corresponding to the outlier in $\mathbf{L}^c$ is actually $\pm\mathcal{Y}_{in}$, depending on which is closer. That is, the outlier in $\mathbf{L}^*$ is merged with the cluster in $\mathbf{L}^c$. Let $\mathbf{A}$ be the observation matrix generated by the responses to triplet queries $\mathcal{T}$ according to the model, under the assumptions **A1** and **A2**, defined in Section III. We provide following result that captures when the optimization in (1) merges outlier to the cluster.

Theorem III.3 (**Conditions for Merging**). *Given $\epsilon > 0$, there exist positive constants c_1, c_2 such that, optimization in* (1) *merges the outlier with the cluster and generate $\mathbf{L}^c$, with probability $1 - c_1(m+1)|\mathcal{T}|\exp\left(-c_2\min(m+1, |\mathcal{T}|)\right)$, when*

$$(1-\epsilon)r(1-2q)(1-2d)\sqrt{(m+1)|\mathcal{T}|} \geq \lambda \geq (1+\epsilon)2\left(\overline{\sigma}\sqrt{n}\right),$$

where $d = 0.5\min\{|\mathcal{Y}_{in} - \mathcal{Y}_{out}|, |\mathcal{Y}_{in} + \mathcal{Y}_{out}|\}/|\mathcal{T}|$ and $\overline{\sigma} = \sqrt{r(1-r+4r(q+d-2qd)(1+2qd-q-d))}$. We note that this result can be viewed as having the model in Section III-A with q' instead of q without the outlier, where $q' = q + d - 2qd$.

Theorem III.3 provides hints on how the value of regularizer is important for the optimization problem in (1) when there is an outlier.

Remark 3. *We provide an analysis of outliers with only one outlier and one cluster. As long as outliers are not related to each other, i.e., when they do not share a common metric and distinctive entries are at random locations, we can easily extend Theorem III.3 to capture multiple outliers.*

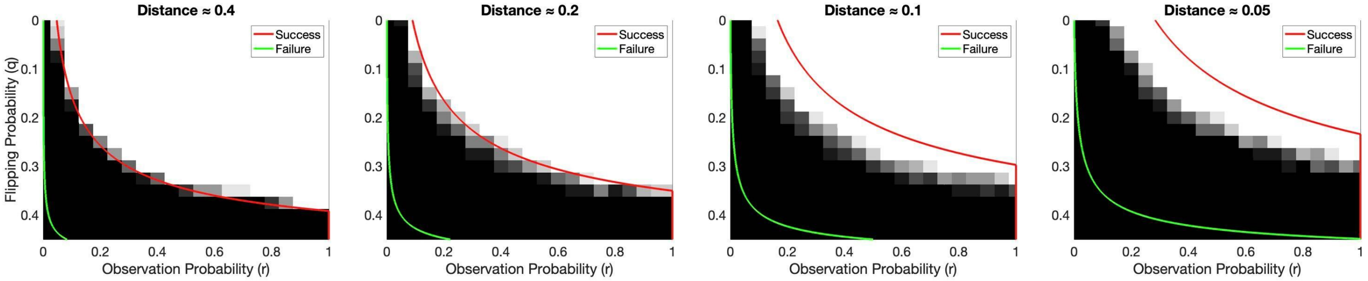
Reflecting on Remark 3, it is not that straightforward what happens even with 2 outliers, when outliers are related and share common structures. One might expect different set of outcomes from the optimization in (1). We can list them as (1) Whole $\mathbf{L}^*$ is recovered correctly and outliers are separated, (2) Outliers merge with the same or different clusters, (3) Outliers emerge as another cluster, and (4) One outlier merges with a cluster and the other one is separated. Therefore, we leave a more comprehensive understanding of outlier behavior for the metric clustering problem as a future work.

IV. Simulations

In this section, we provide simulation results to illustrate our theoretical findings. We run all the simulations on Matlab version R2023b [27]. Note that we use the Alternating Direction Method of Multipliers (ADMM) solver given in [26] (based on [28]) to solve the optimization in (1). First, we illustrate that Σ_{succ} and Σ_{fail} bounds are valid for λ. For this, we keep total number of people as 500, i.e., $m_1 + m_2 = 500$ with equal sizes and sample 50 objects from a normal distribution and generate all possible 58800 triplets. In each trial, we randomly pick 500 triplets out of 58800. Note that there will be sampling error on the number of matched and mismatched queries.

We vary observation probability from 0.05 to 1 in steps of 0.05. That is, portion of the triplets that each user responds on average changes from 0.05 to 1. Similarly, we flip each entry with q, where q is ranging from 0 to 0.45 in steps of 0.025. Then, we apply ADMM solver ([26]) and take sign$(\cdot)$ of the resulting matrix. We repeat each experiment 10 times. To overcome numerical errors due to precision, we consider it as successful recovery if the number of different entries between the resulting matrix and $\mathbf{L}^*$ is no more than 0.1% of all entries. One might also apply k-means algorithm after ADMM solver. Similar techniques are considered in [20], [26].

