Nearly Tight Bounds on Testing of Metric Properties

Yiqiao Bao* Sampath Kannan† Erik Waingarten‡

Abstract

Given a non-negative $n \times n$ matrix viewed as a set of distances between n points, we consider the property testing problem of deciding if it is a metric. We also consider the same problem for two special classes of metrics — tree metrics and ultrametrics. For general metrics, our paper is the first to consider these questions. We prove an upper bound of $O(n^{2/3}/\varepsilon^{4/3})$ on the query complexity for this problem. Our algorithm is simple, but the analysis requires great care in bounding the variance on the number of violating triangles in a sample. When ε is a slowly decreasing function of n (rather than a constant, as is standard), we prove a lower bound of matching dependence on n of $\Omega(n^{2/3})$, ruling out any property testers with $o(n^{2/3})$ query complexity unless their dependence on $1/\varepsilon$ is super-polynomial.

Next, we turn to tree metrics and ultrametrics. While there were known upper and lower bounds, we considerably improve these bounds showing essentially tight bounds of $\tilde{O}(1/\varepsilon)$ on the sample complexity. We also show a lower bound of $\Omega(1/\varepsilon^{4/3})$ on the query complexity. Our upper bounds are derived by doing a more careful analysis of a natural, simple algorithm. For the lower bounds, we construct distributions on NO instances, where it is hard to find a witness showing that these are not ultrametrics.

*University of Pennsylvania.

†Simons Institute, UC Berkeley (on leave from the University of Pennsylvania).

‡University of Pennsylvania.

1 Introduction

Finite metric spaces is a rich topic at the intersection of combinatorics, algorithms, and geometry (see [24, 25, 21], among many other works, for general overviews). In addition to their intrinsic interest, a strong motivation for studying metric spaces from the theoretical computer science perspective is that a metric space, or metric for short, defines a quantitative measure of dissimilarity. Once a metric over the objects, or points, is defined, one can design algorithms to find the most similar objects, partition the objects, and cluster the objects. Formally, a metric space on the ground set $[n]$ is specified by a dissimilarity function $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ which satisfies three axioms (i) the function is symmetric, so $d(i,j) = d(j,i)$, (ii) it is non-negative with $d(i,j) = 0$ if and only if $i = j$, and (iii) it satisfies the triangle inequality: A function $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ satisfies the triangle inequality if and only if $\forall\, i,j,k \in [n]$, $d(i,k) \leq d(i,j) + d(j,k)$.

Property testing [7, 28] is a natural algorithmic framework for studying whether the given matrix constitutes a metric. Here, we consider randomized ε-testing algorithms that receive black-box query access to an unknown and arbitrary $n \times n$ matrix M. The goal is to make as few queries as possible to the entries of the matrix while deciding whether its entries encode a function d which is a metric or is ε-far from a metric (meaning that any metric $d'\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ differs on at least ε-fraction of the n^2 inputs). We remark that "ε-farness" here, which is the most natural from a property testing perspective, provides an ℓ_0-distance guarantee, which has seen a recent surge of interest in approximation algorithms [20, 11, 12, 22]. Our work gives upper and lower bounds on various metric testing problems and can be used as a *sublinear*-time preprocessing step, or quick "sanity check," to the more expensive polynomial-time approximation algorithms.

As we detail in Subsection 1.2, there are a number of prior works that study finite metrics from a property testing perspective. Despite the broad interest, this work is the first to address what is arguably the most basic metric testing question—can finite metrics be efficiently tested? Our main algorithmic result is a non-adaptive metric testing algorithm that makes $O(n^{2/3}/\varepsilon^{4/3})$ queries. As is common in property testing, the algorithm is straightforward and proceeds by executing two checks:

1. It randomly selects a subset of $O(1/\varepsilon)$ points from $[n]$ and a subset of $O(n^{2/3}/\varepsilon^{1/3})$ pairs from $[n] \times [n]$, and checks whether the triangle formed by a point and pair violates the triangle inequality.
2. It samples a random set of $O(n^{1/3}/\varepsilon^{2/3})$ many points and checks whether there are three points in this set that violate the triangle inequality.

Every metric space will trivially pass the above tests since there are no violations of the triangle inequality. The interesting aspect of our analysis, which constitutes the majority of the technical work, establishes that if no violations of the triangle inequality are observed with probability 1/3, the function $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ is (almost) a metric. Furthermore, we show that a lower bound establishing that (unless the dependence on ε becomes super-polynomial) a dependence on $n^{2/3}$ is needed for non-adaptive algorithms with one-sided error.

Next, we turn our attention to the specific classes of tree metrics and ultrametrics. These were first studied by Parnas and Ron [27] from the property testing perspective and also received recent attention from the above-mentioned works on approximation algorithms. Tree metrics (also known as additive metrics) and ultrametrics are specific sub-classes of metric spaces that may be represented with positively-weighted trees. Points in the metric correspond to nodes in a tree and distances are measured by the lengths of paths (see Section 2 for their formal definition). In their work, [27] give an algorithm which samples $O(1/\varepsilon^3)$ random points and queries all pairwise function evaluations (using $O(1/\varepsilon^6)$ queries) to ε-test tree metrics and ultrametrics. We improve upon their dependence in the following way: we show that, for both tree metrics and ultrametrics, it suffices to take $\tilde{O}(1/\varepsilon)$ samples and query all $\tilde{O}(1/\varepsilon^2)$ pairwise evaluations. Furthermore, we show a sample complexity lower bound of $\Omega(1/\varepsilon)$ (i.e., testing algorithms must interact with at least these many points), and a $\Omega(1/\varepsilon^{4/3})$ lower bound on the query complexity. Hence, our algorithms are sample-optimal (up to polylog$(1/\varepsilon)$ factors), and the $\Omega(1/\varepsilon^{4/3})$ query lower bound rules out what is often the "best case" scenario in property testing, which is an $\Theta(1/\varepsilon)$-query tester.

1.1 Our Contributions

Testing Metrics. As mentioned above, our work is the first to prove upper and lower bounds on the query complexity of property testing of metric spaces. We provide three (non-adaptive) testing algorithms for testing metrics, tree metrics, and ultrametrics, along with lower bounds which show that our results are (in certain regimes) the best possible. We begin by stating our main algorithmic theorem, which gives an algorithm for testing metrics using $O(n^{2/3}/\varepsilon^{4/3})$ queries.

THEOREM 1.1. (TESTING METRICS—UPPER BOUND) *For any large enough $n \in \mathbb{N}$ and any $\varepsilon \in (0,1)$, there exists a randomized algorithm that receives query access to an unknown matrix $M \in \mathbb{R}^{n\times n}$ and makes $O(n^{2/3}/\varepsilon^{4/3})$ queries with the following guarantee:*

- *If M defines a metric space on $[n]$, the algorithm outputs "accept" with probability 1.*
- *If M is ε-far from being being a metric, then the algorithm outputs "reject" with probability at least $2/3$.*

Furthermore, the algorithm is non-adaptive (i.e., queries made do not depend on answers to prior queries).

The algorithm that proves Theorem 1.1 is especially appealing from a structural perspective due to its simplicity. By executing only two types of checks and only $O(n^{2/3}/\varepsilon^{4/3})$ queries, the algorithm guarantees that any matrix that passes the test is ε-close to a metric with high probability. Our second result shows that the dependence on the number of points must be $n^{2/3}$ unless one is willing to incur super-polynomial factors in ε.

THEOREM 1.2. (TESTING METRICS—LOWER BOUND) *For any large enough $n \in \mathbb{N}$, let $\varepsilon = n^{-\nu(n)}$ where $\nu(n) = (\log\log\log n + 4)/\log\log n$. Any non-adaptive, one-sided algorithm which can ε-test whether a matrix $M \in \mathbb{R}^{n\times n}$ encodes a metric must make $\Omega(n^{2/3+2\nu(n)/3})$ queries.*

The lower bound implies that, unless the dependence on ε blows up and becomes super-polynomial (in particular, at least $(1/\varepsilon)^{\Omega(1/\nu(n))}$, then a dependence of $n^{2/3}$ is unavoidable. In other words, one cannot hope for an algorithm whose complexity is $O(n^{.65}/\varepsilon^c)$ for any fixed constant $c > 0$. The proof stems from a connection between testing metric spaces and triangle-freeness testing; we will construct matrices $M \in \mathbb{R}^{n\times n}$, which masquerade as metrics to low-query algorithms, by utilizing certain Behrend graphs that have previously appeared for proving lower bounds on testing triangle-freeness in graphs [2, 1]. That work, in Section 6, shows a reduction from one-sided triangle-freeness testers to two-sided triangle freeness testers, which we believe applies mutatis mutandis, so $\Omega(n^{2/3+2\nu(n)/\varepsilon})$ queries are necessary for two-sided non-adaptive metric testing algorithms as well.

We remark that the connection between triangle-freeness testing and metric testing is useful for lower bounds, but not for our upper bounds. As we will discuss, our proof of Theorem 1.1 relies on certain properties of metrics that do not have a direct analog in graphs.

Tree Metrics and Ultrametrics. We now state the main results for testing tree metrics and ultrametrics. As mentioned, our works improve on the algorithms of [27] by improving the sample complexity from $O(1/\varepsilon^3)$ to $\tilde{O}(1/\varepsilon)$, and the query complexity from $O(1/\varepsilon^6)$ to $\tilde{O}(1/\varepsilon^2)$. Our lower bounds establish the following two aspects: (i) our algorithms are sample optimal, meaning any ε-testing must evaluate distances to $\Omega(1/\varepsilon)$ many points, and (ii) one cannot hope for the "best case" scenario of a $O(1/\varepsilon)$-query tester.

THEOREM 1.3. (TESTING TREE METRICS AND ULTRAMETRICS—UPPER BOUND) *For any large enough $n \in \mathbb{N}$ and any $\varepsilon > 0$, there exists a randomized algorithm that receives query access to an unknown matrix $M \in \mathbb{R}^{n\times n}$, using $\tilde{O}(1/\varepsilon)$ samples and $\tilde{O}(1/\varepsilon^2)$ queries, and has the following guarantee:*

- *If M defines a tree metric (or ultrametric) on $[n]$, the algorithm always outputs "accept."*
- *If M is ε-far from being a tree metric (or ultrametric) on $[n]$, the algorithm outputs "reject" with probability at least $2/3$.*

Furthermore, the algorithm is non-adaptive.

Our algorithm is exactly the same as the algorithm of [27], and the improvement lies solely in the analysis. The algorithm takes $\tilde{O}(1/\varepsilon)$ samples and queries all pairwise evaluations using $\tilde{O}(1/\varepsilon^2)$ queries. We show that, if there are no violations to tree metrics (or ultrametrics) in a sample of size $\tilde{O}(1/\varepsilon)$ with probability at least $1/3$, then M is ε-close to a tree metric (or ultrametric).

THEOREM 1.4. (TESTING TREE METRICS AND ULTRAMETRICS—LOWER BOUND) *For any large enough $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$, any non-adaptive algorithm which can ε-test whether a matrix $M \in \mathbb{R}^{n \times n}$ is a tree metric (or ultrametric) must use $\Omega(1/\varepsilon)$ samples and at least $\Omega(1/\varepsilon^{4/3})$ queries.*

1.2 Related Work There are a number of works on sublinear algorithms for metric spaces. The most relevant to this work is [27] who, among other results, gave property testing algorithms for tree metrics and ultrametrics using $O(1/\varepsilon^3)$ samples and $O(1/\varepsilon^6)$ queries. More generally viewing metrics as a property of $n \times n$ matrices, there has been a recent line-of-work on testing of matrix properties [19, 8, 4, 9, 5] (see, also Chapter 8 in the textbook [10]). The works of [23, 26] study property testing of points in a metric, where the algorithm receives query access to a distance oracle, and the proximity parameter ε is with respect to the number of points which must be modified; the properties of interest are dimensionality and embeddability into other metrics. Property testing by accessing vectors directly (for example, when the metric is Euclidean) has also been studied [17, 14], and more generally, there have been various sublinear algorithms in these settings [13, 3, 15, 16].

A recent line of work in approximation algorithms, under the name "metric violation distance," finds the (approximately) most similar metric (with respect to the ℓ_0-distance) in polynomial time [20, 11, 12, 22]. There, the ℓ_0-distance is the number of entries of $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ which are changed, coinciding with the proximity parameter ε studied in property testing. For example, the current best algorithm [11] runs in $O(n^3)$ time and produces a metric space which is a $O(\log n)$-multiplicative factor further than the closest metric. In contrast, property testing is much more efficient but only provides approximate decisions. Hence, our property testing algorithms (both for general metrics, as well as tree metrics or ultrametrics) could be used as a preprocessing step—if a property tester declares a supposed distance-matrix is already too far from a metric (or tree or ultrametric), there may be no use in more expensive approximation algorithms.

1.3 Technical Overview

Metric Testing Upper Bound. Our starting point is that, any matrix M which is ε-far from being a metric must contain $\Omega(\varepsilon n^2)$ triangles of three points $\{i,j,k\}$ whose pairwise distances violate the triangle inequality (Lemma 3.1). These so-called "violating triangles," denoted by the collection T, are evidence that M is not a metric, and our algorithm's goal is to query all three entries of at least one violating triangle if they exist. This already suggests an $O(n^{2/3})$-query tester (assume, for this overview, that $\varepsilon = \Omega(1)$) since the expected number of triangles from T contained within a sample $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s \sim [n]$ is

$$|T| \cdot \mathbf{Pr}\left[\text{a fixed triangle is among } s \text{ random points}\right] = \Omega(\varepsilon n^2) \cdot \Omega\left(\frac{s^3}{n^3}\right),$$

which is a large constant when the size sample s is $\Theta(n^{1/3})$, and leads to an $O(n^{2/3})$ query complexity. The challenge is upper bounding the variance of this random variable. For example, if all violating triangles are

incident on a single vertex or on $O(n)$ pairs of vertices, sampling these requires $\Omega(n)$ queries. Our analysis will rule out such cases by showing matrices M, with these triangle configurations, are close to a metric (even if there are many violating triangles). Furthermore, we show that as long as M is ε-far from being a metric, one of two things happens:

- Either there are $\Omega(\varepsilon n)$ points, each of which participates in $\Omega(n^{4/3})$ violating triangles, or
- There exists a structured set of $\Omega(\varepsilon n^2)$ violating triangles $\tilde{T}$ where each point participates in at most $O(n^{4/3})$ triangles, and each pair of points in at most $O(n^{2/3})$ triangles.

In the first case, we run our first algorithmic check: sample $O(1/\varepsilon)$ points and check the triangle inequality against $O(n^{2/3})$ random pairs (which guarantees one of the $\Omega(\varepsilon n)$ points and corresponding $\Omega(n^{4/3})$ pairs is sampled, see Lemma 3.2). Proving the second case is more challenging (Lemma 3.4), but once we do, the bounds on the number of participating triangles on points and pairs of points help us prove an adequate bound on the variance of the number of violating triangles "caught" by a random $O(n^{1/3})$-size sample (see Claim 3.1).

The second case, Lemma 3.4, proceeds in the following way. Suppose M is ε-far from being a metric and let T be the violating triangles. If $o(\varepsilon n)$ points participate in more than $\Omega(n^{4/3})$ violating triangles, we may "fix" all such points—we change all $O(n)$ entries incident to each of these points for a total of $o(\varepsilon n^2)$ entries (Lemma 3.5). After these changes, the new matrix M' is still $\varepsilon/2$-far from a metric, but each point now participates in $O(n^{4/3})$ violating triangles. We then address the pairs of points (corresponding to entries in M') that participate in too many, $\Omega(n^{2/3})$, violating triangles. We show that there are $o(\varepsilon n^2)$ such pairs since each point participates in few violating triangles (Lemma 3.6), but we may not be able to "fix" these entries—note that there are $n-2$ other points whose distances constrain that of the pair, and these constraints may be irreconcilable by changing only one entry.

Here is where the fact that metrics specify real-valued distances (as opposed to graphs, which only specify Boolean values) becomes useful. Suppose a pair of points $\{i, j\}$ participates in two violating triangles, with points k_1 and k_2 respectively, and further, one cannot change the distance between (i, j) and simultaneously "fix" the triangles $\{i, j, k_1\}$ and $\{i, j, k_2\}$. Then, we show that either $\{i, k_1, k_2\}$ or $\{j, k_1, k_2\}$ must also be a violating triangle (Claim 3.7). This implies that, since i and j each participates in $O(n^{4/3})$ violating triangles, there must exist a setting which fixes all but $O(n^{2/3})$ violating triangles incident on i, j (Lemma 3.8). After performing this modification, the resulting matrix M'' is still $\varepsilon/4$-far, and so must contain $\Omega(\varepsilon n^2)$ violating triangles and both bounds on the number of violating triangles on points and pairs of points hold. These violating triangles may not exist in M, since they may use modified entries, but we show how to map such violating triangles back to violating triangles that do exist in M. (The assumption was that the modified pairs of points participated in many violating triangles and so there will be many such distinct triangles to choose from.) This completes the analysis of the metric testing algorithm.

Metric Testing Lower Bound. The lower bound on metric testing comes from a connection to triangle-freeness testing in graphs. We base our lower bound on the construction in [1], who give tripartite graphs on $3n$ vertices and $\Theta(n^{2-\nu(n)})$ edge-disjoint triangles, so-called Behrend graphs. Our task, which we do in Claim 4.1, is to assign weights to edges (as well as non-edges) of the Behrend graph so that the $\Theta(n^{2-\nu(n)})$ triangles are the only violations to the triangle inequality. A non-adaptive, one-sided lower bound of $\Omega(n^{2/3+2\nu(n)/3})$ proceeds as follows. Fix q queries to entries of a $(3n) \times (3n)$ matrix, and let $\mathbf{M}$ denote the (random) Behrend-based matrix obtained after randomly permuting points. In order to observe a violation of the triangle inequality, there must be a triangle formed by the q queries which contains one of the $\Theta(n^{2-\nu(n)})$ edge-disjoint triangles of the Behrend graph. Among q queries over pairs, there can be at most $O(q^{3/2})$ triangles (because cliques of size $\Theta(\sqrt{q})$ maximize the number of triangles with q edges), and the probability that a fixed triangle of the Behrend graph is mapped to a fixed triangle among the queries under a random permutation is $O(1/n^3)$. Hence, the probability we observe at least one triangle is at most

$$O(q^{3/2}) \cdot O(n^{2-\nu(n)}) \cdot O(1/n^3) = O(q^{3/2}/n^{1+\nu(n)}),$$

which is $o(1)$ when q is $o(n^{2/3+2\nu(n)/3})$. This establishes the one-sided lower bound; to make the lower bound two-sided, there is a reduction from [1] which applies for graphs, and readily applies in our scenario as well.

Tree Metric and Ultrametric Testing Upper Bound. Testing tree metrics and ultrametrics is significantly more efficient, because tree metrics and ultrametrics are more constrained than (general) metrics, making violations easier to find. Consider ultrametrics, which satisfy the following strengthening of the triangle inequality: for any $i, j, k \in [n]$, if $d(i, j)$ is the maximum distance among the three pairs, then

$$d(i, j) = \max\{d(i, k), d(j, k)\}.$$

Our analysis mirrors that of [27] (which obtained a $O(1/\varepsilon^6)$ query complexity) and incorporates one change to obtain the query complexity $\tilde{O}(1/\varepsilon^2)$. Roughly speaking, a set of sampled points $\mathbf{S}$ which does not contain a violation amongst themselves imposes constraints on distances amongst the remaining points in $[n] \setminus \mathbf{S}$. For example, if j and k are two points in $[n] \setminus \mathbf{S}$, and some $i \in \mathbf{S}$ satisfies $M(i, j) \neq M(i, k)$, an ultrametric *must set* $M(j, k)$ to $\max\{M(i, j), M(i, k)\}$; a violation occurs when it does not. Hence, consider the partition of $[n] \setminus \mathbf{S}$ imposes by $\mathbf{S}$, where two points $j, k \in [n] \setminus \mathbf{S}$ belong to different parts if $M(i, j) \neq M(i, k)$ for some $i \in \mathbf{S}$ (Definition 5.2). As parts become smaller, there are more pairs (j, k) in different parts, which makes violations easier to find. [27] argues (using a loose argument which does not optimize ε-factors) that a batch of $O(1/\varepsilon)$ samples added to $\mathbf{S}$ decreases the number of pairs in the same part by $\Omega((\varepsilon n)^2)$, and the analysis follows since one cannot decrease a count of pairs (which are at most n^2 and always non-negative) by $\Omega((\varepsilon n)^2)$ more than $O(1/\varepsilon^2)$ times (see proof of Theorem 3 in [27]). We follow the same plan but show that the expected number of pairs in the same part after a single sample decreases by a multiplicative $(1 - \Omega(\varepsilon))$-factor (Lemma 5.5)—the improved bound of $O(\log(1/\varepsilon)/\varepsilon)$ sample complexity follows analogously.

Tree Metric and Ultrametric Testing Lower Bound We prove a sample complexity lower bound of $\Omega(1/\varepsilon)$ for one-sided error testing algorithms by constructing a distribution over matrices that are far from ultrametrics and tree metrics, where each violating triple contains one out of a set of εn points, making it necessary to sample at least one of these vertices in order to detect a violating triple.

We prove a query complexity lower bound of $\Omega(1/\varepsilon^{4/3})$ for testing ultrametrics, which in turn implies the same query lower bound for tree metrics. This is done by constructing a different distribution where we partition $[n]$ into $1/\varepsilon$ groups each of size εn, such that any violation involves all 3 points from the same group. We also prove that to maximize the probability of finding a violation, a tester is best off making queries on every pair of points in a suitably chosen sample. Finally, we show that to have a decent probability of finding a violation, the sample size must be $\Omega(1/\varepsilon^{2/3})$, leading to the query lower bound of $\Omega(1/\varepsilon^{4/3})$.

2 Preliminaries

We use the standard definitions of metrics, tree metrics, and ultrametrics. In order to be self-contained, we include these definitions in the appendix.

Next, we recall the standard model of property testing which we will use throughout the paper. We state the definitions of property testing as testing properties of matrices. As we will see, a distance function $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ may be encoded by an $n \times n$ matrix. This view will be useful, as our analysis will refer to "blocks", and "block-diagonal" structures which are more intuitive for matrices.