Our plots in Figure 1 show the phase transitions between Σ_{fail} and Σ_{succ} for varying values of λ depending on r and q, while keeping rest of the parameters fixed. As expected from Theorem III.1 and III.2, transition occurs somewhere between failure, Σ_{fail}, and success thresholds, Σ_{succ}. Note that black and white colors represent failure and success regions respectively. We repeat simulation results with different set of metrics $\mathbf{M}_1$ and $\mathbf{M}_2$, where the ratio of mismatched triplets differ depending on $\mathbf{M}_1$ and $\mathbf{M}_2$. We observe that transition curves move towards right in Figure 1 as the ratio of mismatched entries decreases, and successful recovery gets more difficult as expected. We also note from Figures1(a) and (b) that the condition for success of the optimization problem in (1) obtained by our analysis seem to be in general tight.



(a) (b) (c) (d)

Fig. 1. Success (white) and failure (black) regions with decreasing distance values from left to right for $\lambda = 0.9D$. Red and green curves show success and failure thresholds corresponding to $\lambda > \Sigma_{\text{succ}}$ and $\lambda < \Sigma_{\text{fail}}$, respectively. Based on distance values, the number of mismatched queries, i.e., $\mathcal{T}_{mm}$, decreases from left to right.

V. Proofs

We analyze the convex optimization approach in (1) for the metric clustering problem. As we noted, nuclear norm minimization based low rank matrix recovery is extensively used in the literature and considered in different contexts. Our proofs use standard techniques used in the analysis of convex programs and follow similar lines with [26], [29]–[32], inspired by robust PCA analysis [33], [34], within the context of metric clustering problem.

A. Failure Conditions

Without loss of generality, we can suppose that entries of $\mathbf{L}^*_{i,j}$ corresponding to $\mathcal{T}_m$ are all 1. When we exchange second and third items in a triplet, corresponding column in $\mathbf{L}^*$ is multiplied with -1 since $y^i_{t=\{a,b,c\}} = -y^i_{t=\{a,c,b\}}$. Similarly, we can suppose that entries of $\mathbf{L}^*$ corresponding to $\mathcal{T}_{mm}$ are 1 in the first cluster and -1 in the second cluster. Therefore, we have two different clusters in columns of $\mathbf{L}^*$ corresponding to $\mathcal{T}_m$ and $\mathcal{T}_{mm}$. We can write the Lagrange of the optimization in (1) as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{L};\mathbf{N},\mathbf{R}) = &-\langle \mathbf{A},\mathbf{L}\rangle + \lambda\|\mathbf{L}\|_* + \langle \Gamma^+, \mathbf{L}^* - \mathbf{1}\mathbf{1}^T\rangle \\ &- \langle \Gamma^-, \mathbf{L}^* + \mathbf{1}\mathbf{1}^T\rangle.\end{aligned}$$

Any optimal solution $\mathbf{L}^*$ has to satisfy KKT conditions. Therefore, subgradient should be 0.

$$\lambda\partial\|\mathbf{L}^*\|_* - \mathbf{A} + \Gamma^+ - \Gamma^- = 0, \tag{3}$$

where Γ^+ and Γ^- are optimal dual variables. From complementary slackness conditions, we can write $\langle \Gamma^+, \mathbf{L}^* - \mathbf{1}\mathbf{1}^T\rangle = 0$ and $\langle \Gamma^-, \mathbf{L}^* + \mathbf{1}\mathbf{1}^T\rangle = 0$. Then, we have $\Gamma^+_{\{\mathbf{L}^*=1\}} \geq 0, \Gamma^+_{\{\mathbf{L}^*=-1\}} = 0, \Gamma^-_{\{\mathbf{L}^*=-1\}} \geq 0$, and $\Gamma^-_{\{\mathbf{L}^*=1\}} = 0$. Suppose that $\mathbf{L}^* = \mathbf{USV}^T$. We can write the subgradient of $\|\mathbf{L}\|_*$ as $\mathbf{UV}^T + \mathbf{W}$, where $\mathbf{W} \in \mathcal{M}_{\mathbf{UV}^T}\{\mathbf{X} : \mathbf{U}^T\mathbf{X} = \mathbf{XV} = 0, \|\mathbf{X}\| \leq 1\}$. We have 2 different clusters in the population. Suppose R_k is the set of $\{i,j\}$ pairs corresponding to the population in cluster k. C_1 and C_2 are sets of pairs corresponding to the matched and mismatched queries respectively. Suppose $\mu_{ab} = \sqrt{|R_a \cap C_b|}$ for all a, b. Then, we provide following Lemma.