Recall that the goal of property testing algorithms is to provide very efficient, *sublinear time* query algorithms which approximately decide whether an object satisfies a property or is ε-far from satisfying a property. Towards that end, we consider algorithms that test with respect to the ℓ_0-distance which counts the number of entries (i.e., coordinates) where two matrices differ. Throughout the paper, we will encode functions $d\colon [n] \times [n] \to \mathbb{R}$ as $n \times n$ matrices.

DEFINITION 2.1. *Given two $n \times n$ matrices, $A, B \in \mathbb{R}^{n \times n}$, let*

$$\|A - B\|_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\{A_{ij} \neq B_{ij}\}.$$

We will refer to a property $\mathcal{P}$ of $n \times n$ matrices by letting $\mathcal{P}$ denote the subset of matrices that satisfy the property. For a subset $\mathcal{P}$ of $n \times n$ matrices and any $\varepsilon > 0$, we say that an $n \times n$ matrix M is ε-far from $\mathcal{P}$ if

$$d_{\ell_0}(M, \mathcal{P}) = \inf_{A \in \mathcal{P}} \|M - A\|_0 \geq \varepsilon n^2.$$

DEFINITION 2.2. (PROPERTY TESTING ALGORITHM FOR $n \times n$ MATRICES) *For any $n \in \mathbb{N}$ and $\varepsilon > 0$, an ε-testing algorithm which tests a property $\mathcal{P}$ is a randomized algorithm that receives query access to an unknown matrix $M \in \mathbb{R}^{n \times n}$ and satisfies:*

- **Completeness***: If $M \in \mathcal{P}$, the algorithm outputs "accept" with probability at least $2/3$.*
- **Soundness***: If M is ε-far from $\mathcal{P}$, the algorithm outputs "reject" with probability at least $2/3$.*

The algorithm is non-adaptive if the queries made do not depend on the answer to prior queries (i.e., all queries are made in parallel), and the algorithm achieves one-sided error if it accepts inputs $M \in \mathcal{P}$ with probability 1. We design algorithms that measure the following notions of complexity:

- **Sample Complexity***: The sample complexity of an algorithm is the maximum number of distinct rows or columns of the input matrix that are queried by the algorithm.*
- **Query Complexity***: The query complexity of an algorithm is the maximum number of entries of the input matrix that are queried by the algorithm.*

The definition of a metric requires the distance function to be non-negative with $d(i, i) = 0$, symmetric, and satisfy the triangle inequality (or the stronger four-point-condition and three-point-condition in metrics and ultrametrics). Our tests and analysis, which receive as input an arbitrary $n \times n$ matrix, will assume that the matrix is non-negative, symmetric, and zero only along the diagonal. The reason for this assumption is that these conditions may be easily checked algorithmically and incorporated into our testing algorithm. Formally, we state the following lemma, which allows us to design testing algorithms while focusing solely on the "interesting" violations; the lemma follows trivially by checking $O(1/\varepsilon)$ randomly chosen entries.

LEMMA 2.1. *For $n \in \mathbb{N}$, let $\mathcal{P}$ denote any property of $n \times n$ matrices, and let*

$$\mathcal{C} \stackrel{\text{def}}{=} \left\{ M \in \mathbb{R}^{n \times n} : \begin{array}{l} \textit{(i) for all } i, j \in [n],\ M(i,j) = M(j,i) \\ \textit{(ii) for all } i, j \in [n] \textit{ with } i \neq j,\ M(i,j) > 0 \\ \textit{(iii) for all } i \in [n],\ M(i,i) = 0 \end{array} \right\}.$$

Suppose there is an ε-testing algorithm for $\mathcal{P}$ which assumes the input $M \in \mathcal{C}$ and uses $q(n, \varepsilon)$ queries. Then, there is an ε-testing algorithm for $\mathcal{P} \cap \mathcal{C}$ using $O(q(n, \varepsilon)) + O(1/\varepsilon)$ queries.

3 Metric Testing Upper Bound: Theorem 1.1

In this section, we present an algorithm for testing general metrics with query complexity $O(n^{2/3}/\varepsilon^{4/3})$, thereby proving Theorem 1.1. Throughout this section, we will consider $n \in \mathbb{N}$ as the main asymptotic parameter, and let

$$\mathcal{P} = \{M \in \mathcal{C} : M \text{ encodes a metric space over } [n]\}.$$

Recall that the set $\mathcal{C}$ consists of $n \times n$ matrices which we call *clean*, meaning that they are symmetric, non-negative, and zero if and only if on the diagonal. The algorithm is straightforward and executes two types of checks, attempting to find violations of the triangle inequality. We begin by stating the definition which will be needed for the analysis.

DEFINITION 3.1. *A triangle is a set $\{i, j, k\}$ of three distinct indices $i, j, k \in [n]$. A triangle $\{i, j, k\}$ is* violating *for $M \in \mathcal{C}$ if it forms a violation of the triangle inequality. Namely, after re-naming indices so $M(i, j)$ is the maximum among pairwise evaluations,*

$$M(i, j) > M(i, k) + M(k, j).$$

LEMMA 3.1. *For any $\varepsilon \in (0, 1)$, and any $M \in \mathcal{C}$ which is ε-far from $\mathcal{P}$, there are at least $\varepsilon n^2/6$ distinct violating triangles $\{i, j, k\}$ of M.*

Proof. See appendix. ∎

DEFINITION 3.2. (TRIANGLE DEGREE) *For $M \in \mathcal{C}$ and a set of violating triangles T of M, we let $B(T) = ([n], T)$ be the 3-uniform hypergraph whose hyper-edges are violating triangles.*

- *For an index $i \in [n]$, the* vertex-triangle-degree*, $d_T(i)$ is the number of violating triangles containing i,*

$$d_T(i) = |t \in T : i \in t| .$$

- *For a pair $(i, j) \in [n] \times [n]$, the* edge-triangle-degree*, $d_T(i, j)$ is the number of violating triangles containing both i and j,*

$$d_T(i, j) = |t \in T : i, j \in t| .$$

Our test proceeds by executing two sub-routines CHECKHIDEGREE and CHECKVIOLATION which aim to find a violating triangle. We then state the two lemmas concerning these sub-routines and show how they imply the main tester.

Metric Testing Algorithm. The algorithm will aim to find a triangle $\{i, j, k\}$ which forms a violation of the triangle inequality. If it does find one, a violation is a certificate that the input matrix is not a metric and the algorithm will output "reject."

Input: The parameters $n \in \mathbb{N}$ and $\varepsilon \in (1/n, 1)$, as well as query access to the entries of an unknown $n \times n$ matrix M from $\mathcal{C}$ (see Lemma 2.1). For $\varepsilon < 1/n$, the claimed complexity $O(n^{2/3}/\varepsilon^{4/3})$ becomes $O(n^2)$, so one can read the entire matrix.
Output: "accept" or "reject."

1. Execute CHECKHIDEGREE(M, ε). If the sub-routine outputs "reject," then output "reject."
2. Execute CHECKVIOLATION(M, ε). If the sub-routine outputs "reject," then output "reject."
3. If neither sub-routine has output "reject," then output "accept."

The CheckHiDegree Sub-routine.

Input: The parameters $n \in \mathbb{N}$ and $\varepsilon \in (1/n, 1)$, as well as query access to the entries of an unknown $n \times n$ matrix $M \in \mathcal{C}$.
Output: "accept" or "reject."

1. For $u = O(1/\varepsilon)$, take u random samples $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_u \sim [n]$ drawn independently. For $s = O(n^{2/3}/\varepsilon^{1/3})$, take s random sample pairs $(\boldsymbol{j}_1, \boldsymbol{k}_1) \ldots, (\boldsymbol{j}_s, \boldsymbol{k}_s) \sim [n]^2$ drawn independently. Query $M(\boldsymbol{i}_\ell, \boldsymbol{j}_t), M(\boldsymbol{i}_\ell, \boldsymbol{k}_t), M(\boldsymbol{j}_t, \boldsymbol{k}_t)$ for all $\ell \in [u], t \in [s]$.
2. If there exists a triangle among the sampled indices $\{\boldsymbol{i}_\ell, \boldsymbol{j}_t, \boldsymbol{k}_t\}$ which is a violating triangle in M, output "reject." Otherwise, output "accept."

The CheckViolation Sub-routine.

Input: The parameters $n \in \mathbb{N}$ and $\varepsilon \in (1/n, 1)$, as well as query access to the entries of an unknown $n \times n$ matrix $M \in \mathcal{C}$.
Output: "accept" or "reject."

1. For $s = O(n^{1/3}/\varepsilon^{2/3})$, take s random samples $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s \sim [n]$ drawn independently. Query $M(\boldsymbol{i}_\ell, \boldsymbol{i}_k)$ for all $\ell, k \in [s]$.
2. If there exists a triangle among the sampled indices $\{\boldsymbol{i}_\ell, \boldsymbol{i}_k, \boldsymbol{i}_h\}$ which is a violating triangle in M, output "reject." Otherwise, output "accept."

LEMMA 3.2. (CHECKHIDEGREE LEMMA) *For $n \in \mathbb{N}$ and $\varepsilon \in (1/n, 1)$, there exists a randomized algorithm,* CHECKHIDEGREE*, which receives as input an $n \times n$ matrix $M \in \mathcal{C}$ and a parameter ε and has the following guarantees:*

- *If $M \in \mathcal{P}$,* CHECKHIDEGREE(M, ε) *always outputs "accept."*
- *Letting T be the set of violating triangles of M, if there are at least $\varepsilon n/4$ indices $i \in [n]$ such that $d_T(i) \geq \varepsilon^{1/3} n^{4/3}/16$,* CHECKHIDEGREE$(M, \varepsilon)$ *outputs "reject" with probability at least $5/6$.*

The algorithm is non-adaptive, taking $O(1/\varepsilon + n^{2/3}/\varepsilon^{1/3})$ samples and using $O(n^{2/3}/\varepsilon^{4/3})$ queries.

Proof. See appendix. ∎

LEMMA 3.3. (CHECKVIOLATION LEMMA) *For $n \in \mathbb{N}$ and $\varepsilon \in (1/n, 1)$, there exists a randomized algorithm,* CHECKVIOLATION*, which receives as input an $n \times n$ matrix $M \in \mathcal{C}$ and a parameter ε and has the following guarantees:*

- *If $M \in \mathcal{P}$,* CHECKVIOLATION(M, ε) *always outputs "accept."*
- *Letting T be the set of violating triangles of M, if M is ε-far from $\mathcal{P}$ and the set of indices $i \in [n]$ with $d_T(i) \geq \varepsilon^{1/3} n^{4/3}/16$ has size at most $\varepsilon n/4$, the sub-routine outputs "reject" with probability at least $5/6$.*

The algorithm is non-adaptive, taking $O(n^{1/3}/\varepsilon^{2/3})$ samples and using $O(n^{2/3}/\varepsilon^{4/3})$ queries.

Proof. [Proof of Theorem 1.1 Assuming Lemma 3.3] See appendix. ∎

3.1 CheckViolation Sub-routine: Proof of Lemma 3.3 Note that the sub-routine CheckViolation only outputs "reject" when it observes a violating triangle. Therefore, it is easy to establish the first condition of Lemma 3.3, as there are no violating triangles whenever $M \in \mathcal{P}$. It remains to prove that, whenever M is ε-far from $\mathcal{P}$ and at most $\varepsilon n/4$ indices $i \in [n]$ have $d_T(i) \geq \varepsilon^{1/3} n^{4/3}/16$, the sub-routine finds a violation with probability at least 5/6.

Definition 3.3. *For $M \in \mathcal{C}$ and any subset $\tilde{T} \subseteq T$ of violating triangles of M, the random variable $\mathbf{X}(\tilde{T}) \geq 0$ is given by*

$$\mathbf{X}(\tilde{T}) = \sum_{t \in \tilde{T}} \mathbf{1}\left\{t \subset \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s\}\right\}.$$

where $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s \sim [n]$ are indices sampled from CheckViolation.

Claim 3.1. *For any collection of triangles $\tilde{T}$, the expectation of $\mathbf{X}(\tilde{T})$ is bounded by*

$$\mathop{\mathbf{E}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\mathbf{X}(\tilde{T})\right] \geq \Omega\left(\frac{|\tilde{T}| \cdot s^3}{n^3}\right),$$

whenever $3 \leq s \ll n$. Moreover, the variance of the random variable is bounded by

$$\mathop{\mathbf{Var}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\mathbf{X}(\tilde{T})\right] \lesssim \mathop{\mathbf{E}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\mathbf{X}(\tilde{T})\right] + \left(\sum_{i \in [n]} d_{\tilde{T}}(i)^2\right) \cdot \frac{s^5}{n^5} + \left(\sum_{i \neq j} d_{\tilde{T}}(i,j)^2\right) \cdot \frac{s^4}{n^4},$$

where the second term counts pairs of triangles $t, t' \in \tilde{T}$ which intersect at a single vertex, and the third term counts the number of pairs of triangles $t, t' \in \tilde{T}$ which intersect at two vertices.

Proof. By linearity of expectation,

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\mathbf{X}(\tilde{T})\right] &\geq |\tilde{T}| \sum_{1 \leq k_1 < k_2 < k_3 \leq s} 3! \cdot \mathbf{Pr}\left[\begin{array}{c}\boldsymbol{i}_{k_1} = 1, \boldsymbol{i}_{k_2} = 2, \boldsymbol{i}_{k_3} = 3 \\ \text{and occur uniquely}\end{array}\right] \\ &\geq |\tilde{T}| \cdot \binom{s}{3} \cdot 3! \cdot \frac{1}{n^3}\left(1 - \frac{3}{n}\right)^{s-3} = \Omega\left(\frac{|\tilde{T}| \cdot s^3}{n^3}\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathop{\mathbf{Var}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\mathbf{X}(\tilde{T})\right] &= \mathop{\mathbf{E}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\left(\sum_{t \in \tilde{T}} \mathbf{1}\left\{t \subset \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s\}\right\}\right)^2\right] - \mathop{\mathbf{E}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\sum_{t \in \tilde{T}} \mathbf{1}\left\{t \subset \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s\}\right\}\right]^2 \\ &\lesssim \mathop{\mathbf{E}}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_s}\left[\mathbf{X}(\tilde{T})\right] + \left(\sum_{i \in [n]} d_{\tilde{T}}(i)^2\right) \cdot \frac{s^5}{n^5} + \left(\sum_{i \neq j} d_{\tilde{T}}(i,j)^2\right) \cdot \frac{s^4}{n^4}.\end{aligned}$$

□

Since $\mathbf{X}(\tilde{T})$ counts the number of violating triangles from $\tilde{T}$ included in the random sample $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s$, if $\mathbf{X}(\tilde{T}) > 0$, then the algorithm has sampled a violating triangle. Once it makes all pairwise queries, it will find the violating triangle and output "reject." The analysis will find a subset $\tilde{T}$ of violating triangles such that the random variable $\mathbf{X}(\tilde{T})$ is non-zero with high constant probability, which indicates that the random sample contains at least one violating triangle. More specifically, the goal is to find an appropriate set of violating triangles $\tilde{T}$ such that the expectation of $\mathbf{X}(\tilde{T})$ is large while the variance is small. Hence, the $\tilde{T}$ used in the analysis has large cardinality, and both vertex-triangle-degree and edge-triangle-degree of the indices defined over $\tilde{T}$ are bounded.

LEMMA 3.4. *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}$ and let T denote the set of all violating triangles of M. Suppose that the set of indices $i \in [n]$ such that $d_T(i) \geq \varepsilon^{1/3}n^{4/3}/16$ has size at most $\varepsilon n/4$. Then, there exists a subset $\tilde{T}$ of violating triangles of M satisfying*

- $|\tilde{T}| \geq \Omega(\varepsilon n^2)$,
- *For every index $i \in [n], d_{\tilde{T}}(i) \leq O(\varepsilon^{1/3}n^{4/3})$,*
- *for every pair of indices $(i,j) \in [n] \times [n]$, $d_{\tilde{T}}(i,j) \leq O(n^{2/3}/\varepsilon^{1/3})$.*

Proof. [Proof of Lemma 3.3 assuming Lemma 3.4] The sub-routine CHECKVIOLATION(M, ε) samples $c \cdot n^{1/3}/\varepsilon^{2/3}$ indices uniformly at random for some constant c. Since it is guaranteed that there are at least $\Omega(\varepsilon n^2)$ violating triangles in the set $\tilde{T}$, Claim 3.1 implies the expected value of $\mathbf{X}(\tilde{T})$ is a large constant for c large enough. From Lemma 3.4, we derive the following two inequalities:

$$\sum_{i\in[n]} d_{\tilde{T}}(i)^2 \leq O(\varepsilon^{1/3}n^{4/3}) \sum_{i\in[n]} d_{\tilde{T}}(i) = O\left(\varepsilon^{1/3}n^{4/3} \cdot |\tilde{T}|\right),$$

$$\sum_{(i,j)\in[n]\times[n]} d_{\tilde{T}}(i,j)^2 \leq O(n^{2/3}/\varepsilon^{1/3}) \sum_{(i,j)\in[n]\times[n]} d_{\tilde{T}}(i,j) = O\left(n^{2/3}/\varepsilon^{1/3} \cdot |\tilde{T}|\right).$$

Consider (for the sake of analysis) repeating the above randomized trial for $\mathbf{X}(\tilde{T})$ for k independent iterations, letting $\mathbf{X}_1(\tilde{T}), \ldots, \mathbf{X}_k(\tilde{T})$ denote the outcomes of k independent trials. Then, taking the average and applying Chebyshev's inequality,

$$\begin{aligned}\mathbf{Pr}\left[\frac{1}{k}\sum_{\ell=1}^{k}\mathbf{X}_\ell(\tilde{T}) = 0\right] \leq \frac{\mathbf{Var}[\mathbf{X}(\tilde{T})]}{k \cdot \mathbf{E}[\mathbf{X}(\tilde{T})]^2} &\lesssim \frac{\mathbf{E}[\mathbf{X}(\tilde{T})]}{k \cdot \mathbf{E}[\mathbf{X}(\tilde{T})]^2} + \frac{\varepsilon^{1/3}n^{4/3} \cdot |\tilde{T}| \cdot \frac{s^5}{n^5}}{k \cdot \mathbf{E}[\mathbf{X}(\tilde{T})]^2} + \frac{n^{2/3}/\varepsilon^{1/3} \cdot |\tilde{T}| \cdot \frac{s^4}{n^4}}{k \cdot \mathbf{E}[\mathbf{X}(\tilde{T})]^2} \\ &= O\left(\frac{n^3}{k \cdot |\tilde{T}| \cdot s^3}\right) + O\left(\frac{\varepsilon^{1/3}n^{7/3}}{k \cdot |\tilde{T}| \cdot s}\right) + O\left(\frac{n^{8/3}}{\varepsilon^{1/3} \cdot k \cdot |\tilde{T}| \cdot s^2}\right),\end{aligned}$$

which can be made an arbitrarily small constant when $|\tilde{T}| = \Omega(\varepsilon n^2)$, $s = O(n^{1/3}/\varepsilon^{2/3})$ and $k = O(1)$. Notice that the second and the third terms are the bottle-neck.

□

3.2 Finding A Good Subset of Violating Triangles: Proof of Lemma 3.4

LEMMA 3.5. *Suppose $M \in \mathcal{C}$ and is ε-far from $\mathcal{P}$ and there are at most $\varepsilon n/4$ indices $i \in [n]$ such that the vertex-triangle degree $d_T(i) \geq \varepsilon^{1/3}n^{4/3}/16$. Then there exists a matrix $M' \in \mathcal{C}$ which is $\varepsilon/2$-far from $\mathcal{P}$ such that all violating triangles in M' are also in M. Denote this set of violating triangles in M' to be T'. Moreover, for all $i \in [n]$, $d_{T'}(i) < \varepsilon^{1/3}n^{4/3}/16$.*

Proof. Let $m \in \mathbb{R}$ be the value of the maximum entry in M, and let $I \subset [n]$ be the set of indices $i \in [n]$ with $d_T(i) \geq \varepsilon^{1/3}n^{4/3}/16$. We consider the following $n \times n$ matrix M', where we set $M'(i,j) = M(i,j)$ unless, either i or j lie in I, in which case $M'(i,j) = m$. Note that $\|M' - M\|_0 \leq 2|I| \cdot n \leq 2\varepsilon n^2/4 = \varepsilon n^2/2$, so that M' is $\varepsilon/2$-far from $\mathcal{P}$. We apply Lemma 3.1 to M', which shows that there are at least $\varepsilon n^2/12$ distinct violating triangles $\{i,j,k\}$ in M'. Note that, for any violating triangle $\{i,j,k\}$ in M', i,j,k are not in I otherwise $\{i,j,k\}$ is not violating in M'. Hence, the distances between these three vertices are not changed, so this triangle is also violating in M. This shows T', the set of all violating triangles in M', is a subset of T. Thus, for $i \notin I$, $d_{T'}(i) \leq d_T(i) < \varepsilon^{1/3}n^{4/3}/16$. For $i \in I$, no violating triangle in T' contains i, so $d_{T'}(i) = 0$.
□

DEFINITION 3.4. *Let the set of "high-degree-edges" $\overline{E}$ to be the set of pairs of indices (i,j) such that $d_{T'}(i,j) > 10n^{2/3}/\varepsilon^{1/3}$.*

LEMMA 3.6. *For any $i \in [n]$, the number of edges in $\overline{E}$ that contain i is upper bounded by $\varepsilon^{2/3}n^{2/3}/80$. Therefore, the size of $\overline{E}$ is upper bounded by $\varepsilon^{2/3}n^{5/3}/80$, which is at most $\varepsilon n^2/80$ for $\varepsilon > 1/n$.*

Proof. Notice that $d_{T'}(i) = \sum_{e:i\in e} d_{T'}(e)/2 \geq 10n^{2/3}/\varepsilon^{1/3} \cdot |\{e \in \overline{E} : i \in e\}|/2$. By the bound on the vertex-triangle-degree defined over T' given in Lemma 3.5, $d_{T'}(i) \leq \varepsilon^{1/3}n^{4/3}/16$. Thus, $|\{e \in \overline{E} : i \in e\}| \leq \varepsilon^{2/3}n^{2/3}/80$. ∎

DEFINITION 3.5. (UNIQUE TRIANGLES) *For each edge $e \in \overline{E}$, define the unique triangles to e to be the subset $T_e^{(u)} \subset T'$ such that each triangle in $T_e^{(u)}$ only uses edge e and nothing else in $\overline{E}$.*

LEMMA 3.7. *For any edge $e \in \overline{E}$, $|T_e^{(u)}| \geq 9n^{2/3}/\varepsilon^{1/3}$.*

Proof. Let edge $e = (i,j)$. Since $e \in \overline{E}$, e is a high-degree edge: $d_{T'}(e) > 10n^{2/3}/\varepsilon^{1/3}$. By Lemma 3.6, we can upper bound the number of violating triangles that use e and some other high-degree edge $(i,v) \in \overline{E}$ since few high-degree edges are using the vertex i. Quantitatively, we have

$$\left|\left\{\{i,j,v\} \in T' : (i,v) \in \overline{E}\right\}\right| \leq \left|\left\{\{i,j,v\} \subset [n] : (i,v) \in \overline{E}\right\}\right| \leq \varepsilon^{2/3}n^{2/3}/80.$$

And similarly for vertex j, we have

$$\left|\left\{\{i,j,v\} \in T' : (j,v) \in \overline{E}\right\}\right| \leq \left|\left\{\{i,j,v\} \subset [n] : (j,v) \in \overline{E}\right\}\right| \leq \varepsilon^{2/3}n^{2/3}/80.$$

Therefore, the violating triangles that remain and only use edge e can be bounded below.

$$|T_e^{(u)}| = \left|\left\{\{i,j,v\} \in T' : (i,v),(j,v) \notin \overline{E}\right\}\right| \geq d_{T'}(i,j) - \varepsilon^{2/3}n^{2/3}/40 \geq 9n^{2/3}/\varepsilon^{1/3}.$$

∎

Notice that Lemma 3.5 above guarantees the existence of a subset $T' \subset T$ of violating triangles such that the vertex-triangle-degree defined over T' at each vertex is bounded above. The construction of T' is through modifying all entries, or edges, that previously contained a high-degree vertex. The following lemma continues the modification and constructs a subset of violating triangles $\tilde{T} \subset T'$ such that the *edge-triangle-degree* defined over $\tilde{T}$ at each edge is bounded above. The idea is to modify all high-degree entries, or edges. Recall that $\overline{E}$ denotes the set of high-degree edges.

LEMMA 3.8. *Suppose matrix M' and the set of violating triangles T' are as defined in Lemma 3.5. Then there exists a matrix M'' and a subset of violating triangles $\tilde{T} \subset T'$ that satisfy the below properties:*

- *M' and M'' differ on only the entries specified in $\overline{E}$*
- *Suppose T'' is the set of all violating triangles in M'', then $\tilde{T} \subset T'$ has size at least $|T''|$*
- *For all $i,j \in [n]$, $d_{\tilde{T}}(i,j) \leq 10n^{2/3}/\varepsilon^{1/3}$.*

Proof. [Proof of Lemma 3.4 assuming Lemma 3.8] By Lemma 3.6, $|\overline{E}| \leq \varepsilon n^2/80$. Moreover, by construction in Lemma 3.5, M' and M differ on at most $\varepsilon n^2/2$ entries. Since M' and M'' differ on only $2|\overline{E}| \leq \varepsilon n^2/40$

entries, M and M'' differ on at most $21\varepsilon n^2/40$ entries, meaning that M'' is $19\varepsilon/40$-far from $\mathcal{P}$. Again, using Lemma 3.1, there are at least $19\varepsilon n^2/240$ violating triangles in M''. That is, $|T''| \geq 19\varepsilon n^2/240$. Lemma 3.8 promises the existence of a set of violating triangles $\tilde{T} \subset T'$ with size at least $|T''| \geq 19\varepsilon n^2/240$, and $d_{\tilde{T}}(i,j) \leq 10n^{2/3}/\varepsilon^{1/3}$ for all $i,j \in [n]$. Moreover, Lemma 3.5 states that every triangle in T' is violating in M. Thus, $\tilde{T}$ is a subset of violating triangles in M. Lastly, for any index $i \in [n]$, $d_{\tilde{T}}(i) \leq d_{T'}(i) \leq \varepsilon^{1/3}n^{4/3}/16$. ∎

Proof. [Proof of Lemma 3.8] First, we specify the values M'' takes on for each entry in $\overline{E}$ (Claim 3.2). Then, we provide a mapping from T'' to a subset of violating triangles $\tilde{T} \subset T'$ in M' (Claim 3.4) and show that $\tilde{T}$ has the claimed properties.

Recall that a high-degree edge in $\overline{E}$ corresponds to an entry in M' that participates in many violating triangles. The claim below (Claim 3.2) states that there exists some new value $x \in \mathbb{R}_{\geq 0}$ such that, if the entry $e \in \overline{E}$ of the matrix M is changed to the new value, the entry participates in much fewer violating triangles in the new matrix. We give a notation for finding a new value for a high-degree edge. For indices $i,j,k \in [n]$ and for any value $x \in \mathbb{R}_{>0}$, let $\mathbb{I}(i,j,k,x)$ denote the indicator variable in $\{0,1\}$ with

$$\mathbb{I}(i,j,k,x) = 1 \qquad \iff \qquad \left\{ \begin{array}{c} (i,k) \notin \overline{E}, \\ (j,k) \notin \overline{E}, \\ x \notin [|M'(i,k) - M'(j,k)|, M'(i,k) + M'(j,k)] \end{array} \right\}.$$

Claim 3.2. *For each $e = (i,j) \in \overline{E}$, whenever $d_{T'}(i)$ and $d_{T'}(j)$ are both bounded above by $\varepsilon^{1/3}n^{4/3}/16$, there exists a value $x(e) \in \mathbb{R}_{\geq 0}$ such that*

$$\sum_{k \in [n]} \mathbb{I}(i,j,k,x(e)) \leq \varepsilon^{1/6} \cdot n^{2/3}.$$

Assuming Claim 3.2 (which we formally prove shortly), let

$$M''(i,j) = \left\{ \begin{array}{cc} M'(i,j) & (i,j) \notin \overline{E} \\ x(e) & (i,j) = e \in \overline{E} \end{array} \right. ,$$

as the setting $x(e)$ from Claim 3.2, and let T'' denote all violating triangles of M''.

Claim 3.3. *For every edge $e = (i,j) \in \overline{E}$, let A_e be the subset of triangles in T'' that contain the edge e. Then, $d_{T''}(e) = |A_e| \leq 2\varepsilon^{1/6}n^{2/3}$.*

Proof. Following Claim 3.2, we can upper bound the number of triangles in T'' which only use edge e and nothing else in $\overline{E}$. That is,

$$\left| \left\{ \{i,j,k\} \in T'' : (i,k),(j,k) \notin \overline{E} \right\} \right| = \sum_{k \in [n]} \mathbb{I}(i,j,k,x(e)) \leq \varepsilon^{1/6}n^{2/3}.$$

On the other hand, few triangles use e and some other edge in $\overline{E}$, as implied by Lemma 3.6. That is,

$$\left| \left\{ \{i,j,k\} \in [n] : (i,k) \text{ or } (j,k) \in \overline{E} \right\} \right| \leq \varepsilon^{2/3}n^{2/3}/40.$$

Together, these two sets of triangles consist all of A_e. Hence, the edge-triangle degree of e over the set T'' is upper bounded by $d_{T''}(e) = |A_e| \leq \varepsilon^{1/6}n^{2/3} + \varepsilon^{2/3}n^{2/3}/40 \leq 2\varepsilon^{1/6}n^{2/3}$. ∎

With the set of all violating triangles T'' in M'', we then use the below helper lemma to construct a subset of violating triangles $\tilde{T} \subset T'$ that inherits the nice property of T'': the edge-triangle-degree defined over $\tilde{T}$ for each edge e is upper bounded.

CLAIM 3.4. *For each $e = (i, j) \in \overline{E}$, let A_e be the subset of triangles in T'' that contain the edge e. Then, there exists a subset of violating triangles $B_e \subset T_e^{(u)} \subset T'$ such that $|A_e| = |B_e|$.*

Proof. [Proof of Claim 3.4] Notice that $|A_e| = d_{T''}(e) \leq 2\varepsilon^{1/6}n^{2/3}$. By Lemma 3.7, $|T_e^{(u)}| \geq 9n^{2/3}/\varepsilon^{1/3}$. This suggests the existence of $B_e \subset T_e^{(u)} \backslash A_e$ with size $|A_e| \leq 2\varepsilon^{1/6}n^{2/3}$. ∎

Claim 3.4 above hints on how to construct a mapping from T'' to $\tilde{T}$. We initialize an empty set $\tilde{T}$. First, add to $\tilde{T}$ all the triangles in T'' that do not use any of the high-degree edges in $\overline{E}$. Then, for each $e \in \overline{E}$, following Claim 3.4, find the subset $B_e \subset T_e^{(u)}$ and add B_e to $\tilde{T}$. What we are left to show is that $\tilde{T}$ has the properties claimed in Lemma 3.8. That is, $|\tilde{T}| \geq |T''|$ and for any edge $(i, j) \in [n] \times [n]$, $d_{\tilde{T}}(i, j) \leq 10n^{2/3}/\varepsilon^{1/3}$.

CLAIM 3.5. *Define $\tilde{T}$ to be*

$$(\bigcup_{e \in \overline{E}} B_e) \cup \{\{i, j, k\} \in T'' : (i, j), (j, k), (i, k) \notin \overline{E}\}.$$

Then, $|\tilde{T}| \geq |T''|$. Furthermore, $\tilde{T} \subset T'$.

Proof. The inequality follows from the fact that each triangle in T'' corresponds to at least one unique triangle in $\tilde{T}$. Notice that

$$T'' = (\bigcup_{e \in \overline{E}} A_e) \cup \{\{i, j, k\} \in T'' : (i, j), (j, k), (i, k) \notin \overline{E}\}.$$

The set of triangles that do not use any high-degree edge $\overline{E}$ is exactly the same in T'' and $\tilde{T}$. On the other hand, notice that each B_e is disjoint from the set $\{\{i, j, k\} : (i, j), (j, k), (i, k) \notin \overline{E}\}$. Moreover, B_e and $B_{e'}$ are disjoint whenever $e \neq e'$ since B_e is a subset of $T_e^{(u)}$. Since $|A_e| = |B_e|$, for each edge $e \in \overline{E}$ there is some bijective map from A_e to B_e. Thus, each triangle in T'' corresponds to at least one unique triangle in $\tilde{T}$, which suggests $|\tilde{T}| \geq |T''|$. Note that the inequality may be strict because A_e might overlap with $A_{e'}$ for $e \neq e'$, and it only makes the union larger: $|B_e \cup B_{e'}| > |A_e \cup A_{e'}|$.

It is clear that every triangle in B_e for each e is in $T_e^{(u)} \subset T'$. Moreover, each triangle in $\{\{i, j, k\} \in T'' : (i, j), (j, k), (i, k) \notin \overline{E}\}$ has same distance values on all three edges in M' and M''. Since they are violating in M'', they are also violating in M', which indicates $\{\{i, j, k\} \in T'' : (i, j), (j, k), (i, k) \notin \overline{E}\} \subset T'$. ∎

CLAIM 3.6. *For every edge $e = (i, j)$, $d_{\tilde{T}}(e) \leq 10n^{2/3}/\varepsilon^{1/3}$.*

Proof. If $e \notin \overline{E}$, by definition $d_{\tilde{T}}(e) \leq d_{T'}(e) \leq 10n^{2/3}/\varepsilon^{1/3}$. If $e \in \overline{E}$, the only triangles in $\tilde{T}$ that contain e are the triangles in B_e. Hence, $d_{\tilde{T}}(e) = |B_e| = |A_e| = d_{T''}(e) \leq 10n^{2/3}/\varepsilon^{1/3}$. ∎

The two above Claims complete the proof for Lemma 3.8, and we are left to show the correctness of Claim 3.2.

Proof. [Proof of Claim 3.2] The statement follows from the below claim.

CLAIM 3.7. *Suppose for the edge $e = (i, j) \in \overline{E}$ and any value $x(e) \in (0, \infty)$, $\sum_{k \in [n]} \mathbb{I}(i, j, k, x(e)) \geq t$, then either $d_{T'}(i) \geq t^2/8$ or $d_{T'}(j) \geq t^2/8$.*

Suppose Claim 3.7 holds. We take t to be $\varepsilon^{1/6}n^{2/3}$. If for any value $x(e) \in (0, \infty)$, $\sum_{k \in [n]} \mathbb{I}(i, j, k, x(e)) \geq t = \varepsilon^{1/6}n^{2/3}$, then we have a contradiction since the vertex-triangle-degrees of both i and j defined over T' are at most $\varepsilon^{1/3}n^{4/3}/16$. Thus, we are done if the above claim holds.

Proof. [Proof of Claim 3.7] First, the indicator variable $\mathbb{I}(i,j,k,x(e))$ equals 1 if and only if both edges $(i,k),(j,k) \notin \overline{E}$ and the value $x(e)$ does not lie within the interval $[|M'(i,k) - M'(j,k)|, M'(i,k) + M'(j,k)]$. We will then show that if for all $x(e) \in (0,\infty)$,

$$\sum_{k\in[n]} \mathbb{I}(i,j,k,x(e)) \geq t,$$

then there exist at least $t^2/4$ pairs $(k_1,k_2) \in [n] \times [n]$ such that the intersection of intervals

$$[|M'(i,k_1) - M'(j,k_1)|, M'(i,k_1) + M'(j,k_1)] \cap [|M'(i,k_2) - M'(j,k_2)|, M'(i,k_2) + M'(j,k_2)] = \emptyset.$$

The statement is true more generally, if there is a collection of m intervals $I_1,\ldots,I_m$ of the real line, and for all x, there are at least t intervals which do not contain x, then there are two subsets L and R of at least $t/2$ intervals where $l \in L$ and $r \in R$ are disjoint, giving us $t^2/4$ pairs. The proof is algorithmic:

- Consider sorting the intervals by increasing endpoints, and we will let x scan from $-\infty$ to ∞ according to the endpoints of intervals. Let $x_{t/2}$ denote the endpoint of the $t/2$-th interval and let D denote the set of intervals disjoint from $x_{t/2}$.
- Note that, there must be at most $t/2$ intervals which lie to the left of $x_{t/2}$, since we scanned in increasing order of endpoint; and since D is disjoint, there must be at least $t - (t/2) \geq t/2$ intervals whose start is larger than $x_{t/2}$. Thus, we let L denote the $t/2$ intervals of first endpoints, and we let R denote the intervals in D whose start points come after $x_{t/2}$.

Now, consider the implication that two intervals $[|M'(i,k_1)-M'(j,k_1)|, M'(i,k_1)+M'(j,k_1)]$ and $[|M'(i,k_2)-M'(j,k_2)|, M'(i,k_2)+M'(j,k_2)]$ are disjoint, for two vertices k_1,k_2. Without loss of generality, we assume that $|M'(i,k_1) - M'(j,k_1)| \leq M'(i,k_1) + M'(j,k_1) < |M'(i,k_2) - M'(j,k_2)| \leq M'(i,k_2) + M'(j,k_2)$. Then, from the middle inequality, it must be the case that either $M'(i,k_2) > M'(i,k_1) + M'(j,k_1) + M'(j,k_2)$ or $M'(j,k_2) > M'(i,k_1) + M'(j,k_1) + M'(i,k_2)$. Intuitively, it means the 4-cycle $\{i,k_1,j,k_2\}$ is "violating". Suppose we are in the case that $M'(i,k_2)$ is larger than the sum of the other three edges. (The other case that $M'(j,k_2)$ is the largest follows the same analysis.) Then, it must be the case that either $M'(i,k_2) > M'(i,k_1) + M'(k_1,k_2)$ or $M'(k_1,k_2) > M'(j,k_1) + M'(j,k_2)$. In the former case, edge (k_1,k_2) forms a violating triangle with index i while in the latter case, edge (k_1,k_2) forms a violating triangle with index j.

Combining all the above arguments, there are at least $t^2/4$ pairs of pairwise disjoint intervals. Each pair corresponds to a unique edge (k_1,k_2) for some $k_1,k_2 \in [n]$ which forms a violating triangle with either i or j. By pigeon-hole theorem, at least $t^2/8$ distinct edges form some violating triangles with either i or j, implying that $d_{T'}(i) \geq t^2/8$ or $d_{T'}(j) \geq t^2/8$. The claim follows. ∎

∎

∎

4 Metric Testing Lower Bound—Theorem 1.2

We specify a distribution $\mathcal{D}_{\text{no}}$ which is supported on $(3n) \times (3n)$ matrices that are ε-far from being a metric, where $\varepsilon = n^{-\nu(n)}$, for $\nu(n) = (\log\log\log n + 4)/\log\log n$. We show that a non-adaptive algorithm which makes $o(n^{2/3+2\nu(n)/3})$ queries will not be able to find a violation of the triangle inequality with high probability when the input $\mathbf{M} \sim \mathcal{D}_{\text{no}}$. The construction follows closely to the Behrend graph construction used in [1] for showing the hardness of testing triangle-freeness. In particular, we make use of a Salem-Spencer set, which is a dense set of numbers that is free of 3-arithmetic progressions.

LEMMA 4.1. (LEMMA 4 IN [1]) *For any sufficiently large n, there exists a set $X \subset [n]$ of size at least $n^{1-\nu(n)}$, for $\nu(n) = (\log\log\log n + 4)/\log\log n$, so that for all $x, y, z \in X$, $x + y \equiv 2z \mod n$ if and only if $x = y = z$.*

Description of $\mathcal{D}_{\mathbf{no}}$. We describe a single $(3n) \times (3n)$ matrix M, which we will then randomize by re-shuffling indices in $[3n]$. Our construction follows closely that of the Behrend graph (Section 5.3 in [1]), which is a tripartite graph on $3n$ vertices (each part with n vertices) and $n|X|$ edge-disjoint triangles, where set X is from Lemma 4.1. For simplicity, we partition the set of $3n$ indices into A, B, and C, each of size n, and we will associate each $a \in A$ with a number in $[n]$. Hence, we refer to $M(a, b)$ as the entry in the matrix M of row $a \in A$ and column $b \in B$—this causes some ambiguity among the indices, since $a \in A$ both corresponds to a number, and the fact we use "a" means we refer to the subset of rows/columns associated with A (similarly for b and c). We let:

- For any $a \in A$ and $b \in B$, we let $M(a, b) = M(b, a) = 1$ if and only if $b - a \equiv x \mod n$ with $x \in X$, and we let $M(a, b) = M(b, a) = 2$ otherwise.
- For any $b \in B$ and $c \in C$, we let $M(b, c) = M(c, b) = 2$ if and only if $c - b \equiv x \mod n$ with $x \in X$, and we let $M(b, c) = M(c, b) = 3$ otherwise.
- For any $a \in A$ and $c \in C$, we let $M(a, c) = M(c, a) = 4$ if and only if $c - a \equiv 2x \mod n$ with $x \in X$ and $M(a, c) = M(c, a) = 2$ otherwise.
- Every other non-diagonal entry $M(x, y) = 2$ and every diagonal entry $M(x, x) = 0$.

We consider the following important properties of the construction above:

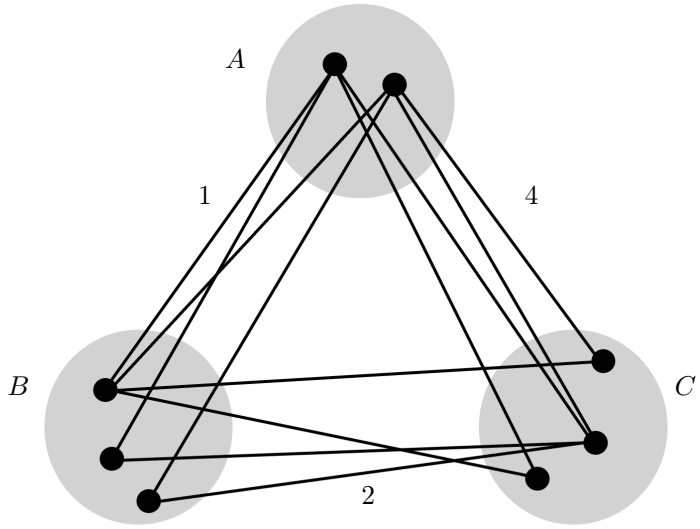
CLAIM 4.1. *For the matrix M constructed above, we have:*

- *Suppose i, j, k are three indices that form a violating triangle. Then, up to a permutation, $M(i, j) = 1$, $M(j, k) = 2$, and $M(k, i) = 4$.*
- *For the permutation where the above holds, there is a single entry $i = a \in A$, $j = b \in B$, and $k = c \in C$.*
- *A violating triangle $\{a, b, c\}$ as above is uniquely associated with $a \in A$ and $x \in X$, so the violating triangle consists of $\{a, a + x \mod n, a + 2x \mod n\}$.*

The above implies that $2n|X|$ modifications on the entries of M are required in order to remove all violations of the triangle inequality. Furthermore, the guarantee of Lemma 4.1 ensures that there are exactly $n|X|$ triples $\{a, b, c\}$ which violate the triangle inequality.

Proof. [Proof of Claim 4.1] First, notice that all distances are $\{0, 1, 2, 3, 4\}$, where distances of 0 are solely for $M(x, x)$. Thus, violations of the triangle inequality come from the possible tuples $(1, 1, 3), (1, 1, 4), (1, 2, 4)$. However, we note that one cannot have violations of the triangle inequality of the form $(1, 1, 3)$ or $(1, 1, 4)$. The reason is that $M(i, j) = 1$ occurs only if indices i, j correspond to some $a \in A$ and $b \in B$. Thus, if $M(i, j) = 1$ and $M(j, k) = 1$, then i, j, k are all among $A \cup B$. However, the only entries set to 3 or 4 are incident on C.

Furthermore, violations of the form $(1, 2, 4)$ are exactly those arising from the $n|X|$ edge-disjoint triangles $\{a, b, c\}$ specified by some $x \in X$, where $b - a \equiv x \mod n$, $c - b \equiv x \mod n$ and $c - a \equiv 2x \mod n$. This follows from the fact that X is three of length-3 arithmetic progressions. Since these violating triangles are edge-disjoint, each entry in M is involved in at most one 3-tuples violating the triangle inequality. Moreover, since at least one entry in $M(i, j), M(i, k), M(j, k)$ has to be modified for the triple $\{i, j, k\}$ to obey metric property, we need to modify at least one edge in each violating triple. That is, at least $n|X|$ entries in M has to be modified for M to be a metric. In fact, exactly $n|X|$ modifications suffice: simply modify all edges with weight 4 to 3. There are $n|X|$ such edges in G, which correspond to $2n|X|$ such entries in M. ∎



See Figure 4 for a description of the above construction, see also Section 2.6 in [29]. We let $\mathbf{M} \sim \mathcal{D}_{\text{no}}$ be obtained from M by re-ordering rows and columns according to a uniformly random permutation $\boldsymbol{\pi}$. By Yao's minimax principle, in order to rule out non-adaptive algorithms that have one-sided error (i.e., always accept metrics), it suffices to rule out any deterministic and non-adaptive algorithm for finding violating triangles in a draw $\mathbf{M} \sim \mathcal{D}_{\text{no}}$.