Lemma 1. *We note that*

$$(\mathbf{UV}^T)_{i,j} = \frac{1}{\omega}\begin{cases}\mu_{11}^{-1}(\mu_{11}+\mu_{22}), \ \textit{if} \ (i,j) \in R_1 \cap C_1 \\ \mu_{21}^{-1}(\mu_{12}+\mu_{21}), \ \textit{if} \ (i,j) \in R_2 \cap C_1 \\ \mu_{12}^{-1}(\mu_{12}+\mu_{21}), \ \textit{if} \ (i,j) \in R_1 \cap C_2 \\ -\mu_{22}^{-1}(\mu_{11}+\mu_{22}), \ \textit{if} \ (i,j) \in R_2 \cap C_2.\end{cases}$$

where $\omega = \sqrt{(\mu_{11}+\mu_{22})^2 + (\mu_{12}+\mu_{21})^2}$.

Now, we can insert the subgradient into (3) and write

$$\lambda\mathbf{UV}^T + \lambda\mathbf{W} - \mathbf{A} + (\Gamma^+ - \Gamma^-) = 0. \tag{4}$$

Given that $\mathbf{W}^T\mathbf{U} = 0$ from the definition of the subgradient, we conclude that $\mathbf{W}^T\mathbf{L} = \mathbf{W}^T\mathbf{U\Sigma V}^T = 0$. Similarly, $\mathbf{LW}^T = \mathbf{U\Sigma V}^T\mathbf{W}^T = 0$ since $\mathbf{V}^T\mathbf{W}^T = 0$. Consider the entries in $R_1 \cap C_1$ with (4) and write their summation as:

$$\begin{aligned}&\text{sum}((\lambda\mathbf{UV}^T)_{\{R_1\cap C_1\}}) + \underbrace{\text{sum}(\lambda\mathbf{W}_{\{R_1\cap C_1\}})}_{0} \\ &- \underbrace{\text{sum}(\mathbf{A}_{\{R_1\cap C_1\}})}_{r(1-2q)} + \underbrace{\text{sum}((\Gamma^+ - \Gamma^-)_{\{R_1\cap C_1\}})}_{\geq 0} = 0\end{aligned}$$

Note that $\text{sum}(\mathbf{A}_{\{R_1\cap C_1\}}) = m_1|\mathcal{T}_m|(1-(1-r)-2rq)$ with probability $1-\exp(-\Omega(m_1|\mathcal{T}_m|))$. When $\lambda(\mu_{11}+\mu_{22})/\omega\mu_{11} - r(1-2q) > 0$, the equation (4) fails. We get similar bounds for the entries corresponding to $R_1 \cap C_2, R_2 \cap C_1$ and $R_2 \cap C_2$. Therefore, the optimization in (1) cannot recover $\mathbf{L}^*$ when $\lambda > r(1-2q)\frac{\omega}{1+\sqrt{\frac{\max(|\mathcal{T}_m|,|\mathcal{T}_{mm}|)\max(m_1,m_2)}{\min(|\mathcal{T}_m|,|\mathcal{T}_{mm}|)\min(m_1,m_2)}}}$. For $m_1 = m_2$, this corresponds to $\lambda > r(1-2q)\sqrt{m * \min(|\mathcal{T}_m|, |\mathcal{T}_{mm}|)}$.

Consider entries corresponding to the set $\mathbf{A}_{-1} \cap \{R_1 \cap C_1\}$. From (4),

$$\begin{aligned}&\mathbf{A}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}} + (\Gamma^- - \Gamma^+)_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}} \\ &= \lambda\mathbf{W}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}} + \lambda(\mathbf{UV}^T)_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}}.\end{aligned}$$

$\mathbf{A}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}} \leq 0$ and $(\Gamma^- - \Gamma^+)_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}} \leq 0$.

Therefore,

$$\|\mathbf{A}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}}\|_F^2$$

$$
\begin{aligned}
&\leq \|\lambda \mathbf{W}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}} + \lambda(\mathbf{U}\mathbf{V}^T)_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}}\|_F^2 \\
&\overset{a}{\leq} \lambda^2 \left(\|\mathbf{W}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}}\|_F^2 + \|(\mathbf{U}\mathbf{V}^T)_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}}\|_F^2\right) \\
&\leq \lambda^2 \|\mathbf{W}\|_F^2 + \lambda^2 \|(\mathbf{U}\mathbf{V}^T)_{\{R_1\cap C_1\}}\|_F^2 \\
&\overset{b}{\leq} \lambda^2 \min(m, |\mathcal{T}|)\, \|\mathbf{W}\|_2^2 + \lambda^2 \|(\mathbf{U}\mathbf{V}^T)_{\{R_1\cap C_1\}}\|_F^2 \\
&\overset{c}{\leq} \lambda^2 (\min(m, |\mathcal{T}|) + 1).
\end{aligned}
$$

Here, (a) follows from triangle inequality and (b) is due to the relation between 2-norm and Frobenius norm of a matrix. Recall that $(\mathbf{U}\mathbf{V}^T)_{i,j} = (\mu_{11}+\mu_{22})/(\omega\mu_{11})$ for $(i,j) \in R_1 \cap C_1$ and $|R_1 \cap C_1| = \mu_{11}^2 = m_1|\mathcal{T}_m|$. Therefore, we have $\|(\mathbf{U}\mathbf{V}^T)_{\{R_1\cap C_1\}}\|_F^2 = (\mu_{11}+\mu_{22})^2/(\omega)^2$. Then, (c) follows from the definition of $\mathcal{M}_{\mathbf{U}\mathbf{V}^T}$, where $\mathbf{W} \in \mathcal{M}_{\mathbf{U}\mathbf{V}^T}$, and the fact that $(\mu_{11}+\mu_{22}) \leq \omega$.

We recall that each entry of $\mathbf{A}$ is -1 within the entries corresponding to $R_1 \cap C_1$ with probability rq. Therefore, $\|\mathbf{A}_{\{\mathbf{A}_{-1}\cap\{R_1\cap C_1\}\}}\|_F^2 = m_1|\mathcal{T}_m|rq$ with probability at least $1-\exp(-\Omega(m_1|\mathcal{T}_m|))$. From (c), we can write $m_1|\mathcal{T}_m|rq \leq \lambda^2(\min(m,|\mathcal{T}|)+1)$. We get similar inequalities using entries corresponding to $\{\mathbf{A}_{-1} \cap R_2 \cap C_1\}$, $\{\mathbf{A}_{-1} \cap R_1 \cap C_2\}$ and $\{\mathbf{A}_1 \cap R_2 \cap C_2\}$. Therefore, we conclude that $\mathbf{L}^*$ cannot be an optimal solution to the optimization in (1), when

$$
\lambda < \sqrt{\frac{\max(m_1, m_2)\max(|\mathcal{T}_m|, |\mathcal{T}_{mm}|)\, rq}{\min(m, |\mathcal{T}|) + 1}}.
$$

Remark 4. *When $m_1 = m_2$, we have $\|(\mathbf{U}\mathbf{V}^T)_{\{R_1\cap C_1\}}\|_F^2 = 0.5$. Therefore, we find that $\mathbf{L}^*$ cannot be an optimal solution to the optimization in* (1)*, when*

$$
\lambda < \sqrt{\frac{m\max(|\mathcal{T}_m|, |\mathcal{T}_{mm}|)\, rq}{2\min(m, |\mathcal{T}|) + 1}}.
$$

B. Proof of Theorem III.1 (Success)

We want to show that $\langle \mathbf{A}, \mathbf{L}^* - \mathbf{L}\rangle + \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) > 0$, for all feasible solutions $\mathbf{L}$. We suppose that $\mathbf{L}^* = \mathbf{U}\mathbf{S}\mathbf{V}^T$ as above and assume 2 different clusters exist in the population. Therefore, we have $(\mathbf{U}\mathbf{U}^T)_{i,j} = \frac{1}{K_{p_i}}$ if $p_i = p_j$ and $(\mathbf{U}\mathbf{U}^T)_{i,j} = 0$ otherwise, and $(\mathbf{V}\mathbf{V}^T)_{i,j} = \frac{1}{L_{q_i}}$ if $q_i = q_j$ and $(\mathbf{V}\mathbf{V}^T)_{i,j} = 0$ otherwise. We can also bound the each entry of $\mathbf{U}\mathbf{V}^T$, simply from Lemma 1:

$$
\|\mathbf{U}\mathbf{V}^T\|_\infty \leq \frac{\sqrt{m_0\tau_0} + \sqrt{\mathcal{M}_0\mathcal{T}_0}}{\omega\sqrt{m_0\tau_0}}. \tag{5}
$$

Note that, for $m_1 = m_2$, we have $\|\mathbf{U}\mathbf{V}^T\|_\infty \leq \frac{1}{\sqrt{m\tau_0}}$. Recall that the subgradient of $\|\mathbf{L}\|_*$ is $\mathbf{U}\mathbf{V}^T + \mathbf{W}$, where $\mathbf{W} \in \mathcal{M}_{\mathbf{U}\mathbf{V}^T}\{\mathbf{X} : \mathbf{U}^T\mathbf{X} = \mathbf{X}\mathbf{V} = 0, \|\mathbf{X}\| \leq 1\}$. We also define following projections:

$$
\begin{aligned}
&\mathcal{P}_L\mathbf{S} = \mathcal{P}_\mathbf{U}\mathbf{S} + \mathbf{S}\mathcal{P}_\mathbf{V} - \mathcal{P}_\mathbf{U}\mathbf{S}\mathcal{P}_\mathbf{V} \\
&\mathcal{P}_{L^\perp}\mathbf{S} = \mathbf{S} - \mathcal{P}_L\mathbf{S},
\end{aligned} \tag{6}
$$