LEMMA 4.2. *Consider any deterministic non-adaptive algorithm that makes $o(n^{2/3+2\nu(n)/3})$ queries. With probability at least $2/3$ over the draw of $\mathbf{M} \sim \mathcal{D}_{no}$, there are no queries $(i,j),(j,k)$, and (i,k) with*

$$\mathbf{M}(i,j) > \mathbf{M}(i,k) + \mathbf{M}(k,j).$$

Since any partial matrix that does not violate the triangle inequality may be completed to one which is a metric (by considering the observed weighted sub-graph and computing shortest paths, as well as a maximum distance), any non-adaptive and one-sided algorithm must make $\Omega(n^{2/3+2\nu(n)/3})$ queries.

Proof. Consider a deterministic non-adaptive algorithm which queries $k = o(n^{2/3+2\nu(n)/3})$ entries, and let $E \subset [n] \times [n]$ be this set of queries—up to a factor of 2, we may assume that the algorithm is symmetric so that it queries (i,j) and (j,i). The probability that a violating triangle is among the queries E, via a union bound, is at most

$$\sum_{\substack{a \in A \\ x \in X}} \mathbf{Pr}_{\boldsymbol{\pi}} \left[\left\{ \begin{array}{c} (\boldsymbol{\pi}(a), \boldsymbol{\pi}(a+x)), \\ (\boldsymbol{\pi}(a+x), \boldsymbol{\pi}(a+2x)), \\ (\boldsymbol{\pi}(a), \boldsymbol{\pi}(a+2x)) \end{array} \right\} \subset E \right],$$

where the summation is over the $n|X|$ possible violating triangles. If we let $T(E)$ denote the set of triangles among the E queries, the above expression becomes

$$n|X| \cdot O\left(\frac{1}{n^3}\right) \cdot |T(E)|,$$

since the probability that any fixed triangle is mapped to a fixed triangle in $T(E)$ under $\boldsymbol{\pi}$ is $O(1/n^3)$. Since the maximum number of triangles in $|E|$ edges is $O(|E|^{3/2})$, the probability is upper bounded by

$$O\left(\frac{|X|}{n^2} \cdot |E|^{3/2}\right) = O\left(\frac{|E|^{3/2}}{n^{1+\nu(n)}}\right),$$

which is $o(1)$ if $|E| = o(n^{2/3+2\nu(n)/3})$.

□

5 Ultrametric and Tree Metric Testing Upper Bound: Theorem 1.3

In this section, we present algorithms for testing ultrametrics and tree metrics, both using $\tilde{O}(1/\varepsilon)$ samples and $\tilde{O}(1/\varepsilon^2)$ queries, thereby proving Theorem 1.3. We let

$$\mathcal{P}^U = \{M \in \mathcal{C}, M \text{ encodes an ultrametric space over } [n]\},$$
$$\mathcal{P}^T = \{M \in \mathcal{C}, M \text{ encodes a tree metric space over } [n]\}.$$

5.1 Ultrametric Testing Upper Bound As per Lemma 2.1, input matrices $M \in \mathcal{C}$ are already symmetric, non-negative, and zero only on the diagonal. As in Section 3, algorithms will have one-sided error, meaning that the tester's task is finding a certificate that the input matrix is not in $\mathcal{P}^U$. We first define the type of violation the algorithm seeks.

Definition 5.1. (Ultrametric Violating Triple) *Given $M \in \mathcal{C}$, the triple $\{i, j, k\}$ is a violation of ultrametric if, after renaming so $M(i, j)$ is maximum among the 3 pairwise distances,*

$$M(i, j) > M(i, k) \text{ and } M(i, j) > M(j, k).$$

We present the algorithm for testing ultrametrics, UltraTesting.

UltraTesting Algorithm. The algorithm aims to find a violating triple of indices $\{i, j, k\}$ of the ultrametric property. If a violating triple is found, it constitutes a certificate that the matrix is not ultrametric, and the algorithm outputs "reject".

Input: The parameters $n \in \mathbb{N}$ and $\varepsilon \in (0, 1)$, as well as query access to the entries of an unknown $n \times n$ matrix M from $\mathcal{C}$ (see Lemma 2.1).

Output: "accept" or "reject."

1. For $s = \Theta(\log(1/\varepsilon)/\varepsilon)$, take s random samples $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s \sim [n]$ drawn independently. Query $M(\boldsymbol{i}_\ell, \boldsymbol{i}_k)$ for all $\ell, k \in [s]$.
2. If there exists a triple among the sampled indices $\{\boldsymbol{i}_\ell, \boldsymbol{i}_k, \boldsymbol{i}_h\}$ which is a violating triple in M, output "reject."

Lemma 5.1. (Ultrametric Testing Algorithm Lemma) *For $n \in \mathbb{N}$ and $\varepsilon \in (0, 1)$, the algorithm* UltraTesting *receives as input an $n \times n$ matrix $M \in \mathcal{C}$ and a parameter ε and has the following guarantees:*

- *If $M \in \mathcal{P}^U$,* UltraTesting(M, ε) *always outputs "accept."*
- *If $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^U$, the algorithm outputs "reject" with probability at least* $2/3$.

The algorithm is non-adaptive, taking $O(\log(1/\varepsilon)/\varepsilon)$ samples and using $O(\log(1/\varepsilon)^2/\varepsilon^2)$ queries.

5.1.1 Ultrametric Testing Algorithm: Proof of Lemma 5.1 For the sake of analysis, we will divide the samples obtained in Line 1 of UltraTesting into two distinct groups.

- The first group consists of a set $\mathbf{S} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2}\}$ of the first $O(\log(1/\varepsilon)/\varepsilon)$ samples. As in [27], we use the set $\mathbf{S}$ to define a "skeleton partition" (Definition 3.5 in [27] and Definition 5.2 below).
- The second group consists of the remaining $O(\log(1/\varepsilon)/\varepsilon)$ samples viewed as a set of consecutive pairs, since the violations we find are formed by such some pairs $(\boldsymbol{i}_\ell, \boldsymbol{i}_{\ell+1})$ violating a constraint imposed by the skeleton partition of $\mathbf{S}$ in the first group.

We begin by defining the skeleton partition and corresponding equivalence classes formed by a fixed subset $S \in [n]$. Our analysis will track how the equivalence classes evolve as the (random) set changes with each sample.

Definition 5.2. (Skeleton Partition and Equivalence Classes) *Let $M \in \mathcal{C}$ and $S \subset [n]$. We say that S is consistent if the $|S| \times |S|$ submatrix $M_{|S\times S}$ encodes an ultrametric. For a consistent set S, we define:*

- **Consistent Points**. *A point $j \in [n] \setminus S$ is a consistent point if $S \cup \{j\}$ is consistent.*
- **Skeleton Partition**. *A skeleton partition of S is a partition $P_1, \ldots, P_\ell$ of the consistent points in $[n] \setminus S$ where two indices $j, k \in [n] \setminus S$ are in the same part P_p, or equivalence class, if and only if $M(j, i) = M(k, i)$ for all $i \in S$.*
- **Separator Set**. *If $j, k \in [n] \setminus S$ are in different equivalence classes, their separator set $\mathrm{SEP}(j, k) \subset S$ is given by*
$$\mathrm{SEP}(j, k) = \{i \in S : M(i, j) \neq M(i, k)\}.$$
- **Separator Corruption**. *If $j, k \in [n] \setminus S$ are in different equivalence classes, the pair (j, k) is a separator corruption if there exists $i \in \mathrm{SEP}(j, k)$ where*
$$M(j, k) \neq \max\{M(i, j), M(i, k)\}. \tag{5.1}$$
We let $\mathrm{SC}(M, S) \subset ([n] \setminus S) \times ([n] \setminus S)$ denote the set of pairs which are separator corruptions.

Definition 5.2 suggests our analysis approach. The algorithm will sample a set $\mathbf{S}$ of indices for the first group—if the set $\mathbf{S}$ is not consistent, there is a violation and we are already done. Assume that $\mathbf{S}$ is consistent. We aim to sample a pair $(\boldsymbol{j}, \boldsymbol{k})$ in the second group forming a separator corruption. Note that (5.1) forms a violation of the ultrametric property, so if this occurs, the algorithm can safely output "reject." We define one more type of "easy-to-detect" violation to the ultrametric property.

Definition 5.3. (Easy-to-Detect Corruption) *For $M \in \mathcal{C}$ and a consistent set $S \subset [n]$, let $(j, k) \in ([n] \setminus S) \times ([n] \setminus S)$ be in the same equivalence class. The pair (j, k) form an easy-to-detect corruption if there exists $i \in S$ where*
$$M(j, k) > M(j, i) = M(k, i).$$
Let $\mathrm{EC}(M, S) \subset ([n] \setminus S) \times ([n] \setminus S)$ be set of easy-to-detect corruptions in M with respect to S.

From the analysis perspective, we prove the following lemma, which implies the correctness of our algorithm. We state Lemma 5.2, as it will directly imply Lemma 5.1. The proof is a straightforward consequence of two lemmas (Lemma 5.3 and Lemma 5.4), which encapsulates our improvement over the analysis of [27].

Lemma 5.2. *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}^U$. Then, with probability at least $5/6$ over the draw of $\mathbf{S} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2}\} \subset [n]$ where $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2} \sim [n]$ and $s = O(\log(1/\varepsilon)/\varepsilon)$, one of the followings holds:*

- $\mathbf{S}$ *is not consistent;*
- *There are at least* $\varepsilon n/32$ *points in* $[n] \setminus \mathbf{S}$ *which are not consistent with* S*;*
- $|\mathrm{SC}(M, \mathbf{S})| + |\mathrm{EC}(M, \mathbf{S})| \geq \varepsilon n^2/8$.

Proof. [Proof of Lemma 5.1 assuming Lemma 5.2] First, if $M \in \mathcal{P}^U$, UltraTesting never observe a violation and outputs "accept." If $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^U$, we use Lemma 5.2 to deduce that with probability at least 5/6, the set $\mathbf{S} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2}\}$ obtained from the first $s/2$ samples in Line 1 is either already inconsistent (in which case among the entries of $\mathbf{S}$, there already is a violation), or it will become easy to sample a violation. Assume that $\mathbf{S}$ is consistent and that either there are $\varepsilon n/32$ inconsistent points in $[n] \setminus \mathbf{S}$, or $|\mathrm{SC}(M, \mathbf{S})| + |\mathrm{EC}(M, \mathbf{S})| \geq \varepsilon n^2/8$. We turn to the second group of $s/2$ samples, which are divided into pairs $(\boldsymbol{j}, \boldsymbol{k})$. We note that in order for a violation to be avoided, the $\Theta(\log(1/\varepsilon)/\varepsilon)$ pairs of samples must all avoid the $\varepsilon n/32$ inconsistent points, and, at the same time, avoid the $\varepsilon n^2/8$ entries in $\mathrm{SC}(M, \mathbf{S})$ and $\mathrm{EC}(M, \mathbf{S})$. The probability this occurs is at most

$$(1 - \Omega(\varepsilon))^{\Theta(\log(1/\varepsilon)/\varepsilon)} + (1 - \Omega(\varepsilon))^{\Theta(\log(1/\varepsilon)/\varepsilon)} \leq 1/6.$$

for an appropriate choice of the constant in front of the sample size.

Thus, by union bound over all events, the probability that the algorithm does not find any violation will be at most 1/3, which completes the proof. ∎

We turn our attention to proving Lemma 5.2, which will follow by lower-bounding the sizes of the separator corruptions $\mathrm{SC}(M, \mathbf{S})$ and easy-to-detect corruptions $\mathrm{EC}(M, \mathbf{S})$. Instead of analyzing $\mathrm{SC}(M, \mathbf{S})$ and $\mathrm{EC}(M, \mathbf{S})$ directly, we will look at a related quantity defined over certain types of equivalence classes.

Definition 5.4. (Easy, Versatile, and Active Parts) *Let* $M \in \mathcal{C}$ *be* ε*-far from* $\mathcal{P}^U$ *and* $S \subset [n]$ *be consistent, and let* $P_1, \ldots, P_\ell$ *denote the skeleton partition.*

- **Easy Part**. *A part* P_ℓ *is* easy *if*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{j}, \boldsymbol{k} \sim P_\ell} [(\boldsymbol{j}, \boldsymbol{k}) \in \mathrm{EC}(M, \mathbf{S})] \geq \frac{1}{2}.$$

  That is, at least half of the entries in the submatrix $M_{|P_\ell \times P_\ell}$ *are easy-to-detect corruptions.*

- **Versatile Part**. *Consider a part* P_ℓ *with* βn *indices for* $\beta = \Omega(1)$. *We say* P_ℓ *is* versatile *if the following hold:*
  - *there exists a* $(\beta n) \times (\beta n)$ *matrix* $M^*_{P_\ell}$ *encoding an ultrametric whose entries are at most* $\min_{i \in S, j \in P_l}\{M(i, j)\}$.
  - $\|M^*_{P_\ell} - M_{|P_\ell \times P_\ell}\| \leq \beta \cdot \varepsilon n^2/2$.

  That is, intuitively, no more than $\beta \cdot \varepsilon n^2/2$ *entries of the submatrix* $M_{|P_\ell \times P_\ell}$ *need to be modified so that* $M_{|P_\ell \times P_\ell}$ *is fixed to be an ultrametric consistent with* S.

- **Active Part**. *A part* P_ℓ *is* active *if it is neither easy nor versatile. We let*

$$A(M, S) = \{(j, k) \in P_\ell \text{ for some active part } P_\ell\}$$
$$\alpha(M, S) = \mathop{\mathbf{Pr}}_{\boldsymbol{j} \sim [n]} [\boldsymbol{j} \in P_\ell \text{ for some active part } P_\ell]$$

LEMMA 5.3. *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}^U$ and $S \subset [n]$ be consistent. If $|A(M,S)| \leq \varepsilon n^2/8$, then either there are at least $\varepsilon n/32$ inconsistent points in $[n] \setminus S$, or $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \geq \varepsilon n^2/8$.*

Proof. We prove the contra-positive by showing that, given M being ε-far from $\mathcal{P}^U$ and a consistent subset $S \subset [n]$ which satisfy (i) $|A(M,S)| \leq \varepsilon n^2/8$, (ii) $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \leq \varepsilon n^2/8$, and (iii) at most $\varepsilon n/32$ inconsistent points in $[n] \setminus S$, there exists an $n \times n$ matrix $\tilde{M} \in \mathcal{P}^U$ which differs from M on fewer than εn^2 entries. This implies that M is ε-close to $\mathcal{P}^U$, which is a contradiction. The matrix $\tilde{M}$ is constructed as follows:

1. For all $i, i' \in S$ we set $\tilde{M}(i,i')$ to $M(i,i')$.
2. If $j \in [n] \setminus S$ is among the $\varepsilon n/32$ inconsistent points, we let $\tilde{M}(j,k)$ be some arbitrarily large value for all $k \in [n]$. This corresponds to effectively removing j while keeping an ultrametric. For the points $j \in [n] \setminus S$ which are consistent, set $\tilde{M}(j,i)$ to be $M(j,i)$ for all $i \in S$.
3. For consistent $j, k \in [n] \setminus S$ lying in different equivalence classes:
   - If $(j,k) \notin \mathrm{SC}(M,S)$, set $\tilde{M}(j,k) = M(j,k)$.
   - If $(j,k) \in \mathrm{SC}(M,S)$, find $i \in \mathrm{SEP}(j,k)$ and set $\tilde{M}(j,k) = \max\{M(i,j), M(i,k)\}$. (As we will see, the specific choice of i will not matter).
4. For consistent $j, k \in [n] \setminus S$ lying in the same equivalence class P_ℓ:
   - If P_ℓ is an Easy Part or an Active Part, set $\tilde{M}(j,k)$ to be the minimum positive entry in M.
   - If P_ℓ is a Versatile Part, find the matrix $M^*_{P_\ell}$ as in Definition 5.4, and set $\tilde{M}(j,k) = M^*_{P_\ell}(j,k)$.

We now show that $\|\tilde{M} - M\|_0 < \varepsilon n^2$ and $\tilde{M} \in \mathcal{P}^U$. First, note that M and $\tilde{M}$ differ on at most

$$\frac{\varepsilon n}{32} \times 2n + |\mathrm{SC}(M,S)| + \sum_{P_\ell \text{ Easy}} |P_\ell|^2 + |A(M,S)| + \sum_{P_\ell \text{ Versatile}} |P_\ell| \cdot \varepsilon n/2$$

entries. Notice that, $2|\mathrm{EC}(M,S)|$ upper bounds the sum of $|P_\ell|^2$ over Easy Part P_ℓ (since each Easy Part P_ℓ has at least half of its entries in $\mathrm{EC}(M,S)$. Furthermore, parts partition consistent points in $[n] \setminus \mathbf{S}$, so the sum of $|P_\ell|$ over Versatile parts is at most n. Finally, the assumption $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \leq \varepsilon n^2/8$ as well as $|A(M,S)| \leq \varepsilon n^2/8$ give the desired bound: $\|\tilde{M} - M\|_0 \leq 15/16 \cdot \varepsilon n^2$.

The rest of the proof which shows that $\tilde{M} \in \mathcal{P}^U$ has been moved to the appendix, in the interest of space. We prove $\tilde{M} \in \mathcal{P}^U$ by considering an arbitrary triple $\{i,j,k\} \subset [n]$ and showing that it does not form a violation in $\tilde{M}$.

The above argument shows that $\tilde{M}$ encodes an ultrametric and is $(15/16 \cdot \varepsilon)$-close to M, which leads to a contradiction. Hence, if $|A(M,S)| \leq \varepsilon n^2/8$, then either there are at least $\varepsilon n^2/32$ inconsistent points in $[n] \setminus S$ or $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \geq \varepsilon n^2/8$.

∎

LEMMA 5.4. *Suppose $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^U$. With probability at least $5/6$, over the draw of $\mathbf{S} \subset [n]$ of size $\Omega(\log(1/\varepsilon)/\varepsilon)$, $|A(M,\mathbf{S})| \leq \varepsilon n^2/8$.*

Proof. [Proof of Lemma 5.2 assuming Lemmas 5.4] The proof is a straightforward application of Lemma 5.4 and Lemma 5.3. In particular, by Lemma 5.4, a draw of $\mathbf{S}$ will satisfy $|A(M,\mathbf{S})| \leq \varepsilon n^2/8$ with probability at least $5/6$. Then, either $\mathbf{S}$ is inconsistent, or we may apply Lemma 5.3, which implies either there are at least $\varepsilon n/32$ indices which incur violations with S or $|\mathrm{SC}(M,\mathbf{S})| + |\mathrm{EC}(M,\mathbf{S})| \geq \varepsilon n^2/8$. ∎

5.1.2 Bounding the Number of Active Entries: Proof of Lemma 5.4

LEMMA 5.5. *Suppose $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^U$ and suppose $S = \{i_1, ..., i_{j-1}\} \subset [n]$. Over the randomness of the index $\boldsymbol{i}_j$ sampled from $[n]$, $\mathbf{E}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})|] \leq |A(M,S)| \cdot \exp(-\frac{\varepsilon}{16})$.*

Proof. [Proof of Lemma 5.4 assuming Lemma 5.5]

Suppose $\boldsymbol{i}_1, ..., \boldsymbol{i}_s \in [n]$ are drawn independently where $s = 16\log(48/\varepsilon)/\varepsilon$. Let $\mathbf{S}$ be this set of random samples. Then the law of total expectation,

$$\begin{aligned}
\log\left(\mathop{\mathbf{E}}_{\boldsymbol{i}_1,...,\boldsymbol{i}_s}[|A(M,\mathbf{S})|]\right) &= \log\left(\mathbf{E}\left[|A(M, \{\boldsymbol{i}_1, ..., \boldsymbol{i}_{s-1}\} \cup \{\boldsymbol{i}_s\})|\right]\right) \\
&= \log\left(\mathbf{E}\left[\mathop{\mathbf{E}}_{\boldsymbol{i}_s}[|A(M, \{\boldsymbol{i}_1, ..., \boldsymbol{i}_{s-1}\} \cup \{\boldsymbol{i}_s\})||\boldsymbol{i}_1, ..., \boldsymbol{i}_{s-1}]\right]\right) \\
&\leq \log\left(\mathbf{E}\left[|A(M, \{\boldsymbol{i}_1, ..., \boldsymbol{i}_{s-1}\})| \cdot \exp(-\frac{\varepsilon}{16})\right]\right) \\
&= \log\left(\mathop{\mathbf{E}}_{\boldsymbol{i}_1,...,\boldsymbol{i}_{s-1}}[|A(M, \{\boldsymbol{i}_1, ..., \boldsymbol{i}_{s-1}\})|]\right) - \frac{\varepsilon}{16}.
\end{aligned}$$

By induction,

$$\log\left(\mathop{\mathbf{E}}_{\boldsymbol{i}_1,...,\boldsymbol{i}_s}[|A(M,\mathbf{S})|]\right) = \log(\mathbf{E}[|A(M,\emptyset)|]) - \frac{\varepsilon \cdot s}{16} = \log(n^2) - \log(48/\varepsilon) = \log(\varepsilon n^2/48).$$

Thus, $\mathbf{E}_{\boldsymbol{i}_1,...,\boldsymbol{i}_s}[|A(M,\mathbf{S})|] = \varepsilon n^2/48$. By Markov's inequality, with probability at least 5/6, over the draw of $\{\boldsymbol{i}_1, ..., \boldsymbol{i}_s\}$, $A(|M,\mathbf{S}|) \leq \varepsilon n^2/8$.

□

LEMMA 5.6. *Suppose $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^U$. Suppose $S = \{i_1, ..., i_{j-1}\} \subset [n]$. Over the randomness of the next sampled index $\boldsymbol{i}_j$, if $\boldsymbol{i}_j$ is sampled from an Active Part P with size βn,*

$$\mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in P] \leq |A(M,S)| - \beta \cdot \varepsilon n^2/16.$$

Proof. [Proof of Lemma 5.5 assuming Lemma 5.6]

Let $\mathcal{A}(M,S)$ be the set of all Active Parts defined over the set S. Suppose the part P_l with $\beta_l n$ points is active. A point is chosen from this part with probability β_l, and as a result, $\mathbf{E}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in P_l] \leq |A(M,S)| - \beta_l \cdot \varepsilon n^2/16$. Thus, we have the conditional expectation

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in \mathcal{A}(M,S)] &= \sum_{l:P_l \in \mathcal{A}(M,S)} \mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in P_l] \cdot \mathbf{Pr}(\boldsymbol{i}_j \in P_l | \boldsymbol{i}_j \in \mathcal{A}(M,S)) \\
&\leq |A(M,S)| - \sum_{l:P_l \in \mathcal{A}(M,S)} \beta_l \cdot \varepsilon n^2/16 \cdot \frac{\beta_l}{\sum_{k:P_k \in \mathcal{A}(M,S)} \beta_k} \\
&= |A(M,S)| - \frac{\sum_{l:P_l \in \mathcal{A}(M,S)} \beta_l^2 \cdot \varepsilon n^2/16}{\sum_{l:P_l \in \mathcal{A}(M,S)} \beta_l}
\end{aligned}$$

On the other hand, we have

$$|A(M,S)| = \sum_{l:P_l \in \mathcal{A}(M,S)} (\beta_l n)^2,$$

$$\alpha(M,S) = \mathop{\mathbf{Pr}}_{\boldsymbol{j}\sim[n]}[\boldsymbol{j} \in P_\ell \text{ for some active part } P_\ell] = \sum_{l:P_l\in\mathcal{A}(M,S)} \beta_l.$$

Hence, as index $\boldsymbol{i}_j$ is sampled from some Active Part, $\mathbf{E}_{\boldsymbol{i}_j}[|A(M,S\cup\{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in \mathcal{A}(M,S)] \le |A(M,S)| - |A(M,S)| \cdot \frac{\varepsilon}{16\cdot\alpha(M,S)}$. In general when index $\boldsymbol{i}_j$ is sampled uniformly at random,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M,S\cup\{\boldsymbol{i}_j\})|] \\
=&\mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M,S\cup\{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in \mathcal{A}(M,S)] \cdot \mathbf{Pr}(\boldsymbol{i}_j \in \mathcal{A}(M,S)) \\
&+ \mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M,S\cup\{\boldsymbol{i}_j\})||\boldsymbol{i}_j \notin \mathcal{A}(M,S)] \cdot \mathbf{Pr}(\boldsymbol{i}_j \notin \mathcal{A}(M,S)) \\
\le&\left(|A(M,S)| - |A(M,S)| \cdot \frac{\varepsilon}{16\cdot\alpha(M,S)}\right)\cdot\alpha(M,S) + |A(M,S)|\cdot(1-\alpha(M,S)) \\
=&|A(M,S)|\cdot\left(1-\frac{\varepsilon}{16}\right) \le |A(M,S)|\cdot\exp\left(-\frac{\varepsilon}{16}\right)
\end{aligned}$$

□

Proof. [Proof of Lemma 5.6] The Active Part P corresponds to a sub-square-matrix $B = M_{|P\times P}$ of M. Consider all the entries in B that are not in $\mathrm{EC}(M,S)$. Suppose the set of values these entries take on is $\{v_1, v_2, ..., v_k\}$, and further suppose v_1 is the value that most entries take on. Suppose $\{v_{k+1}, ..., v_l\}$ is the set of values the entries in $\mathrm{EC}(M,S)\cap B$ take on. Following the definition of easy-to-detect corruption $\mathrm{EC}(M,S)$, any value in $\{v_{k+1}, ..., v_l\}$ is larger than any value in $\{v_1, v_2, ..., v_k\}$, so $\{v_1, ..., v_l\}$ is a set of non-repetitive values. Let r_τ to denote the number of entries taking on value v_τ in the block matrix B. Hence, $\sum_{\tau=1}^{l} r_\tau = (\beta n)^2$. The assumption indicates that $r_1 \ge r_\tau, \forall\tau\in\{2,...,k\}$, and $\sum_{\tau=k+1}^{l} r_\tau$ is the number of easy-to-detect-corruptions in B. That is,

$$|B\cap\mathrm{EC}(M,S)| = \sum_{\tau=k+1}^{l} r_\tau.$$

Since P is an Active Part, it is not an Easy Part. By definition, no more than half of the entries in B are in $\mathrm{EC}(M,S)$. Hence, $\sum_{\tau=k+1}^{l} r_\tau$ is no larger than $(\beta n)^2/2$.

Now, focus on the row/column in B that corresponds to index $\boldsymbol{i}_j$ in P. Suppose in the row, there are $r_l^{(\boldsymbol{i}_j)}$ entries whose values are v_l. Then, once $\boldsymbol{i}_j$ is sampled, some pairs of indices are separated into different parts. In particular, for indices $x, y \in P$, if $M(\boldsymbol{i}_j, x) \ne M(\boldsymbol{i}_j, y)$, the pair x, y that previously both belonged to part P now fall into two different parts. The number of such pairs of indices that are separated is $\sum_{\tau=1}^{l} r_\tau^{(\boldsymbol{i}_j)}(\beta n - r_\tau^{(\boldsymbol{i}_j)})/2$. Over the randomness of sampling any index $\boldsymbol{i}$ in P, the expected number of separated pairs created is

$$\frac{1}{\beta n}\sum_{\boldsymbol{i}\in P}\sum_{\tau=1}^{l} r_\tau^{(\boldsymbol{i})}(\beta n - r_\tau^{(\boldsymbol{i})})/2 = \frac{1}{2\beta n}\sum_{\tau=1}^{l}\sum_{\boldsymbol{i}\in P} r_\tau^{(\boldsymbol{i})}(\beta n - r_\tau^{(\boldsymbol{i})}).$$

Notice here that the number of separated pairs is exactly the number of entries moving from the block matrix B to off-diagonal, and these entries no longer belong to Active Parts. Therefore, $|A(M,S)| - |A(M,S\cup\{\boldsymbol{i}\})|$ is at least the number of separated pairs created by sampling an index $\boldsymbol{i}\in P$.

Hence, the goal is to find a lower bound for the number of separated pairs created by an index $\boldsymbol{i}\in P$. Since a pair of indices are separated if their distance value to the sampled index $\boldsymbol{i}$ is different, we focus on the values in the matrix. Imagine we have a graph with βn vertices, and two vertices $\boldsymbol{i}_x, \boldsymbol{i}_y$ are connected by an edge if and only if the entry $(\boldsymbol{i}_x, \boldsymbol{i}_y)$ in B equals v_τ. Then there are $r_\tau/2$ edges in this graph, and the degree of a

vertex $\boldsymbol{i}$ is $r_\tau^{(\boldsymbol{i})}$. Thus,

$$\sum_{\boldsymbol{i}\in P} r_\tau^{(\boldsymbol{i})}(\beta n - r_\tau^{(\boldsymbol{i})}) = \beta n \cdot \sum_{\boldsymbol{i}\in P} r_\tau^{(\boldsymbol{i})} - \sum_{\boldsymbol{i}\in P}(r_\tau^{(\boldsymbol{i})})^2 = \beta n \cdot r_\tau - \sum_{\boldsymbol{i}\in P}(r_\tau^{(\boldsymbol{i})})^2.$$

De Caen [18] gives an upper bound on the sum of squared degrees, that

$$\sum_{\boldsymbol{i}\in P}(r_\tau^{(\boldsymbol{i})})^2 \le E(2E/(V-1)+V-2) \approx r_\tau^2/(2\beta n) + r_\tau \cdot \beta n/2.$$

On the other hand, since $r_1 \ge r_\tau, \forall \tau \in \{2,...,k\}$, an upper bound for $\sum_{\tau=1}^{l} r_\tau^2$ is

$$\sum_{\tau=1}^{l} r_\tau^2 \le \left(\sum_{\tau=k+1}^{l} r_\tau\right)^2 + \left((\beta n)^2 - \sum_{\tau=k+1}^{l} r_\tau\right) r_1,$$

where, recall that, the quantity $\sum_{\tau=k+1}^{l} r_\tau$ is the total number of easy-to-detect corruptions in B. To sum up, $|A(M,S)| - |A(M, S\cup\{\boldsymbol{i}\})|$ is at least the number of pairs being separated. On the other hand, in expectation over the randomness of the sampled index $\boldsymbol{i} \in P$, the number of pairs being separated is at least

$$\begin{aligned}
&\frac{1}{2\beta n}\sum_{\tau=1}^{l}\sum_{\boldsymbol{i}\in P} r_\tau^{(\boldsymbol{i})}(\beta n - r_\tau^{(\boldsymbol{i})}) \\
\ge &\frac{1}{2\beta n}\sum_{\tau=1}^{l} r_\tau \cdot \beta n - \frac{r_\tau^2}{2\beta n} - r_\tau \cdot \beta n/2 \\
= &\frac{1}{2\beta n}(\beta n)^3 - \frac{1}{4(\beta n)^2}\sum_{\tau=1}^{l} r_\tau^2 - \frac{(\beta n)^2}{4} \\
\ge &\frac{(\beta n)^2}{4} - \frac{1}{4(\beta n)^2}\left(\left(\sum_{\tau=k+1}^{l} r_\tau\right)^2 - r_1 \cdot \sum_{\tau=k+1}^{l} r_\tau\right) - \frac{r_1}{4}
\end{aligned}$$

To further bound this quantity, we consider two cases for the size of r_1, i.e. the number of entries taking on value v_1.

- CASE $r_1 \le (\beta n)^2/2$: Following the assumption that P is not an Easy Part, $\sum_{\tau=k+1}^{l} r_\tau$, which equals $|\mathrm{EC}(M,S) \cap B|$, shall not exceed $|B|/2 = (\beta n)^2/2$. Thus, the above expected number of separated pairs is at least $(\beta n)^2/4 - (\beta n)^4/(16(\beta n)^2) - r_1/4 = 3(\beta n)^2/16 - r_1/4$. As a consequence,

$$8 \cdot \left(|A(M,S)| - \mathop{\mathbf{E}}_{\boldsymbol{i}}[|A(M, S\cup\{\boldsymbol{i}\})||\boldsymbol{i}\in P]\right) \ge 3(\beta n)^2/2 - 2r_1 \ge (\beta n)^2 - r_1.$$

- CASE $r_1 > (\beta n)^2/2$: Notice that $\sum_{\tau=k+1}^{l} r_\tau \le (\beta n)^2 - r_1 < (\beta n)^2/2$. In this case, the above expected number of separated pairs is minimized when $\sum_{\tau=k+1}^{l} r_\tau$ is 0 or $(\beta n)^2 - r_1$ by convexity. As a result, $|A(M,S)| - \mathbf{E}_{\boldsymbol{i}}[|A(M, S\cup\{\boldsymbol{i}\})||\boldsymbol{i}\in P] \ge \max\{(\beta n)^2/4 - r_1/4, r_1/2\}$. Again, it implies that

$$8 \cdot \left(|A(M,S)| - \mathop{\mathbf{E}}_{\boldsymbol{i}}[|A(M, S\cup\{\boldsymbol{i}\})||\boldsymbol{i}\in P]\right) \ge (\beta n)^2 - r_1.$$

Now we show by contradiction that, since P is not a Versatile Part, $\left(\sum_{\tau=2}^{l} r_\tau\right) = (\beta n)^2 - r_1$ must be at least $\beta \cdot \varepsilon n^2/2$.

CLAIM 5.1. *Since P is an Active Part, $\sum_{\tau=2}^{l} r_\tau = (\beta n)^2 - r_1 \geq \beta \cdot \varepsilon n^2/2$.*

Proof. Suppose $\sum_{\tau=2}^{l} r_\tau = (\beta n)^2 - r_1 < \beta \cdot \varepsilon n^2/2$. Create a $(\beta n) \times (\beta n)$ matrix M_P^* where diagonal entries are 0 and every other entry is set to v_1. Then, $\|M_P^* - B\| \leq \sum_{\tau=2}^{l} r_\tau < \beta \cdot n^2/2$. Moreover, M_P^* encodes an ultrametric. Lastly, as the value v_1 is taken on by some entry $M(u, u') \notin \mathrm{EC}(M, S)$, for any index $i \in S$, $M(i, u) = M(i, u') \geq M(u, u') = v_1$. Moreover, for any $u'' \in P$, $M(i, u) = M(i, u'')$. Thus, $v_1 \leq \min_{i \in S, u \in P}\{M(i, u)\}$. Therefore, M_P^* satisfies all conditions that make P a Versatile Part. However, because P is an Active Part, by contradiction, we must have $\sum_{\tau=2}^{l} r_\tau = (\beta n)^2 - r_1 \geq \beta \cdot \varepsilon n^2/2$. ∎

As a summary for the above arguments, we have

$$8 \cdot \left(|A(M, S)| - \mathop{\mathbf{E}}_{\boldsymbol{i}}[|A(M, S \cup \{\boldsymbol{i}\})| | \boldsymbol{i} \in P] \right) \geq (\beta n)^2 - r_1 \geq \beta \cdot \varepsilon n^2/2.$$

∎

5.2 Tree Metric Testing Upper Bound In this section, we finish the proof of Theorem 1.3 by presenting a tree metric testing algorithm TREETESTING. We first define the type of violation the algorithm seeks.

DEFINITION 5.5. (TREE METRIC VIOLATING QUADRUPLE) *For $i, j, k, l \in [n]$, the quadruple $\{i, j, k, l\}$ is a violation for tree metrics in $M \in \mathcal{C}$ if, after re-naming so $M(i, j) + M(k, l)$ is the maximum among the three matchings*

$$\{M(i, j) + M(k, l), M(i, k) + M(j, l), M(i, l) + M(j, k)\},$$

$$M(i, j) + M(k, l) > M(i, k) + M(j, l)) \text{ and } M(i, j) + M(k, l) > M(i, l) + M(j, k).$$

The testing algorithm TREETESTING below is very similar to the ultrametric tester we presented earlier. The analysis follows an analogous path.

TreeTesting Algorithm: The algorithm aims to find a violating quadruple of indices $\{i, j, k, l\}$ to the tree metric property Definition 5.5. If such a quadruple is found, it constitutes a certificate that the matrix is not a tree metric, and the algorithm outputs "reject".

Input: The parameters $n \in \mathbb{N}$ and $\varepsilon \in (0, 1)$, as well as query access to the entries of an unknown $n \times n$ matrix M from $\mathcal{C}$ (see Lemma 2.1).

Output: "accept" or "reject"

1. For $s = O(\log(1/\varepsilon)/\varepsilon)$, take s random samples $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_s \sim [n]$ drawn independently. Query $M(\boldsymbol{i}_\ell, \boldsymbol{i}_k)$ for all $\ell, k \in [s]$.
2. If there exists a violating quadruple among the sampled indices $\{\boldsymbol{i}_a, \boldsymbol{i}_b, \boldsymbol{i}_c, \boldsymbol{i}_d\}$, output "reject."

LEMMA 5.7. (TREE METRIC TESTING ALGORITHM LEMMA) *For $n \in \mathbb{N}$ and $\varepsilon \in (0, 1)$, there exists a randomized algorithm, TREETESTING, which receives as input an $n \times n$ matrix $M \in \mathcal{C}$ and a parameter ε and has the following guarantees:*

- *If $M \in \mathcal{P}^T$, TREETESTING(M, ε) always outputs "accept."*

- *If $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^T$, the algorithm outputs "reject" with probability at least 2/3.*

The algorithm is non-adaptive, taking $O(\log(1/\varepsilon)/\varepsilon)$ samples and using $O(\log(1/\varepsilon)^2/\varepsilon^2)$ queries.

Proof. [Proof of Theorem 1.3 with Lemma 5.1 and assuming Lemma 5.7] By Lemma 2.1, it suffices to show Theorem 1.3 holds for $M \in \mathcal{C}$. For $M \in \mathcal{P}^T$ (respectively $M \in \mathcal{P}^U$), TreeTesting(M, ε) (resp. UltraTesting(M, ε)) outputs "accept" with probability 1, so the algorithm outputs "accept" if M encodes a tree metric space (resp. ultrametric space). If M is ε-far from $\mathcal{P}^T$ (resp. ε-far from $\mathcal{P}^U$), the TreeTesting(M, ε) algorithm (resp. UltraTesting(M, ε) algorithm) outputs "reject" with probability at least 2/3. Both algorithms are non-adaptive with one-sided error and sample complexity $O(\log(1/\varepsilon)/\varepsilon)$ and query complexity $O(\log(1/\varepsilon)^2/\varepsilon^2)$.

□

Again, we divide the samples selected in Line 1 in TreeTesting into two distinct groups for analysis purposes.

- The first group consists of a set $\mathbf{S} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2}\}$ of the first $O(\log(1/\varepsilon)/\varepsilon)$ samples. Again, we use the set $\mathbf{S}$ to define a "skeleton partition" (Definition 3.5 in [27] and Definition 5.6 below).
- The second group consists of the remaining $O(\log(1/\varepsilon)/\varepsilon)$ samples, which are divided into pairs $(\boldsymbol{i}_\ell, \boldsymbol{i}_{\ell+1})$ of indices. The violations we find are formed by a pair $(\boldsymbol{i}_\ell, \boldsymbol{i}_{\ell+1})$ among the second group which violates a constraint imposed by the skeleton partition of $\mathbf{S}$ in the first group.

As before, the first group of samples $\mathbf{S}$ puts structural constraints on the remaining points by partitioning them into equivalence classes. The notions of separator corruptions and easy-to-detect corruptions as well as parts follow in a very similar fashion.

Definition 5.6. (Skeleton Partition and Equivalence Classes) *Let $M \in \mathcal{C}$ and $S \subset [n]$. We say that S is consistent if the $|S| \times |S|$ submatrix $M_{|S \times S}$ encodes a tree metric. For a consistent set S, we let:*

- **Consistent Points**. *A point $j \in [n] \setminus S$ is a consistent point if $S \cup \{j\}$ is consistent.*
- **Skeleton Partition**. *A skeleton partition of S is a partition $P_1, \ldots, P_\ell$ of consistent points in $[n] \setminus S$ where two points $i, j \in [n] \setminus S$ are in the same part P_ℓ, or equivalence class, if and only if $M(i,u) - M(i,v) = M(j,u) - M(j,v)$ for all $u, v \in S$.*
- **Separator Set**. *If $i, j \in [n] \setminus S$ are in different equivalence classes, their separator set $\mathrm{SEP}(i,j) \subset S$ is given by*
$$\mathrm{SEP}(i,j) = \left\{(u,v) \in S^2 : M(i,u) - M(i,v) \neq M(j,u) - M(j,v)\right\}.$$
- **Separator Corruption**. *If $i, j \in [n] \setminus S$ are in different equivalence classes defined over S, the pair (i,j) is a separator corruption if there exists $(u,v) \in \mathrm{SEP}(i,j)$ where*
$$M(i,j) + M(u,v) \neq \max\{M(i,u) + M(j,v), M(i,v) + M(j,u)\}.$$
We let $\mathrm{SC}(M,S) \subset ([n] \setminus S) \times ([n] \setminus S)$ denote the set of pairs that are separator corruptions.

The equation in the Skeleton Partition definition conveys the facts that, out of the three matchings induced by i, j, u, v, the matching $M(i,u) + M(j,v)$ equals the matching $M(i,v) + M(j,u)$. In order for the four points to obey the four-point condition, we only need $M(i,j) + M(u,v)$ to be no more than this quantity. On the other hand, if two points i, j fall into different classes, then $M(i,j)$ can be deduced exactly using the separator

of pair (i, j). The definition of Separator Corruption indicates that if $M(i, j)$ does not equal this deducted quantity for some (u, v) in the separator set of (i, j), the quadruple $\{i, j, u, v\}$ is a violation.

The idea for the algorithm is similar to the ultrametric testing algorithm. If the first group of samples $\mathbf{S}$ is not consistent, then there is already a violation in the first group and we are done. If $\mathbf{S}$ is consistent, on the other hand, we use $\mathbf{S}$ to make a skeleton partition on the remaining points. We aim to sample a separator corruption $(\boldsymbol{i}, \boldsymbol{j})$ which forms a violation with $\mathbf{S}$. There is, as before, another type of corruption, the easy-to-detect corruption.

Definition 5.7. (Easy-to-Detect Corruption) *For $M \in \mathcal{C}$ and a consistent set $S \subset [n]$, let $(i, j) \in ([n] \setminus S) \times ([n] \setminus S)$ be in the same equivalence class. The pair (i, j) form an easy-to-detect corruption if there exists $u, v \in S$ where*

$$M(i, j) + M(u, v) > M(i, u) + M(j, v) = M(i, v) + M(j, u).$$

Let $\mathrm{EC}(M, S) \subset ([n] \setminus S) \times ([n] \setminus S)$ be set of easy-to-detect corruptions in M with respect to S.

Now, we formalize the idea that after sampling $\mathbf{S}$, we aim to find many separator corruptions and easy-to-detect corruptions, which act as certificates that M is not a tree metric. The below Lemma 5.8 is an exact replicate of Lemma 5.2 in the ultrametric section. Thus, the proof of Tree Metric Testing Lemma 5.7 assuming Lemma 5.8 is exactly as the proof of Ultrametric Testing Lemma 5.1 using Lemma 5.2.

Lemma 5.8. *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}^T$. Then, with probability at least $5/6$ over the draw of $\mathbf{S} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2}\} \subset [n]$ where $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{s/2} \sim [n]$ and $s = O(\log(1/\varepsilon)/\varepsilon)$, one of the followings holds:*

- $\mathbf{S}$ *is not consistent;*
- *There are at least $\varepsilon n/32$ inconsistent points in $[n] \setminus \mathbf{S}$;*
- $|\mathrm{SC}(M, \mathbf{S})| + |\mathrm{EC}(M, \mathbf{S})| \geq \varepsilon n^2/8$.

Proof. [Proof of Lemma 5.7 assuming Lemma 5.8] The proof is exactly the same as that of Lemma 5.1. ∎

To show Lemma 5.8, we re-use the ideas of categorizing different parts and analyze the dynamics between the parts.