for a matrix $\mathbf{S}$ with correct dimensions, where $\mathcal{P}_\mathbf{U} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{P}_\mathbf{V} = \mathbf{V}\mathbf{V}^T$. It is easy to show that $\mathcal{P}_{L^\perp}\mathbf{X} \in \mathcal{M}_{\mathbf{U}\mathbf{V}^T}$. Then, we note that $\mathbf{U}\mathbf{V}^T + \frac{1}{\lambda}\mathcal{P}_{L^\perp}\mathbf{X}$ is a subgradient of $\|\mathbf{L}^*\|$ for any $\|\mathbf{X}\| \leq \lambda$. Therefore, we can write

$$
\lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) \geq \langle \lambda\mathbf{U}\mathbf{V}^T + \mathcal{P}_{L^\perp}\mathbf{X}, \mathbf{L} - \mathbf{L}^*\rangle
$$

for any feasible $\mathbf{L}$. We can reorganize this expression using (6) and write it in the following form:

$$
\lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) \geq \langle \lambda\mathbf{U}\mathbf{V}^T + \mathbf{X} - \mathcal{P}_L\mathbf{X}, \mathbf{L} - \mathbf{L}^*\rangle. \tag{7}
$$

Recall that we want to show that

$$
\langle \mathbf{A}, \mathbf{L}^* - \mathbf{L}\rangle + \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) > 0
$$

holds with high probability for all feasible solutions $\mathbf{L}$.

$$
\begin{aligned}
&\langle \mathbf{A}, \mathbf{L}^* - \mathbf{L}\rangle + \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) \\
&= \langle \mathbb{E}(\mathbf{A}), \mathbf{L}^* - \mathbf{L}\rangle + \langle \mathbf{A} - \mathbb{E}(\mathbf{A}), \mathbf{L}^* - \mathbf{L}\rangle + \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) \\
&\overset{a}{\geq} r(1-2q)\|\mathbf{L} - \mathbf{L}^*\|_1 + \langle \mathbf{A} - \mathbb{E}(\mathbf{A}), \mathbf{L}^* - \mathbf{L}\rangle \\
&+ \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^*\|_*) \\
&\overset{b}{\geq} r(1-2q)\|\mathbf{L} - \mathbf{L}^*\|_1 + \langle \mathcal{P}_L(\mathbf{A} - \mathbb{E}(\mathbf{A})) - \lambda\mathbf{U}\mathbf{V}^T, \mathbf{L}^* - \mathbf{L}\rangle \\
&\overset{c}{\geq} (r(1-2q) - \lambda\frac{\sqrt{m_0\tau_0} + \sqrt{\mathcal{M}_0\mathcal{T}_0}}{\omega\sqrt{m_0\tau_0}} - \epsilon'')\|\mathbf{L} - \mathbf{L}^*\|_1
\end{aligned}
$$

Here, (a) follows from the fact that $\mathbb{E}(\mathbf{A}_{i,j}) = r(1-2q)\mathbf{L}^*_{i,j}$. We can apply Lemma 2 for $\mathbf{S} = \mathbf{A}$ and (b) follows from (7) for $\lambda \geq (2\sigma + \epsilon')\sqrt{n}$. Lastly, (c) is from (5) and Theorem V.1. We set $r(1-2q) - \lambda\frac{\sqrt{m_0\tau_0}+\sqrt{\mathcal{M}_0\mathcal{T}_0}}{\omega\sqrt{m_0\tau_0}} > \epsilon''$ and find that $\mathbf{L}^*$ is the optimal solution, where ϵ'' is chosen to be a sufficiently small constant. In particular, we set $\epsilon'' = \frac{1}{2}\left(r(1-2q) - \lambda\frac{\sqrt{m_0\tau_0}+\sqrt{\mathcal{M}_0\mathcal{T}_0}}{\omega\sqrt{m_0\tau_0}}\right)$.

Lemma 2. *Let $\mathbf{S}$ be a random matrix with independent entries and $|\mathbf{S}_{i,j}| \leq 1$ for all i, j, and the variance of each entry is at most σ^2. Then, we can bound $\|\mathbf{S} - \mathbb{E}(\mathbf{S})\|$ as follows:*

$$
\|\mathbf{S} - \mathbb{E}(\mathbf{S})\| \leq (2\sigma + \epsilon')\sqrt{n}
$$

with probability $1 - \exp(-\Omega(n))$.

Proof: $\mathbf{S}$ is a random matrix with independent entries and note that $|\mathbf{S}_{i,j} - \mathbb{E}(\mathbf{S}_{i,j})| \leq 2$. Using standard results on random matrix theory (Theorem 1.4 in [35]), we can bound the operator norm of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ as $\|\mathbf{X}\| \leq (2\sigma+\epsilon')\sqrt{n}$, with probability $1-\exp(-\Omega(n))$. Now, we use the trick called dilations [30], [36] and consider $\mathbf{X}$ as $\mathbf{X} = [0\ \mathbf{S}; \mathbf{S}^T\ 0]$ and write that $\|\mathbf{S}\| = \|\mathbf{X}\| \leq (2\sigma + \epsilon')\sqrt{n}$, where $n = m + |\mathcal{T}|$.

Theorem V.1. *Let $\mathbf{A}$ be the observation matrix as described previously in Section III-A. We have that*

$$
\|\mathcal{P}_L\mathbf{Z}\|_\infty \leq \epsilon'',
$$

where $\mathbf{Z} = \mathbf{A} - \mathbb{E}(\mathbf{A})$, with probability at least $1 - 6m|\mathcal{T}|\exp(-\frac{2}{9}\epsilon''^2\min\{m_0, \tau_0\})$,

Proof: We first characterize the distribution of entries of $\mathbf{Z}$. For this, we define following random variable Y and call its distribution $\Delta(r, q)$.

$$Y = \begin{cases} 1 - r(1-2q), \ w.p. \ r(1-q) \\ -1 - r(1-2q), \ w.p. \ rq \\ -r(1-2q), \ w.p \ 1-r. \end{cases}$$

Here, we note that $\mathbf{Z}_{i,j}\text{sign}(\mathbf{L}^*) \sim \Delta(r, q)$. Therefore, each entry of $\mathbf{P_U Z}$ will be the average of m_i i.i.d mean zero random variables. Now, we can write following expressions using Hoeffding's Inequality (see (8) for the restatement of Hoeffding's Inequality):

$$\begin{aligned} &\mathbb{P}[|(\mathbf{P_U Z})_{i,j}| \geq \epsilon''] \leq 2\exp -2\epsilon''^2 m_0 \\ &\mathbb{P}[|(\mathbf{Z P_V})_{i,j}| \geq \epsilon''] \leq 2\exp -2\epsilon''^2 \tau_0 \\ &\mathbb{P}[|(\mathbf{P_U Z P_V})_{i,j}| \geq \epsilon''] \leq 2\exp -2\epsilon''^2 m_0\tau_0 \end{aligned}$$

Note that $\|\mathcal{P}_L \mathbf{Z}\|_\infty \leq \|\mathbf{P_U Z}\|_\infty + \|\mathbf{Z P_V}\|_\infty + \|\mathbf{P_U} S \mathbf{P_V}\|_\infty$. Then, apply the union bound together with Hoeffding's Inequality based bounds derived above to complete the proof.

Hoeffding's Inequality [37]: Let $X_1, X_2 \ldots, X_N$ be independent random variables such that $a_i \leq X_i \leq b_i$ and let $S_N := \sum_{i=1}^N X_i$, then for all $t > 0$,

$$\Pr(|S_N - \mathbb{E}(S_N)| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right). \quad (8)$$

Lemma 3. *Let $q < \frac{1}{2}$ and $\mathbf{A}$ represents the observation matrix generated by the responses to triplet queries in $\mathcal{T}$ according to the model defined in Section III-A. If $\lambda \geq (1+\delta)\Sigma_{succ}$ and $D = r(1-2q)\omega / \left(1 + \sqrt{\mathcal{T}_0 \mathcal{M}_0/(m_0\tau_0)}\right) \geq \lambda$, then $\mathbf{L}^*$ is the unique optimal solution to the optimization in (1) with probability $1 - \exp(-\Omega(n)) - 6m|\mathcal{S}|\exp(-\Omega(\min\{m_0, \tau_0\}))$.*

Proof: It follows from applying union bound for (c) and Lemma 2.

C. Proof for Outliers

We want to show that$\langle \mathbf{A}, \mathbf{L}^c - \mathbf{L}\rangle + \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^c\|_*) > 0$ holds with high probability for all feasible solutions $\mathbf{L}$, where entries of $\mathbf{A}$ are responses to triplet comparisons. Note that we have, for any $\|\mathbf{X}\| \leq \lambda$,

$$\langle \mathbf{A}, \mathbf{L}^c - \mathbf{L}\rangle + \langle \lambda \mathbf{U}\mathbf{V}^T + \mathbf{X} - \mathcal{P}_T\mathbf{X}, \mathbf{L} - \mathbf{L}^c\rangle > 0. \quad (9)$$