Definition 5.8. (Easy, Versatile, and Active Parts) *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}^T$ and $S \subset [n]$ be consistent, and let $P_1, \ldots, P_\ell$ denote the skeleton partition.*

- **Easy Part**. *A part P_ℓ is* easy *if*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{j}, \boldsymbol{k} \sim P_\ell} [(\boldsymbol{j}, \boldsymbol{k}) \in \mathrm{EC}(M, \mathbf{S})] \geq \frac{1}{2}.$$

- **Versatile Part**. *Consider a part P_ℓ with βn indices (for $\beta > 0$). We say P_ℓ is* versatile *if the following hold:*
  - *there exists a $(\beta n) \times (\beta n)$ matrix $M^*_{P_\ell}$ such that the metric on subset $S \cup P_\ell$ defined by*

$$D(i, j) = \begin{cases} M(i, j) & \text{if } i \text{ or } j \text{ or both in } S \\ M^*_{P_\ell}(i, j) & \text{if } i, j \in P_\ell \end{cases}$$

  is a tree metric

– $\|M^*_{P_\ell} - M_{|P_\ell \times P_\ell}\| \leq \beta \cdot \varepsilon n^2/2$.

That is, intuitively, no more than $\beta \cdot \varepsilon n^2/2$ entries of the sub-block-matrix $M_{|P_\ell \times P_\ell}$ need to be modified so that $M_{|P_\ell \times P_\ell}$ is fixed into a tree metric consistent with S.

- **Active Part**. *A part P_ℓ is* active *if it is neither easy nor versatile. We let*

$$A(M,S) = \{(j,k) \in P_\ell \text{ for some active part } P_\ell\}$$
$$\alpha(M,S) = \mathop{\mathbf{Pr}}_{\boldsymbol{j} \sim [n]}[\boldsymbol{j} \in P_\ell \text{ for some active part } P_\ell]$$

LEMMA 5.9. *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}^T$ and $S \subset [n]$ be consistent. If $|A(M,S)| \leq \varepsilon n^2/8$, then either there are at least $\varepsilon n/32$ inconsistent points in $[n] \setminus S$, or $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \geq \varepsilon n^2/8$.*

Proof. We prove the contra-positive. That is, given $M \in \mathcal{C}$ and a consistent $S \subset [n]$ which satisfies (i) $|A(M,S)| \leq \varepsilon n^2/8$, (ii) $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \leq \varepsilon n^2/8$, and (iii) at most $\varepsilon n/32$ inconsistent points in $[n] \setminus S$, there exists an $n \times n$ matrix $\tilde{M} \in \mathcal{P}^T$ that differs from M on fewer than εn^2 entries. The matrix $\tilde{M}$ is constructed as follows:

1. For all $i, i' \in S$ we set $\tilde{M}(i,i')$ to $M(i,i')$.
2. If $j \in [n] \setminus S$ is among the $\varepsilon n/32$ inconsistent points, we let $\tilde{M}(j,k)$ be arbitrarily large for all $k \in [n]$, which corresponds to effectively removing j while keeping an tree metric. The points $j \in [n] \setminus S$ which are consistent have $\tilde{M}(j,i)$ be set to $M(j,i)$ for $i \in S$.
3. For $j, k \in [n] \setminus S$ which are consistent and lie in different equivalence classes:
   - If $(j,k) \notin \mathrm{SC}(M,S)$, set $\tilde{M}(j,k) = M(j,k)$.
   - If $(j,k) \in \mathrm{SC}(M,S)$, find $(u,v) \in \mathrm{SEP}(j,k)$ and set $\tilde{M}(j,k) = \max\{M(i,u) + M(j,v), M(i,v) + M(j,u)\} - M(u,v)$. (As we will see, the specific choice of (u,v) will not matter).
4. For $j, k \in [n] \setminus S$ which are consistent and in the same equivalence class P_ℓ:
   - If P_ℓ is an Easy Part, find a $u \in S$ and set $\tilde{M}(j,k) = M(u,j) + M(u,k) - \min_{j',k' \in P_\ell}\{M(u,j') + M(u,k')\}$.
   - If P_ℓ is an Active Part, find a $u \in S$ and set $\tilde{M}(j,k) = M(u,j) + M(u,k) - \min_{j',k' \in P_\ell}\{M(u,j') + M(u,k')\}$.
   - If P_ℓ is a Versatile Part, find the matrix $M^*_{P_\ell}$ as in Definition 5.8, and set $\tilde{M}(j,k) = M^*_{P_\ell}(j,k)$.

We now show that $\|\tilde{M} - M\|_0 < \varepsilon n^2$. By the same argument as in Lemma 5.3 in ultrametric testing, M and $\tilde{M}$ differ on at most

$$\frac{\varepsilon n}{32} \times 2n + |\mathrm{SC}(M,S)| + \sum_{P_\ell \text{ easy}} |P_\ell|^2 + |A(M,S)| + \sum_{P_\ell \text{ versatile}} |P_\ell| \cdot \varepsilon n/2 \leq 15/16 \cdot \varepsilon n^2$$

entries. With similar ideas as in Lemma 5.3, any quadruple $\{i,j,k,l\}$ does not form a violation in $\tilde{M}$, which makes $\tilde{M}$ a tree metric. Since M is ε-far from $\mathcal{P}^T$, we must have either at least $\varepsilon n/32$ inconsistent indices in $[n] \setminus S$ or $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \geq \varepsilon n^2/8$.

∎

LEMMA 5.10. *Suppose $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^T$. With probability at least $5/6$, over the draw of $\mathbf{S} \subset [n]$ of size at least $\tilde{\Omega}(1/\varepsilon)$, $|A(M,\mathbf{S})| \leq \varepsilon n^2/8$.*

Proof. [Proof of Lemma 5.8 assuming Lemma 5.10] The proof is exactly the same as the proof of Lemma 5.2. ∎

LEMMA 5.11. *Suppose $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^T$. Suppose $S = \{i_1, ..., i_{j-1}\} \subset [n]$. Over the randomness of the index $\boldsymbol{i}_j$ sampled from $[n]$, $\mathbf{E}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})|] \leq |A(M, S)| \cdot \exp(-\frac{\varepsilon}{16})$.*

Proof. [Proof of Lemma 5.10 assuming Lemma 5.11] The proof is exactly the same as the proof of Lemma 5.4. ∎

LEMMA 5.12. *Suppose $M \in \mathcal{C}$ is ε-far from $\mathcal{P}^T$. Suppose $S \subset [n]$. Over the randomness of the next sampled index $\boldsymbol{i}_j$, if $\boldsymbol{i}_j$ is sampled from an Active Part P with size βn,*

$$\mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})||\boldsymbol{i}_j \in P] \leq |A(M, S)| - \beta \cdot \varepsilon n^2/16.$$

Proof. [Proof of Lemma 5.11 assuming Lemma 5.12] The proof is exactly the same as the proof of Lemma 5.5. ∎

The chain of implications between lemmas and the proofs of the lemmas are all the same as in the ultrametric case, with only the proof of the last Lemma 5.12 being different. We now present this proof.

Proof. [Proof of Lemma 5.12]

Let B be the sub-block-matrix $M_{|P \times P}$ representing the part P. Suppose the $(j,k)^{th}$ entry of block B gives the distance $M(i_j, i_k)$ between indices i_j and i_k. We fix an arbitrary sampled point $u \in S$ and define a $\beta n \times \beta n$ masking matrix MASK where the $(j,k)^{th}$ entry of MASK equals $M(u, i_j) + M(u, i_k)$. That is,

$$B = \begin{bmatrix} M(i_1,i_1) & M(i_1,i_2) & \cdots & M(i_1,i_{\beta n}) \\ M(i_2,i_1) & M(i_2,i_2) & \cdots & M(i_2,i_{\beta n}) \\ M(i_3,i_1) & M(i_3,i_2) & \cdots & M(i_3,i_{\beta n}) \\ \vdots & \vdots & \ddots & \vdots \\ M(i_{\beta n},i_1) & M(i_{\beta n},i_2) & \cdots & M(i_{\beta n},i_{\beta n}) \end{bmatrix},$$

$$\text{MASK} = \begin{bmatrix} M(u,i_1)+M(u,i_1) & M(u,i_2)+M(u,i_1) & \cdots & M(u,i_{\beta n})+M(u,i_1) \\ M(u,i_1)+M(u,i_2) & M(u,i_2)+M(u,i_2) & \cdots & M(u,i_{\beta n})+M(u,i_2) \\ M(u,i_1)+M(u,i_3) & M(u,i_2)+M(u,i_3) & \cdots & M(u,i_{\beta n})+M(u,i_3) \\ \vdots & \vdots & \ddots & \vdots \\ M(u,i_1)+M(u,i_{\beta n}) & M(u,i_2)+M(u,i_{\beta n}) & \cdots & M(u,i_{\beta n})+M(u,i_{\beta n}) \end{bmatrix}.$$

Lemma 5.6 in the ultrametric testing section considers the values in the block matrix B. Here, instead, we consider the values in the square matrix $(B - \text{MASK})$. The motivation is that, when we sample some index $\boldsymbol{i}_j$ from the part P, other indices in P are partitioned into equivalence classes according to the values in the

j^{th} row/column of $(B - \text{MASK})$. In particular, for indices i_a, i_b,

$$
\begin{aligned}
& i_a, i_b \text{ are separated into two equivalence classes according to } S \cup \{\boldsymbol{i}_j\} \\
\Leftrightarrow & \exists\, w \in S, M(\boldsymbol{i}_j, i_a) + M(w, i_b) \neq M(w, i_a) + M(\boldsymbol{i}_j, i_b) \\
\Leftrightarrow & M(\boldsymbol{i}_j, i_a) + M(w, i_b) + M(u, i_b) \neq M(w, i_a) + M(\boldsymbol{i}_j, i_b) + M(u, i_b) = M(w, i_b) + M(\boldsymbol{i}_j, i_b) + M(u, i_a) \\
& \text{(equality follows from the fact that } i_a, i_b \text{ are in the same equivalence class under } S) \\
\Leftrightarrow & M(\boldsymbol{i}_j, i_a) + M(u, i_b) \neq M(\boldsymbol{i}_j, i_b) + M(u, i_a) \\
\Leftrightarrow & M(\boldsymbol{i}_j, i_a) - M(u, \boldsymbol{i}_j) - M(u, i_a) \neq M(\boldsymbol{i}_j, i_b) - M(u, \boldsymbol{i}_j) - M(u, i_b) \\
\Leftrightarrow & \text{ the } (\boldsymbol{i}_j, i_a) \text{ and } (\boldsymbol{i}_j, i_b) \text{ entry of the matrix } B - \text{MASK are different.}
\end{aligned}
$$

Reusing the argument idea of Lemma 5.6, among the entries in $B \setminus \text{EC}(M, S)$, suppose the set of the values these entries take on in $(B - \text{MASK})$ is $V = \{v_1, v_2, ..., v_k\}$, and further suppose v_1 is the value that most entries take on. Suppose $V' = \{v_{k+1}, ..., v_l\}$ is the set of values the easy-to-detect corrupted entries in B take on in the matrix $(B - \text{MASK})$. We first show that any value in V' is larger than any value in V, so as a result $\{v_1, ..., v_k, v_{k+1}, ..., v_l\}$ is indeed a set of non-repetitive values.

CLAIM 5.2. *Suppose $v' \in V'$ and $v \in V$. Then $v' > v$.*

Proof. Suppose the entry (i_a, i_b) in $(B - \text{MASK})$ takes on the value v', which implies that $M(i_a, i_b) \in \text{EC}(M, S)$ and $v' = M(i_a, i_b) - M(u, i_a) - M(u, i_b)$. Suppose the entry (i_c, i_d) in $(B - \text{MASK})$ takes on the value v, which implies that $M(i_c, i_d)$ is not in $\text{EC}(M, S)$. Moreover, $v = M(i_c, i_d) - M(u, i_c) - M(u, i_d)$. Since $M(i_a, i_b)$ is an easy-to-detect corruption, there exist some $x, y \in S$ such that $M(i_a, i_b) + M(x, y) > M(i_a, x) + M(i_b, y)$. On the other hand, as $M(i_c, i_d)$ is not an easy-to-detect corruption, for the pair x, y, $M(i_c, i_d) + M(x, y) \leq M(i_c, x) + M(i_d, y)$.

Following the above facts, we have the inequalities

$$
\begin{aligned}
v' =& M(i_a, i_b) - M(u, i_a) - M(u, i_b) && (5.2)\\
>& M(i_a, x) + M(i_b, y) - M(x, y) - M(u, i_a) - M(u, i_b) && (5.3)\\
=& M(i_c, i_d) + M(i_a, x) + M(i_b, y) - (M(x, y) + M(i_c, i_d)) - M(u, i_a) - M(u, i_b) && (5.4)\\
\geq& M(i_c, i_d) + M(i_a, x) + M(i_b, y) - (M(i_c, x) + M(i_d, y)) - M(u, i_a) - M(u, i_b) && (5.5)\\
=& M(i_c, i_d) + M(i_a, x) + M(i_b, y) - (M(i_c, x) + M(i_a, u)) - (M(i_d, y) + M(i_b, u)) && (5.6)\\
=& M(i_c, i_d) + M(i_a, x) + M(i_b, y) - (M(i_c, u) + M(i_a, x)) - (M(i_d, u) + M(i_b, y)) && (5.7)\\
=& M(i_c, i_d) - M(i_c, u) - M(i_d, u) = v, && (5.8)
\end{aligned}
$$

where (2) and (4) follow from the above facts on easy-to-detect corruptions, (6) follows from the definition of i_c, i_d, i_a, i_b being in the same part, and the other equations are merely rearrangements of terms. ∎

By exactly the same argument as in Lemma 5.6, if we let r_1 to denote the number of entries taking on value v_1, then over the randomness of sampling an index $\boldsymbol{i}_j$ from P, the expected number of pairs of indices in P that are separated is at least $((\beta n)^2 - r_1)/8$. Moreover, $|A(M, S)| - |A(M, S \cup \{\boldsymbol{i}_j\})|$ is at least the number of pairs being separated. What we left in this proof is to find a bound for r_1.

Recall that in Lemma 5.6, we argued that there is a block matrix encoding ultrametric with all entries taking on value v_1. The same argument applies for $(B - \text{MASK})$. If we set a $(\beta n) \times (\beta n)$ matrix to be an all-v_1 matrix, then after un-masking this matrix, we obtain a new block matrix B' which is a tree metric and consistent with S. Notice that this matrix only differs with B on $(\beta n)^2 - r_1$ entries.

CLAIM 5.3. *Since P is an Active Part, $\left(\sum_{\tau=2}^{l} r_\tau\right) = (\beta n)^2 - r_1 \geq \beta \cdot \varepsilon n^2/2$.*

Proof. We prove this by contradiction. Suppose $\left(\sum_{\tau=2}^{l} r_\tau\right) = (\beta n)^2 - r_1 < \beta \cdot \varepsilon n^2/2$. Create a $\beta n \times \beta n$ square matrix M_P^*, which is the sum of MASK and an all-v_1 matrix of dimension $\beta n \times \beta n$. Then the metric on subset $S \cup P$ defined by

$$D(i,j) = \begin{cases} M(i,j) & \text{if } i \text{ or } j \text{ or both in } S \\ M_P^*(i,j) & \text{if } i,j \in P \end{cases}$$

is a tree metric. Moreover, $\|M_P^* - B\| \leq \sum_{\tau=2}^{l} r_\tau < \beta \cdot \varepsilon n^2/2$. Therefore, M_P^* satisfies all conditions that make P a Versatile Part. However, by assumption, P is Active, which leads to a contradiction. ∎

As a summary, we have the sequence of inequalities

$$8 \cdot \left(|A(M,S)| - \mathop{\mathbf{E}}_{\boldsymbol{i}_j}[|A(M, S \cup \{\boldsymbol{i}_j\})| | \boldsymbol{i}_j \in P] \right) \geq (\beta n)^2 - r_1 \geq \beta \cdot \varepsilon n^2/2.$$

∎

6 Ultrametric and Tree Metric Testing Lower Bound—Theorem 1.4

In this section, we present two distributions $\mathcal{D}_s$ and $\mathcal{D}_q$, both supported on $n \times n$ square matrices and ε-far from both ultrametrics and tree metrics. Any testing algorithm that samples $o(1/\varepsilon)$ indices will not find any violation to either ultrametric or tree metric property in $\mathbf{M} \sim \mathcal{D}_s$ with probability at least 2/3; any testing algorithm that makes $o(1/\varepsilon^{4/3})$ queries will not find any violation to either ultrametric or tree metric property in a draw $\mathbf{M} \sim \mathcal{D}_q$ with probability at least 2/3.

6.1 Sample Complexity Lower Bound Distribution $\mathcal{D}_s$

Description of $\mathcal{D}_s$. We describe a single $n \times n$ matrix M, which we will then randomize by re-shuffling indices in n. Divide the n indices into two groups, $G = \{x_1, x_2, ..., x_r\}$, $B = \{y_1, ..., y_s\}$, where $|G| = r = n - \lfloor \varepsilon n \rfloor$ and $|B| = s = \lfloor \varepsilon n \rfloor$. Set the diagonal entries of M to 0 and the remaining entries as follows,

$$M(x_i, x_j) = 2n, M(x_i, y_j) = M(y_j, x_i) = 2n + i, M(y_i, y_j) = 2n, \ \forall\, x_i, x_j \in G, y_i, y_j \in B.$$

Claim 6.1. *Matrix M constructed as above satisfies the below properties.*

- *Suppose the triple $\{a, b, c\}$ is a violation of ultrametric property in M. Then after re-naming a, b, and c, $a, b \in G, c \in B$.*
- *Suppose the quadruple $\{a, b, c, d\}$ is a violation of the tree property in M. Then after re-naming a, b, c, and d, $a, b, c \in G, d \in B$.*

Proof. For the first item, suppose $a, b, c \in G$, then all three pairwise distances equal $2n$. If one index is in G and another two are in B, suppose the index in G is x_i for some $i \in [r]$. Then the three pairwise distances equal $\{2n, 2n + i, 2n + i\}$. Lastly, if all three indices are in B, then all three pairwise distances equal $2n$. None of the above triples are a violation of the ultrametric property.

Similarly, for the second item, if all four indices are in G, then all three induced matchings equal $2n + 2n$. If two indices are in G, which we denote as x_i, x_j for some $i, j \in [r]$, and two other indices in B, then the three induced matchings equal $2n + 2n, (2n + i) + (2n + j), (2n + j) + (2n + i)$. If only one index is in G, which we denote as x_i for some $i \in [r]$, and three other indices are in B, all three induced matchings equal $2n + (2n + i)$. Lastly, if all four indices are in B, all three induced matchings equal $2n + 2n$. None of the above quadruples are violations of the tree metric four-point-condition. ∎

CLAIM 6.2. *The matrix M is $\Omega(\varepsilon)$-far from an ultrametric and $\Omega(\varepsilon)$-far from a tree metric.*

Proof. First, the $n \times n$ matrix $\tilde{M}$ with every entry equal to $2n$ encodes an ultrametric and a tree metric. Moreover, $\|M - \tilde{M}\|_0 = 2\lfloor \varepsilon n \rfloor (n - \lfloor \varepsilon n \rfloor) = \Omega(\varepsilon n^2)$.

On the other hand, for $i < j \in [r], l \in [s]$, all triples of the form $\{x_i, x_j, y_l\}$ violate the ultrametric property, since $M(x_i, x_j) = 2n < M(x_i, y_l) = 2n + i < M(x_j, y_l) = 2n + j$. There are $\binom{n-\lfloor \varepsilon n \rfloor}{2} \lfloor \varepsilon n \rfloor = \Omega(\varepsilon n^3)$ such violating triples. Each pair (x_i, y_l) or (y_l, x_i) participates in $n - \lfloor \varepsilon n \rfloor$ many triples; each pair (x_i, x_j) participates in $\lfloor \varepsilon n \rfloor$ such triples. Suppose matrix M' encodes an ultrametric and S' is the set of entries M and M' differ upon. Then, all $\Omega(\varepsilon n^3)$ violating triples need to be covered by some pairs of indices in S', which indicates $\|M' - M\|_0 = |S'| = \Omega(\varepsilon n^2)$.

Similarly, for $i, j, k \in [r], l \in [s]$, any quadruple of the form (x_i, x_j, x_k, y_l) violates the four-point condition for tree metric property. There are $\Omega(\varepsilon n^4)$ such quadruples. Each pair (x_i, y_l) participates in at most $\Omega(n^2)$ such quadruples, and each pair (x_i, x_j) participates in at most $\Omega(\varepsilon n^2)$ such quadruples. Suppose matrix M'' encodes a tree metric and S'' is the set of entries M and M'' differ upon. Then, all $\Omega(\varepsilon n^4)$ violating quadruples need to be covered by some pairs of indices in S'', which indicates $\|M'' - M\|_0 = |S''| = \Omega(\varepsilon n^2)$. ∎

We let $\mathbf{M} \sim \mathcal{D}_s$ be obtained from M by re-ordering rows and columns according to a uniformly random permutation $\boldsymbol{\pi}$. By Yao's minimax principle, it suffices to rule out any deterministic ultrametric testing algorithm for finding violating triples and any deterministic tree metric testing algorithm for finding violating quadruples in a draw $\mathbf{M} \sim \mathcal{D}_s$.

LEMMA 6.1. *Consider any deterministic non-adaptive testing algorithm which samples $o(1/\varepsilon)$ indices. With probability at least $2/3$ over the draw of $\mathbf{M} \sim \mathcal{D}_s$, there are no indices i, j, k with*

$$\mathbf{M}(i, j) > \max\{\mathbf{M}(i, k), \mathbf{M}(k, j)\};$$

Moreover, with probability at least $2/3$ over the draw of $\mathbf{M} \sim \mathcal{D}_s$, there are no indices i, j, k, l with

$$\mathbf{M}(i, j) + \mathbf{M}(k, l) > \max\{\mathbf{M}(i, k) + \mathbf{M}(j, l), \mathbf{M}(i, l) + \mathbf{M}(j, k)\}.$$

For partial matrix M' that contains the pairwise distance between each pair of indices in a subset $S \subset [n]$ and does not violate the ultrametric three-point-condition (or tree metric four-point-condition), M' may be completed to one which is an ultrametric (or tree metric). Thus, any non-adaptive testing algorithm must sample $\Omega(1/\varepsilon)$ indices.