For any given matrix $\mathbf{L} \in \mathbb{R}^{(m+1)\times|\mathcal{T}|}$, we can write $\mathbf{L}$ as $\mathbf{L} = [\mathbf{L}^{non}; \mathbf{L}^o]$, where $\mathbf{L}^{non} \in \mathbb{R}^{m \times |\mathcal{T}|}$ and $(\mathbf{L}^o)^T \in \mathbb{R}^{|\mathcal{T}|}$ are entries corresponding to positions of responses from the cluster and the outlier respectively. Similarly, note that $\mathbf{A} = [\mathbf{A}^{non}; \mathbf{A}^o]$. Therefore, we have

$$\begin{aligned} &\langle [\mathbf{A}^{non}; \mathbf{A}^o], \mathbf{L}^c - \mathbf{L}\rangle + \lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^c\|_*) \\ =&\langle [\mathbb{E}(\mathbf{A}^{non}); \mathbb{E}(\mathbf{A}^o)], \mathbf{L}^c - \mathbf{L}\rangle + \langle \mathbf{A} - \mathbb{E}(\mathbf{A}), \mathbf{L}^c - \mathbf{L}\rangle \\ &+\lambda(\|\mathbf{L}\|_* - \|\mathbf{L}^c\|_*) \\ \overset{a}{\geq}&\langle \mathbb{E}(\mathbf{A}^{non}), (\mathbf{L}^c - \mathbf{L})^{non}\rangle + \langle \mathbb{E}(\mathbf{A}^o), (\mathbf{L}^c - \mathbf{L})^o\rangle \\ &+\langle \mathcal{P}_L(\mathbf{A} - \mathbb{E}(\mathbf{A})) - \lambda\mathbf{U}\mathbf{V}^T, \mathbf{L}^c - \mathbf{L}\rangle \\ \overset{b}{\geq}&(r(1-2q) - \lambda/\sqrt{(m+1)|\mathcal{T}|} - \epsilon'')\|(\mathbf{L} - \mathbf{L}^c)^{non}\|_1 \\ &+(r(1-2q)(1-2d) - \lambda/\sqrt{(m+1)|\mathcal{T}|} - \epsilon'')\|(\mathbf{L} - \mathbf{L}^c)^o\|_1 \end{aligned}$$

Since $\|\mathbf{A} - \mathbb{E}(\mathbf{A})\| \leq (2\overline{\sigma} + \epsilon')\sqrt{n}$ from Lemma 2 for $\mathbf{S} = \mathbf{A}$, (a) follows from (9). We have $\mathbb{E}(\mathbf{A}^{non}_{i,j}) = r(1-2q)\mathbf{L}^{non^*}_{i,j}$ and $\mathbb{E}(\mathbf{A}^o_{i,j}) = r(1-2q)(1-2d)\mathbf{L}^{o^*}_{i,j}$. Note that $\mathbf{U}\mathbf{V}^T = 1/\sqrt{(m+1)|\mathcal{T}|}$ for single cluster. Then, (b) follows from Proposition 2.

Proposition 2. *Let $\mathbf{A}^{non}$ and $\mathbf{A}^o$ be as described previously above. We have that*

$$\|\mathcal{P}_L \mathbf{Z}\|_\infty \leq \epsilon'',$$

where $\mathbf{Z} = [\mathbf{A}^{non}; \mathbf{A}^o] - [\mathbb{E}(\mathbf{A}^{non}); \mathbb{E}(\mathbf{A}^o)]$, with probability at least $1 - 6(m+1)|\mathcal{T}|\exp(-\frac{2}{9}\epsilon''^2 \min\{m+1, \mathcal{T}\})$,

Proof: We first characterize the distribution of entries of $\mathbf{Z}$. For this, we define following random variable Y' and call its distribution $\Delta(r, q_1, q_2)$.

$$Y' = \begin{cases} -1 - r(1-2q_1)(1-2q_2), \ w.p. \ r(1-q_1)q_2 + rq_1(1-q_2) \\ 1 - r(1-2q_1)(1-2q_2), \ w.p. \ rq_1q_2 + r(1-q_1)(1-q_2) \\ -r(1-2q_1)(1-2q_2), \ w.p. \ 1-r. \end{cases}$$

Note that $\mathbf{Z}_{i,j}\text{sign}(\mathbf{L}^*) \sim \Delta(r, q, d)$ for entries of the outlier and $\mathbf{Z}_{i,j}\text{sign}(\mathbf{L}^*) \sim \Delta(r, q, 0)$ for the rest. Therefore, each entry of $\mathbf{P_U Z}$ will be the average of $m+1$ mean zero random variables. Now, we can write following expressions using Hoeffding's Inequality (see (8) for the restatement of Hoeffding's Inequality):

$$\begin{aligned} &\mathbb{P}[|(\mathbf{P_U Z})_{i,j}| \geq \epsilon''] \leq 2\exp -2\epsilon''^2(m+1)) \\ &\mathbb{P}[|(\mathbf{Z P_V})_{i,j}| \geq \epsilon''] \leq 2\exp -2\epsilon''^2|\mathcal{T}| \\ &\mathbb{P}[|(\mathbf{P_U Z P_V})_{i,j}| \geq \epsilon''] \leq 2\exp -2\epsilon''^2(m+1)|\mathcal{T}|, \end{aligned}$$

The rest follows similar lines with the proof of Theorem V.1.