Proof. For the first claim, it suffices to show that for any testing algorithm which samples $o(1/\varepsilon)$ indices S, with probability at least $2/3$ over the draw of $\mathbf{M} \sim \mathcal{D}_s$, $S \cap B = \emptyset$. Then, by Claim 6.1, all triples in S are not violations of the ultrametric property. Thus, the probability that a violation of ultrametric property is found among the indices S, by a union bound, is at most

$$\sum_{b \in B} \mathbf{Pr}_{\boldsymbol{\pi}} (\boldsymbol{\pi}(b) \in S)) = |B| \cdot \frac{|S|}{n} = o(1)$$

when $|S| = o(1/\varepsilon)$.

For the second claim, the same argument as above holds. That is, by Claim 6.1, the probability that a violation of tree metric property is found among the indices S is at most the probability that some index in B fall into S. By the same union bound, this probability is at most $o(1)$ when $|S| = o(1/\varepsilon)$. ∎

6.2 Query Complexity Lower Bound Distribution $\mathcal{D}_q$ Notice that any ultrametric is also a tree metric. We show below a distribution $\mathcal{D}_q$ such that any testing algorithm that makes $o(1/\varepsilon^{4/3})$ queries will not find any violation of the ultrametric property in a draw $\mathbf{M} \sim \mathcal{D}_q$ with probability at least 2/3. Thus, any testing algorithm that makes $o(1/\varepsilon^{4/3})$ queries will not find any violation of the tree metric property in a draw $\mathbf{M} \sim \mathcal{D}_q$ with probability at least 2/3.

Description of $\mathcal{D}_q$. As before, we describe a single $n \times n$ matrix M, which we will then randomize by re-shuffling indices in n. First, let $r = \lfloor \varepsilon n \rfloor$. We define a $r \times r$ block matrix D and then use multiple block diagonal matrices D to construct M. First, D is a randomized $r \times r$ matrix with 0 on the diagonal; for $i < j \in [r]$, set entry $D(i,j) = D(j,i)$ to be 1 with probability 1/2 and 2 with probability 1/2. Partition $[n]$ into l groups of indices $[n] = (\cup_{j=1}^{l} I_j) \cup R$ for $l = \lfloor 1/\varepsilon \rfloor$ such that each group I_j has r indices, and possibly a ramainder group R with less than r indices. In particular, for $j \in [l]$, I_j contain all integers in $((j-1)r, jr]$. For each $j \in [l]$, sample a random $r \times r$ matrix D_j and set $M_{|I_j \times I_j} = D_j$. This is equivalent to letting l random block matrices D_j to be the block diagonal matrices of M. Set all other entries, i.e. the entries that are not in the block diagonal matrices D_j and the entries in the remainder group $M_{|R \times R}$, to be 10.

Below is a demonstration of the matrix M, which has l random block diagonal matrices of dimension $r \times r$; the entries that are not in the block diagonal matrices are 10.

$$M = \left[\begin{array}{c|c|c|c|c|c} D_1 & * & * & \cdots & * & * \\ \hline * & D_2 & * & \cdots & * & * \\ \hline * & * & D_3 & \cdots & * & * \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline * & * & * & \cdots & D_l & * \\ \hline * & * & * & \cdots & * & R \end{array}\right].$$

Claim 6.3. *The matrix D is $\Omega(1)$-far from an ultrametric. In consequence, the matrix M is $\Omega(\varepsilon)$-far from an ultrametric.*

Proof. For indices $i < j < k \in [r]$, each entry $D(i,j), D(i,k), D(j,k)$ takes on value 1 with probability 1/2 and 2 with probability 1/2. Out of 8 arrangements of values that happen with equal probability, 3 of them incur a violation of the ultrametric property. Thus, with probability 3/8 the triple violates the ultrametric property. In the matrix D, there are $\Omega(r^3)$ such violating triples. Each pair of indices participates in $\Omega(r)$ such violating triples. Suppose the $r \times r$ matrix $\tilde{D}$ encodes an ultrametric; further suppose E is the set of entries $\tilde{D}$ and D differ on. Then, each violating triple in D needs to be covered by some entry in E. This suggests $|E| = \Omega(r^2)$, so D is $\Omega(1)$-far from ultrametric.

Suppose $n \times n$ matrix $\tilde{M}$ encodes an ultrametric, then $\tilde{M}$ and M differ on $\Omega(r^2)$ entries on each block diagonal matrix, and there are $\Omega(1/\varepsilon)$ block diagonal matrices. This shows that $\|M - \tilde{M}\|_0 \geq \Omega(\varepsilon n^2)$ for every $\tilde{M} \in \mathcal{P}^U$. In fact, no more than εn^2 entries need to be modified for M to be an ultrametric, as $\|M - M'\|_0 \leq \varepsilon n^2$ for M' being a $n \times n$ matrix with each entry set to 10. ∎

Claim 6.4. *Suppose the triple $\{a,b,c\}$ is a violation of the ultrametric property in M. Then three indices are in the same index group I_j for some $j \in [l]$.*

Proof. We prove this by showing the contra-positive is true. Suppose less than 3 indices are in the same index group I_j for any $j \in [l]$. Then out of the three distances $M(a,b), M(a,c), M(b,c)$, at least two of them equal 10. The remaining distance is no larger than 10. Thus, the ultrametric is not violated for $\{a,b,c\}$. ∎

We let $\mathbf{M} \sim \mathcal{D}_q$ be obtained from M by re-ordering rows and columns according to a uniformly random

permutation $\boldsymbol{\pi}$. Again, it suffices to rule out any deterministic ultrametric testing algorithm for finding violating triples in a draw $\mathbf{M} \sim \mathcal{D}_s$.

LEMMA 6.2. *Consider any deterministic non-adaptive algorithm that makes $o(1/\varepsilon^{4/3})$ queries. With probability at least $2/3$ over the draw of $\mathbf{M} \sim \mathcal{D}_q$, no violation to the ultrametric can be found.*

Since any partial matrix that does not violate the ultrametric property may be completed to one that is ultrametric, any non-adaptive testing algorithm must make $\Omega(1/\varepsilon^{4/3})$ queries.

Proof. Suppose a testing algorithm samples a set of indices S of size s. The algorithm only recognizes a violation if at least three indices in S fall into the same index group I_j for some $j \in [l]$. Moreover, if a subset $U \subset S$ with at least three indices fall into the same index group I_j, the algorithm needs to make queries that form a cycle in U to recognize a violation. Formally we have the below claim.

CLAIM 6.5. *Suppose for a subset of indices $U \subset [n]$, E' is the set of pairs whose distances are known and E' is cycle-free. That is, suppose a graph G' is on vertices U and $(a,b) \in U^2$ is an edge if and only if $(a,b) \in E'$, and suppose G' is cycle-free. Then the known distances have no violation to ultrametric.*

Proof. Define G' as proposed in the claim and for each edge (a,b) in G', set its weight to be the known distance between a, b. Consider the partial matrix $M'_{|U\times U}$ whose entries are the known distances in E'. There exists a way to complete $M'_{|U\times U}$ to another matrix $M_{|U\times U}$ which encodes an ultrametric. Namely, first, we identify all disconnected components in G', and between each pair of components, find a vertex in each component and set their edge weight to be an arbitrary positive value. Now G' is a tree. Then, for every non-edge $(a,b) \notin E'$, find the unique simple path from a to b in the graph G' and set $M_{|U\times U}(a,b)$ to be the weight of the heaviest edge on the path. ∎

Upon sampling the indices S, the testing algorithm makes queries among a subset of pairs $E \subset S^2$. Create a graph G whose vertex set is S and edge set E. Let $\mathcal{C}(G)$ be the set of cycles in G. For a cycle $C \subset [n]$, let $\mathbb{I}(C)$ denote the indicator variable in $\{0,1\}$ with

$$\mathbb{I}(C) \qquad \Longleftrightarrow \qquad \left\{ \begin{array}{c} C \in \mathcal{C}(G), \\ \exists\, j \in [l],\, \forall\, v \in C,\, v \in \boldsymbol{\pi}(I_j) \end{array} \right\}$$

Then, the probability that a violation of ultrametric is detect by the algorithm, by a union bound, is at most

$$\sum_{C \in \mathcal{C}(G)} \mathbb{I}(C) \le \sum_{k=3}^{\infty} \sum_{\substack{C \in \mathcal{C}(G), \\ |C|=k}} 2\varepsilon^{k-1} \tag{6.9}$$

where the inequality follows from when cycle C has k indices, all k indices fall into the same index group with probability $\lfloor 1/\varepsilon \rfloor \cdot \binom{\lfloor \varepsilon n \rfloor}{k} / \binom{n}{k} \le 2\varepsilon^{k-1}$ as there are $\Omega(1/\varepsilon)$ index groups of $\lfloor \varepsilon n \rfloor$ indices. An algorithm aims to pick a query graph structure to maximize this quantity. The following claim shows that among the sampled indices S, making a "clique-like" query graph structure is optimal for a testing algorithm.

LEMMA 6.3. *Suppose a testing algorithm will make m queries. To maximize the quantity 6.9, the algorithm should select s indices such that $\binom{s}{2} \ge m > \binom{s-1}{2}$ and make m queries on the graph with these s indices. That is, a graph with m edges that is closest to a clique maximizes the quantity.*

Assuming Lemma 6.3, with m queries, the probability that a violation to ultrametric is detected is maximized when the algorithm samples s indices such that $\binom{s}{2} = m$ and makes m queries on the pairwise distances. If an

algorithm samples a set S of $o(1/\varepsilon^{2/3})$ indices and query pairwise distances, then a violation is detected when three indices in S fall into the same index group. For each triple, all three indices fall into the same group with probability at most $2\varepsilon^2$. Thus, by a union bound, the probability that the algorithm detects a violation is at most

$$\sum_{i,j,k\in S} \mathbf{Pr}\left(\exists\, j \in [l],\, i,j,k \in \boldsymbol{\pi}(I_j)\right) \leq o(1/\varepsilon^2)\cdot 2\varepsilon^2 = o(1)$$

when $|S| = o(1/\varepsilon^{2/3})$. Such clique-like query structure makes $o(1/\varepsilon^{4/3})$ queries, and for any other query structure, by Lemma 6.3, the probability of detecting a violation is no larger than $o(1)$. This shows that any non-adaptive testing algorithm must make $\Omega(1/\varepsilon^{4/3})$ queries to detect a violation of ultrametric property with high probability.

Proof. [Proof of Lemma 6.3] Given the choice to make m queries, an algorithm wants to maximize the quantity $\sum_{k=3}^{\infty} \varepsilon^{k-1}\cdot|C_k(G)|$ where $C_k(G)$ denotes the set of cycles of length k in the query graph G. Let s be an integer defined as in the statement. We start with an arbitrary graph G with $s' > s$ vertices and m edges, and show that G can be modified to G' with strictly less vertices such that $\sum_{k=3}^{\infty} \varepsilon^{k-1}\cdot|C_k(G')| > \sum_{k=3}^{\infty} \varepsilon^{k-1}\cdot|C_k(G)|$. This graph modification is done by merging two vertices and adding edges between non-adjacent vertices.

First, we modify $G = (V, E)$ such that its diameter is at most 2. Suppose initially the diameter of G is larger than 2. Then, there exists two non-adjacent vertices x, y such that the neighbor of x, denoted by $N_x(G)$, is disjoint from that of y, $N_y(G)$. Create a new graph G' on the vertex set $(V \cup \{z\}) \setminus \{x, y\}$ and an edge set E' :

$$E' = \{(u,v) : (u,v) \in E, u, v \notin \{x,y\}\} \cup \{(v,z) : (v,x) \in E\} \cup \{(v,z) : (v,y) \in E\}.$$

The following statements hold:

- For $C \in \mathcal{C}(G)$, if $x, y \notin C$, $C \in \mathcal{C}(G')$.
- For $C \in \mathcal{C}(G)$, if $|C \cap \{x, y\}| = 1$, $C \in \mathcal{C}(G')$. That is, any cycle in G that only contains one of x, y still remains in G'.
- For $C \in \mathcal{C}(G)$,if $x, y \in C$, then C must be of the form $\{x \to v_1 \to \cdots \to v_i \to y \to u_1 \to \cdots \to u_j \to x\}$ where $v_1 \neq v_i$, $u_1 \neq u_j$. Thus, the cycles $C_1 = \{z \to v_1 \to \cdots \to v_i \to z\}$ and $C_2 = \{z \to u_1 \to \cdots \to u_i \to z\}$ are in $\mathcal{C}(G')$ and $|C_1| < |C|$ and $|C_2| < |C|$.

Therefore, $\sum_{k=3}^{\infty} \varepsilon^{k-1}\cdot|C_k(G')| > \sum_{k=3}^{\infty} \varepsilon^{k-1}\cdot|C_k(G)|$.

Hence, suppose that $G = (V, E)$ is a graph with s' vertices with diameter at most 2. Since $s' > s$ where integer s satisfies $\binom{s}{2} \geq m$, we must have that at least $\binom{s'}{2} - m \geq \binom{s'}{2} - \binom{s'-1}{2} = s' - 1$ pairs of vertices are non-adjacent in G. That is, G has at least $s' - 1$ non-edges. Suppose $(x, y) \notin E$, and suppose $|N_x(G) \cap N_y(G)| = \nu$. Notice that $\nu > 0$ as $diam(G) \leq 2$. Consider the effect of adding an edge (x, y) into the graph as well as the effect of merging the pair of vertices (x, y) into a single vertex z.

Create edge (x, y)**:** Define G'' to be a graph on vertex set V and edge set $E \cup \{(x, y)\}$. Then $|\mathcal{C}(G'')| > |\mathcal{C}(G)|$. In particular, we gain ν new triangles from the ν common neighbors of $x, y \in G$. That is, $|C_3(G'')| = |C_3(G)| + \nu$ so $|\mathcal{C}(G'')| \geq |\mathcal{C}(G)| + \nu$.

Merge vertices (x, y)**:** Define G'' to be a graph on vertex set $(V \cup \{z\}) \setminus \{x, y\}$ and edge set E'', where

$$E'' = \{(u,v) : (u,v) \in E, u, v \notin \{x,y\}\} \cup \{(v,z) : (v,x) \in E\} \cup \{(v,z) : (v,y) \in E\}.$$

For $C \in \mathcal{C}(G)$ and $|C \cap \{x, y\}| \leq 1$, $C \in \mathcal{C}(G'')$; for $C \in \mathcal{C}(G)$, $x, y \in C$, and $|C| \geq 5$, C corresponds to at least one cycle $\tilde{C} \in \mathcal{C}(G'')$ where $|\tilde{C}| < 5$. For $C \in \mathcal{C}(G)$, $x, y \in C$, and $|C| = 4$, C is of the form $\{x \to v \to y \to u \to x\}$. Such cycles are vanishing due to the merge of (x, y). Moreover, for triangles of

the form $\{x,u,v\},\{y,u,v\} \in C_3(G)$ for some $u,v \in V$, $\{z,u,v\} \in C_3(G'')$: after the merge, two triangles become one. These types of triangles are the only affected ones. Moreover, since $|N_x(G) \cap N_y(G)| = \nu$, $|E''| = |E| - \nu$. In summary, among the ν common neighbors in $N_x(G) \cap N_y(G)$, if p pairs of them are adjacent, then merging (x,y) causes p triangles vanishing, $\binom{\nu}{2}$ cycles of length 4 vanishing, but the number of edges decrease by ν.

Now we will show how to create G' base on G to increase the objective function. Suppose the set of non-edges in G is

$$(x_1,y_1),\ldots,(x_\alpha,y_\alpha),\ \forall\, i \in [\alpha],\, x_i,y_i \in V,\, (x_i,y_i) \notin E.$$

Recall that $\alpha \geq s' - 1$. Each pair $(x_i,y_i) \notin E$ is associated with two quantities

$$\begin{aligned}\nu_i &:= |N_{x_i}(G) \cap N_{y_i}(G)|,\\ p_i &:= |\{(u,v) \in E : u,v \in (N_{x_i}(G) \cap N_{y_i}(G))\}|.\end{aligned}$$

That is, ν_i is the number of common neighbors of x_i, y_i, and p_i is the number of pairs of common neighbors that are adjacent. Notice that $p_\alpha \leq \binom{\nu_\alpha}{2}$. We assume the pairs of non-edges are sorted in the way that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_\alpha$. The construction of G' is done by merging a single pair (x_α, y_α) and adding edges to some of the previous non-adjacent pairs.

By the previous case analysis, merging (x_α,y_α) into one single new vertex z causes p_α triangles and $\binom{\nu_\alpha}{2}$ cycles of length 4 vanishing; but we also spares ν_α edges that can be added to elsewhere in the graph. Let $G'' = (V'', E'')$ denote the new graph with x_α, y_α merged into a new vertex x. (G'' is defined formally in Claim 6.6 below.) Then $|C_3(G'')| \geq |C_3(G)| - p_\alpha \geq |C_3(G)| - \binom{\nu_\alpha}{2}$ and $|C_4(G'')| \geq |C_4(G)| - \binom{\nu_\alpha}{2}$. Moreover, for $j > 4$, $|C_j(G'')| \geq |C_j(G)|$. In terms of the number of edges, $|E''| = |E| - \nu_\alpha$.

A natural next step is to create ν_α edges between some of the non-edges $(x_1,y_1),\ldots,(x_{\alpha-1},y_{\alpha-1})$ and make some of them adjacent. However, note that a by-product of merging (x_α,y_α) into a new single vertex z in G'' is that, some of the non-edges $(x_1,y_1),\ldots,(x_{\alpha-1},y_{\alpha-1}) \in G$ become adjacent in G''. Consider the possible scenario where two non-edges (x_i,y_i) and (x_α,y_α) in G share one same endpoint $x_i = x_\alpha$ and the pair $(y_\alpha,y_i) \in E$ is adjacent in G. Thus, the non-edge $(x_i,y_i) = (x_\alpha,y_i) \in G$ corresponds to $(z,y_i) \in G''$ when merging x_α and y_α as a new vertex $z \in G''$. In this case, it makes no sense to use the ν_α-edge-budget to create an edge between (x_i,y_i) in G'', as the vertex $x_i = x_\alpha$ corresponds to z in G'' which is already adjacent to y_i. Below, we give an upper bound on the decrease in the non-edges after the merge.

CLAIM 6.6. *Define G'' to be a graph on vertex set $V'' = (V \cup \{z\}) \setminus \{x_\alpha, y_\alpha\}$ and edge set E'' where*

$$E'' = \{(u,v) : (u,v) \in E, u,v \notin \{x_\alpha,y_\alpha\}\} \cup \{(v,z) : (v,x_\alpha) \in E\} \cup \{(v,z) : (v,y_\alpha) \in E\}.$$

Then there are at least ν_α non-edges in G''.

Proof. Suppose non-edge $(x_i,y_i) \notin E$ in G becomes adjacent in E''. Then we must be in one of the below two cases

$$\Big\{x_i = x_\alpha \text{ and } (y_\alpha,y_i) \in E\Big\} \qquad \text{or} \qquad \Big\{y_i = y_\alpha \text{ and } (x_\alpha,x_i) \in E\Big\}.$$

Consider the former case. Since $(y_\alpha,y_i) \in E$, y_i was not adjacent to $x_\alpha = x_i$ as $(x_i,y_i) \notin E$. That is, $y_i \in N_{y_\alpha}(G) \setminus N_{x_\alpha}(G)$. Thus, $|\{(y_\alpha,y_i) \in E : y_i \in V\}| \leq |N_{y_\alpha}(G) \setminus N_{x_\alpha}(G)|$. Similarly, $|\{(x_\alpha,x_i) \in E : x_i \in V\}| \leq |N_{x_\alpha}(G) \setminus N_{y_\alpha}(G)|$. In summary, $|\{(x_i,y_i) \notin E, (x_i,y_i) \in E''\}| \leq |N_{y_\alpha}(G) \setminus N_{x_\alpha}(G)| + |N_{x_\alpha}(G) \setminus N_{y_\alpha}(G)| \leq s' - 2 - \nu_\alpha$. There are $\alpha - 1$ non-edges $(i_1,j_1),\ldots,(i_{\alpha-1},j_{\alpha-1})$ in G. Thus, in G'', there are at least $\alpha - 1 - (s' - 2 - \nu_\alpha) = \alpha - s' + 1 + \nu_\alpha \geq \nu_\alpha$ non-adjacent pairs. ∎

The above argument shows that there are still at least ν_α pairs of vertices in G' that remain to be non-adjacent, and the ν_α spare edges can be added to ν_α of the non-adjacent pairs. After renaming, let these ν_α non-adjacent pairs be

$$E^{(NA)} = \{(x_i,y_i) \notin E'' : i \in [\nu_\alpha], x_i,y_i \in V''\}.$$

Define graph G' on the vertex set $V' = V'' = (V \cup \{z\}) \setminus \{x_\alpha, y_\alpha\}$ and edge set E' :

$$\begin{aligned} E' =& \{(u,v) : (u,v) \in E, u,v \notin \{x_\alpha, y_\alpha\}\} \cup \{(v,z) : (v,x_\alpha) \in E\} \\ & \cup \{(v,z) : (v,y_\alpha) \in E\} \cup \{(x,y) : (x,y) \in E^{(NA)}\}. \end{aligned}$$

Notice that, for each new edge $(x_i, y_i) \in E^{(NA)}$, $N_{x_i}(G') \cap N_{y_i}(G') = (((N_{x_i}(G) \cap N_{y_i}(G))) \setminus \{x,y\}) \cup \{z\}$ if either x or $y \in (N_{x_i}(G) \cap N_{y_i}(G))$; if $x, y \notin (N_{x_i}(G) \cap N_{y_i}(G))$, then $N_{x_i}(G') \cap N_{y_i}(G') = N_{x_i}(G) \cap N_{y_i}(G)$. In both cases, $|N_{x_i}(G') \cap N_{y_i}(G')| \geq \nu_\alpha - 1$. As a consequence, for each such pair $(x_i, y_i) \in E^{(NA)}$, adding this edge in G' creates at least $\nu_\alpha - 1$ new triangles in G'. In total, adding ν_α spare edges creates at least $\nu_\alpha \cdot (\nu_\alpha - 1)$ new triangles. Thus, $|C_3(G')| \geq |C_3(G'')| + \nu_\alpha \cdot (\nu_\alpha - 1)$ and the number of cycles with larger length can only increase.