VI. Conclusions and Future Work

We study the problem of clustering populations based on shared metric within subgroups by using answers to triplet comparison queries. We provide analysis of convex optimization approach showing success and failure conditions, and behavior of outliers with a simple setting. We leave the analysis of populations with multiple metrics and better understanding of outlier behavior as future work.

VII. Acknowledgment

We thank Zhi Wang and Yi Chen for their valuable feedback.

References

[1] H. Wu, Q. Zhou, R. Nie, and J. Cao, "Effective metric learning with co-occurrence embedding for collaborative recommendations," *Neural Networks*, vol. 124, pp. 308–318, 2020.

[2] X. Li and Y. Tang, "A social recommendation based on metric learning and network embedding," in *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 2020, pp. 55–60.

[3] Q. Liu, W. Li, Z. Chen, and B. Hua, "Deep metric learning for image retrieval in smart city development," *Sustainable Cities and Society*, vol. 73, p. 103067, 2021.

[4] D. J. HOPKINS and H. NOEL, "Trump and the shifting meaning of "conservative": Using activists' pairwise comparisons to measure politicians' perceived ideologies," *American Political Science Review*, vol. 116, no. 3, p. 1133–1140, 2022.

[5] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[6] W. G. Jacoby and D. J. Ciuk, "Multidimensional scaling: An introduction," *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, pp. 375–412, 2018.

[7] R. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. I," *Psychometrika*, vol. 27, no. 2, pp. 125–140, June 1962.

[8] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. ii," *Psychometrika*, vol. 27, pp. 219–246, 1962.

[9] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie, "Generalized non-metric multidimensional scaling," in *Artificial Intelligence and Statistics*. PMLR, 2007, pp. 11–18.

[10] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, vol. 10, no. 2, 2009.

[11] B. Mason, L. Jain, and R. Nowak, "Learning low-dimensional metrics," *Advances in neural information processing systems*, vol. 30, 2017.

[12] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "A new kernelization framework for mahalanobis distance learning algorithms," *Neurocomputing*, vol. 73, no. 10-12, pp. 1570–1579, 2010.

[13] A. Xu and M. Davenport, "Simultaneous preference and metric learning from paired comparisons," *Advances in Neural Information Processing Systems*, vol. 33, pp. 454–465, 2020.

[14] G. Canal, B. Mason, R. K. Vinayak, and R. Nowak, "One for all: Simultaneous metric and preference learning over multiple users." in *NeurIPS*, 2022.

[15] Z. Wang, G. So, and R. K. Vinayak, "Metric learning from limited pairwise preference comparisons," in *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[16] A. Bellet, A. Habrard, and M. Sebban, *Metric learning*. Morgan & Claypool Publishers, 2015.

[17] B. Kulis, A. C. Surendran, and J. C. Platt, "Fast low-rank semidefinite programming for embedding and clustering," in *Artificial Intelligence and Statistics*. PMLR, 2007, pp. 235–242.

[18] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit," *Advances in neural information processing systems*, vol. 23, 2010.

[19] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu, "Clustering partially observed graphs via convex optimization," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2213–2238, 2014.

[20] R. K. Vinayak, S. Oymak, and B. Hassibi, "Graph clustering with missing data: Convex algorithms and analysis," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[21] B. P. Ames and S. A. Vavasis, "Nuclear norm minimization for the planted clique and biclique problems," *Mathematical programming*, vol. 129, no. 1, pp. 69–89, 2011.

[22] B. P. Ames and S. Vavasis, "Convex optimization for the planted k-disjoint-clique problem," *Mathematical Programming*, vol. 1, no. 143, pp. 299–337, 2014.

[23] R. K. Vinayak and B. Hassibi, "Similarity clustering in the presence of outliers: Exact recovery via convex program," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 91–95.

[24] R. K. Vinayak, T. Zrnic, and B. Hassibi, "Tensor-based crowdsourced clustering via triangle queries," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2322–2326.

[25] D. G. Mixon, S. Villar, and R. Ward, "Clustering subgaussian mixtures by semidefinite programming," *Information and Inference: A Journal of the IMA*, vol. 6, no. 4, pp. 389–415, 2017.

[26] S. H. Lim, Y. Chen, and H. Xu, "A convex optimization framework for bi-clustering," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1679–1688.

[27] MATLAB, *version R2023b*. Natick, Massachusetts: The MathWorks Inc., 2023.

[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

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

[30] J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying, "Jointly clustering rows and columns of binary matrices: algorithms and trade-offs," *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, pp. 29–41, 2014.

[31] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.

[32] R. K. Vinayak and B. Hassibi, "Crowdsourced clustering: Querying edges vs triangles," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[33] E. J. Candes and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 227–254, 2006.

[34] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[35] V. H. Vu, "Spectral norm of random matrices," *Combinatorica*, vol. 27, no. 6, pp. 721–736, 2007.

[36] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, pp. 389–434, 2012.

[37] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.