In summary, from graph G to G'', we have

$$\begin{aligned} |C_3(G'')| &\geq |C_3(G)| - \nu_\alpha(\nu_\alpha - 1)2, \\ |C_4(G'')| &\geq |C_4(G)| - \nu_\alpha(\nu_\alpha - 1)2, \\ |C_j(G'')| &\geq |C_j(G)|, \forall j > 4, \\ |E''| &= |E| - \nu_\alpha. \end{aligned}$$

From graph G'' to G', we have

$$\begin{aligned} |C_3(G')| &\geq |C_3(G'')| + \nu_\alpha \cdot (\nu_\alpha - 1), \\ |C_j(G')| &\geq |C_j(G'')|, \forall j \geq 4, \\ |E'| &= |E''| + \nu_\alpha. \end{aligned}$$

Thus, $|C_3(G')| \geq |C_3(G)| + \binom{\nu_\alpha}{2}$, $|C_4(G')| \geq |C_4(G)| - \binom{\nu_\alpha}{2}$, $|C_j(G')| \geq |C_j(G)|$ for all $j > 4$ and the number of larger cycles in G' is at least that of G. Recall that $\nu_\alpha > 0$. This shows that, with the same number of edges m, $\sum_{k=3}^{\infty} \varepsilon^{k-1} \cdot |C_k(G')| > \sum_{k=3}^{\infty} \varepsilon^{k-1} \cdot |C_k(G)|$. □

□

Acknowledgement

Erik Waingarten would like to thank the support from the National Science Foundation (NSF) under Grant No. CCF-2337993.

References

[1] Noga Alon, Tali Kaufman, Michael Krivelevich, and Dana Ron, *Testing triangle-freeness in general graphs*, SIAM Journal on Discrete Mathematics, 22(2) (2008).

[2] Noga Alon, *Testing subgraphs in large graphs*, Random Structures & Algorithms, 21(3–4) (2002).

[3] Mihai Bădoiu, Artur Czumaj, Piotr Indyk, and Christian Sohler, *Facility location in sublinear time*, in Proceedings of the 32nd International Colloquium on Automata, Languages, and Programming (ICALP '2005), 2005.

[4] Ainesh Bakshi, Nadiia Chepurko, and Rajesh Jayaram, *Testing positive semi-definiteness via random submatrices*, in Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS '2020), 2020.

[5] Rajarshi Bhattacharjee, Gregory Dexter, Petros Drineas, Cameron Musco, and Archan Ray, *Sublinear time eigenvalue approximation via random sampling*, Algorithmica, 86:1764–1829 (2024).

[6] Jon Bentley, *Multidimensional binary search trees used for associative searching*, Communications of the ACM, 18(9):509–517 (1975).

[7] Manuel Blum, Michael Luby, and Ronitt Rubinfeld, *Self-testing/correcting with applications to numerical problems*, Journal of Computer and System Sciences, 47(3):549–595 (1993).
[8] Maria-Florina Balcan, Yi Li, David P. Woodruff, and Hongyang Zhang, *Testing matrix rank, optimally*, in Proceedings of the 30th ACM-SIAM Symposium on Discrete Algorithms (SODA '2019), 2019.
[9] Nader H. Bshouty, *On property testing of the binary rank*, in Proceedings of the 48th International Symposium on Mathematical Foundations of Computer Science (MFCS '2023), 2023.
[10] Arnab Bhattacharyya and Yuichi Yoshida, *Property Testing: Problems and Techniques*, Springer Singapore, 2022.
[11] Vincent Cohen-Addad, Chenglin Fan, Euiwoong Lee, and Arnaud de Mesmay, *Fitting metric and ultrametrics with minimum disagreements*, in Proceedings of the 63rd Annual IEEE Symposium on Foundations of Computer Science (FOCS '2022), 2022.
[12] Moses Charikar and Ruiquan Gao, *Improved approximations for ultrametric violation distance*, in Proceedings of the 34th ACM-SIAM Symposium on Discrete Algorithms (SODA '2023), 2023.
[13] Bernard Chazelle, Ding Liu, and Avner Magen, *Sublinear geometric algorithms*, in Proceedings of the 35th ACM Symposium on the Theory of Computing (STOC '2003), 2003.
[14] Artur Czumaj and Christian Sohler, *Property testing with geometric queries*, in Proceedings of the 9th European Symposium on Algorithms (ESA '2001), 2001.
[15] Artur Czumaj and Christian Sohler, *Estimating the weight of metric minimum spanning trees in sublinear time*, SIAM Journal on Computing, 39(3):904–922 (2009).
[16] Artur Czumaj and Christian Sohler, *Sublinear time approximation of the cost of a metric k-nearest neighbor graph*, in Proceedings of the 31st ACM-SIAM Symposium on Discrete Algorithms (SODA '2020), 2020.
[17] Artur Czumaj, Christian Sohler, and Martin Ziegler, *Property testing in computational geometry*, in Proceedings of the 8th European Symposium on Algorithms (ESA '2000), 2000.
[18] Dominique de Caen, *An upper bound on the sum of squares of degrees in a graph*, Discrete Mathematics, 185(1):245–248 (1998).
[19] Eldar Fischer and Ilan Newman, *Testing of matrix properties*, in Proceedings of the 33rd ACM Symposium on the Theory of Computing (STOC '2001), 2001.
[20] Chenglin Fan, Benjamin Raichel, and Gregory van Buskirk, *Metric violation distance: hardness and approximation*, in Proceedings of the 29th ACM-SIAM Symposium on Discrete Algorithms (SODA '2018), 2018.
[21] Piotr Indyk, Jiří Matoušek, and Anastasios Sidiropoulos, *Low-distortion embeddings of finite metric spaces*, in Handbook of Discrete and Computational Geometry, Chapman and Hall/CRC, 2017.
[22] Evangelos Kipouridis, *Fitting tree metrics with minimum disagreements*, in Proceedings of the 31st European Symposium on Algorithms (ESA '2023), 2023.
[23] Robert Krauthgamer and Ori Sasson, *Property testing of data dimensionality*, in Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms (SODA '2003), 2003.
[24] Nathan Linial, *Finite metric spaces: combinatorics, geometry, algorithms*, in Proceedings of the 18th ACM Symposium on Computational Geometry (SoCG '2002), 2002.
[25] Jiří Matoušek, *Lectures on Discrete Geometry*, vol. 212 of Graduate Texts in Mathematics, Springer, 2002.
[26] Krzysztof Onak, *Testing properties of sets of points in metric spaces*, in Proceedings of the 35th International Colloquium on Automata, Languages, and Programming (ICALP '2008), 2008.
[27] Michal Parnas and Dana Ron, *Testing metric properties*, Information and Computation, 187(2):155–195 (2003).
[28] Ronitt Rubinfeld and Madhu Sudan, *Robust characterization of polynomials with applications to program testing*, SIAM Journal on Computing, 25(2):252–271 (1996).
[29] Yufei Zhao, *Graph Theory and Additive Combinatorics*, Cambridge University Press, 2023.

A Definitions of Metric Spaces

Definition A.1. (Metric Space) *A metric space defined over $[n]$ is specified by a distance function $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ which satisfies:*

- *$d(i,j) = 0$ if and only if $i = j$, and $d(i,j) = d(j,i)$ for all $i, j \in [n]$.*
- **Triangle Inequality***: for all $i, j, k \in [n]$, $d(i,j) \leq d(i,k) + d(j,k)$.*

Furthermore, if the only condition unsatisfied above is that $i \neq j$ satisfies $d(i,j) = 0$, then d defines a pseudometric *over $[n]$.*

DEFINITION A.2. (TREE METRIC) *A metric space $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ defines a tree metric over $[n]$ if the function d may be realized in the following way:*

- *There is a weighted tree $T = (V, E)$ with weights $w\colon E \to (0, \infty)$ and an injection $\phi\colon [n] \to V$.*
- *The distance $d(i, j)$ is given by the length of the path (i.e., sum of edge weights) between $\phi(i)$ and $\phi(j)$ in T.*

An alternative (and equivalent) definition, due to [6], is that d is a tree metric, in addition to being a metric, satisfies:

$$d(i,j) + d(k,\ell) \leq \max\{d(i,k) + d(j,\ell), d(i,\ell) + d(j,k)\} \qquad \text{for all } i,j,k,\ell \in [n].$$

DEFINITION A.3. (ULTRAMETRIC) *A metric space $d\colon [n] \times [n] \to \mathbb{R}_{\geq 0}$ defines an ultrametric over $[n]$ if the function d may be realized in the following way:*

- *There is a rooted tree $T = (V, E)$ with a root $r \in V$ and weights $w\colon E \to (0, \infty)$ such the sum of weights from r to each leaf is the same. Furthermore, there is an injection $\phi\colon [n] \to V$ which maps to the leaves of the tree.*
- *The distance $d(i, j)$ is given by the length of the path (i.e., sum of edge weights) between $\phi(i)$ and $\phi(j)$ in T.*

An alternative (and equivalent) definition, is that d defines an ultrametric if, in addition to being a metric, satisfies:

$$d(i,j) \leq \max\{d(i,k), d(j,k)\} \qquad \text{for all } i,j,k \in [n],$$

B Proof of Theorem 1.1

THEOREM 1.1. (TESTING METRICS—UPPER BOUND) *For any large enough $n \in \mathbb{N}$ and any $\varepsilon \in (0, 1)$, there exists a randomized algorithm which receives query access to an unknown matrix $M \in \mathbb{R}^{n \times n}$ and makes $O(n^{2/3}/\varepsilon^{4/3})$ queries with the following guarantee:*

- *If M defines a metric space on $[n]$, the algorithm outputs "accept" with probability 1.*
- *If M is ε-far from being being a metric, then the algorithm output "reject" with probability at least $2/3$.*

Furthermore, the algorithm is non-adaptive (i.e., queries made do not depend on answers to prior queries).

Proof. By Lemma 2.1, it suffices to show the theorem holds for $M \in \mathcal{C}$. Note that, if $\varepsilon < 1/n$, the claimed complexity $O(n^{2/3}/\varepsilon^{4/3})$ is $O(n^2)$, so that we read the entire matrix. If $1/\varepsilon > n$, run the Metric Testing Algorithm. For $M \in \mathcal{P}$, both sub-routines output "accept" with probability 1, so the algorithm outputs "accept" if M encodes a metric space. If M is ε-far from $\mathcal{P}$, let T be the set of all violating triangles in M and let $I = \{i \in [n] : d_T(i) \geq \varepsilon^{1/3} n^{4/3}/16\}$. If $|I| \geq \varepsilon n/4$, then sub-routine CHECKHIDEGREE(M, ε) outputs "reject" with probability at least $5/6$. On the other hand, if $|I| < \varepsilon n/4$, sub-routine CHECKVIOLATION(M, ε) outputs "reject" with probability at least $5/6$. By union bound, the algorithm outputs "reject" with probability at least $2/3$ when $M \in \mathcal{C}$ is ε-far from $\mathcal{P}$. ∎

C Proofs from Section 3 (Metric Testing Upper Bound)

LEMMA 3.1. *For any $\varepsilon \in (0,1)$, and any $M \in \mathcal{C}$ which is ε-far from $\mathcal{P}$, there are at least $\varepsilon n^2/6$ distinct violating triangles $\{i,j,k\}$ of M.*

Proof. Suppose M contains less than $\varepsilon n^2/6$ triangles which are violating for M. Then, we show how to modify less than εn^2 entries on matrix M to get a matrix M' which encodes a metric space over $[n]$. Let S be the subset of $[n] \times [n]$ such that for all $(i,j) \in S$, (i,j) does not participate in any violating triangle. Then, $|S| > n^2 - \varepsilon n^2$ since a triangle $\{i,j,k\}$ has 3 (unordered) pairs of indices and hence, 6 entries in M. Consider the weighted undirected graph G on $[n]$ which adds edges $(i,j) \in S$ with the weight $M(i,j)$. If G is disconnected, add edges of the maximum weight to connect it. We then consider the matrix M', where entry $M'(i,j)$ denotes the length of the shortest path between i and j along edges in G. The matrix M' encodes a metric and differs with M on less than εn^2 entries. ∎

LEMMA 3.2. (CHECKHIDEGREE LEMMA) *For $n \in \mathbb{N}$ and $\varepsilon \in (1/n, 1)$, there exists a randomized algorithm,* CHECKHIDEGREE*, which receives as input an $n \times n$ matrix $M \in \mathcal{C}$ and a parameter ε and has the following guarantees:*

- *If $M \in \mathcal{P}$,* CHECKHIDEGREE(M, ε) *always outputs "accept."*
- *Letting T be the set of violating triangles of M, if there are at least $\varepsilon n/4$ indices $i \in [n]$ such that $d_T(i) \geq \varepsilon^{1/3} n^{4/3}/16$,* CHECKHIDEGREE$(M, \varepsilon)$ *outputs "reject" with probability at least* $5/6$.

The algorithm is non-adaptive, taking $O(1/\varepsilon + n^{2/3}/\varepsilon^{1/3})$ samples and using $O(n^{2/3}/\varepsilon^{4/3})$ queries.

Proof. The sub-routine CHECKHIDEGREE(M, ε) selects a random subset $\mathbf{U} \subset [n]$ of size $12/\varepsilon$ by independently sampling from $[n]$, and random subset $\mathbf{E} \subset [n] \times [n]$ of size $48n^{2/3}/\varepsilon^{1/3}$ by repeatedly sampling from $[n] \times [n]$. For each index $i \in \mathbf{U}$ and each pair of indices $(j,k) \in \mathbf{E}$, the sub-routine checks whether the triangle $\{i,j,k\}$ is a violating triangle of M. If it is a violating triangle, the sub-routine outputs "reject". Otherwise, it outputs "accept". The first item follows directly from the procedure as any matrix in $\mathcal{P}$ does not contain violating triangles. On the other hand, let $i \in [n]$ such that $d_T(i) \geq \varepsilon^{1/3} n^{4/3}/16$ be called the high-degree indices. In order for the algorithm to output "accept," it must avoid sampling a high-degree index, or if it does sample a high-degree index, must avoid the corresponding $\varepsilon^{1/3} n^{4/3}/16$ pairs of indices which form the violation. The probability this occurs is at most

$$\left(1 - \frac{\varepsilon}{4}\right)^{12/\varepsilon} + \left(1 - \frac{\varepsilon^{1/3}}{16n^{2/3}}\right)^{48n^{2/3}/\varepsilon^{1/3}} \leq 1/6.$$

∎

D Proof from Section 5 (Ultrametric and Tree Metric Testing Upper Bound)

LEMMA 5.3. *Let $M \in \mathcal{C}$ be ε-far from $\mathcal{P}^U$ and $S \subset [n]$ be consistent. If $|A(M,S)| \leq \varepsilon n^2/8$, then either there are at least $\varepsilon n/32$ inconsistent points in $[n] \setminus S$, or $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \geq \varepsilon n^2/8$.*

Proof. In Section 5, we have outlined the proof of Lemma 5.3. In particular, we want to show the contrapositive of the statement is true: given $M \in \mathcal{C}$ and a consistent $S \subset [n]$ which satisfies $|A(M,S)| \leq \varepsilon n^2/8$, $|\mathrm{SC}(M,S)| + |\mathrm{EC}(M,S)| \leq \varepsilon n^2/8$ and at most $\varepsilon n/32$ inconsistent points in $[n] \setminus S$, there exists an $n \times n$ matrix $\tilde{M} \in \mathcal{P}^U$ which differs from M on fewer than εn^2 entries. The matrix $\tilde{M}$ is constructed in Section 5 and it is shown that $\|\tilde{M} - M\|_0 < \varepsilon n^2$.

It remains to show that $\tilde{M} \in \mathcal{P}^U$. We do this by considering an arbitrary triple $i, j, k \in [n]$, and showing that it does not form a violation in $\tilde{M}$. First, if at least one of three i, j or k had been inconsistent with S, then because we set distances from this point to be arbitrarily large, the triangle i, j, k is no longer violated. A second easy case occurs when all $i, j, k \in S$. In this case, the consistency of S, and the fact $\tilde{M}_{|S\times S} = M_{|S\times S}$ implies there are no such violations.

- Suppose that $i, j \in S$ and that $k \in [n] \setminus S$, and in addition, k was consistent for S. In this case, all pairwise values among $\{i, j, k\}$ in $\tilde{M}$ are exactly the same as those in M. Since k was consistent for S, the set $S \cup \{k\}$ has no violations in M, and thus no violations in $\tilde{M}$.
- Suppose $i \in S$ and that $j, k \in [n] \setminus S$. We consider two sub-cases, according to whether or not j and k belong to the same part.
  - Suppose j and k belong to different parts. Then, Item 3 covers this case: If $(j, k) \notin \mathrm{SC}(M, S)$, then $M(j, k) = \max\{M(i', j), M(i', k)\}$ for all $i' \in S$, and hence also for i. Since $\tilde{M}(j, k) = M(j, k)$, this is not a violation. If $(j, k) \in \mathrm{SC}(M, S)$, then $\tilde{M}(j, k) = \max\{M(i', j), M(i', k)\}$ for some $i' \in \mathrm{SEP}(j, k)$ (which may be different from i). We assume without loss of generality that $M(i', k) > M(i', j)$. Note that $\{i, i', k\}$ and $\{i, i', j\}$ are not violating triples in $\tilde{M}$ as these are covered by Item D. Since j, k are consistent with S, $\tilde{M}(i', j) = M(i', j)$, $\tilde{M}(i', k) = M(i', k)$ for all $i' \in S$. We will now show that the maximum of $\tilde{M}(j, k)$, $\tilde{M}(i, j)$ and $\tilde{M}(i, k)$ is not unique, i.e., either
    $$\text{(i) } \tilde{M}(j, k) = \max\left\{\tilde{M}(i, j), \tilde{M}(i, k)\right\} \quad \text{or} \quad \text{(ii) } \tilde{M}(j, k) \leq \tilde{M}(i, j) = \tilde{M}(i, k).$$
    - Suppose $M(i', i) < M(i', j) < M(i', k)$. First, notice that $\tilde{M}(j, k) = \max\{M(i', j), M(i', k)\}$ by the second bullet in Item 3, which is equal to $M(i', k) = \tilde{M}(i', k)$ by assumption of this case. Because $\{i, i', k\}$ is not violating in $\tilde{M}$ and $\tilde{M}(i', k) = M(i', k) > M(i', i) = \tilde{M}(i', i)$, we must have $\tilde{M}(i', k) = \tilde{M}(i, k)$. Similarly, since the triple $\{i, i', j\}$ is not violating in $\tilde{M}$ and $\tilde{M}(i', j) = M(i', j) > M(i', i) = \tilde{M}(i', i)$, we must have $M(i', j) = \tilde{M}(i', j) = \tilde{M}(i, j)$. In summary, $\tilde{M}(j, k) = \tilde{M}(i', k) = \tilde{M}(i, k)$. By assumption, $M(i', k) > M(i', j) = \tilde{M}(i, j)$. Thus, (i) holds.
    - Suppose $M(i', i) = M(i', j) < M(i', k)$. By the same first step, $\tilde{M}(j, k) = M(i', k)$. Since the triple $\{i, i', k\}$ is not violating in M and $M(i', k) > M(i', i)$, we must have $M(i', k) = M(i, k)$. Moreover, since the triple $\{i, i', j\}$ is not violating in M and $M(i', j) = M(i', i)$, we must have $M(i', j) \geq M(i, j) = \tilde{M}(i, j)$. By assumption, $M(i', k) > M(i', j)$, which is at least $M(i, j) = \tilde{M}(i, j)$. In summary, $\tilde{M}(j, k) = M(i, k) = \tilde{M}(i, k)$, and $\tilde{M}(j, k) = M(i', k) > \tilde{M}(i, j)$. Thus, (i) holds.
    - Suppose $M(i', j) < M(i', i) < M(i', k)$. By the same first step, $\tilde{M}(j, k) = M(i', k)$. Since the triple $\{i, i', k\}$ is not violating in M and $M(i', k) > M(i', i)$, we must have $M(i', k) = M(i, k)$. Similarly, since the triple $\{i, i', j\}$ is not violating in M and $M(i', i) > M(i', j)$, we must have $M(i', i) = M(i, j)$. In summary, $\tilde{M}(j, k) = M(i', k) = M(i, k) = \tilde{M}(i, k)$, and by assumption, $M(i', k) > M(i', i)$ which is equal to $M(i, j) = \tilde{M}(i, j)$. Thus, (i) holds.
    - Suppose $M(i', j) < M(i', i) = M(i', k)$. By the same first step, $\tilde{M}(j, k) = M(i', k)$. Since the triple $\{i, i', j\}$ is not violating in M and $M(i', i) > M(i', j)$ by assumption, we must have $M(i', i) = M(i, j)$. Since the triple $\{i, i', k\}$ is not violating in M and $M(i', k) = M(i', i)$, we must have $M(i', k) \geq M(i, k)$. In summary, we have $\tilde{M}(k, j) = M(i', k) = M(i, j) = \tilde{M}(i, j)$, and we have $M(i', k) \geq M(i, k) = \tilde{M}(i, k)$. Thus, (i) holds.
    - Suppose $M(i', j) < M(i', k) < M(i', i)$. By the same first step, $\tilde{M}(j, k) = M(i', k)$. Since the triple $\{i, i', k\}$ is not violating in M and $M(i', k) < M(i', i)$, we must have $M(i', i) = M(i, k)$. Similarly, since the triple $\{i, i', j\}$ is not violating in M and $M(i', j) < M(i', i)$, we must have $M(i', i) = M(i, j)$. In summary, $\tilde{M}(k, j) = M(i', k) < M(i', i)$, which is equal to $M(i, k) = \tilde{M}(i, k)$ and also equal to $M(i, j) = \tilde{M}(i, j)$. Thus, (ii) holds.

    Thus $\{i, j, k\}$ is not violating in $\tilde{M}$.

– If j and k belong to the same part P, we do a case analysis based on the category of P.

∗ Suppose P is an Easy Part or an Active Part in M. Then $\tilde{M}(j,k)$ is set to be x. Since j,k are in the same part, $M(i',j) = M(i',k)$ for all $i' \in S$ so in particular, $M(i,j) = M(i,k)$. By definition, x is the minimum positive entry in M, so $M(i,j) = M(i,k) \geq x = \tilde{M}(j,k)$. Since j,k are consistent with S, by item 2, $\tilde{M}(i,j) = M(i,j)$ and $\tilde{M}(i,k) = M(i,k)$. Thus, the triple $\{i,j,k\}$ is not violating in $\tilde{M}$.

∗ Suppose P is a Versatile Part. Then $\tilde{M}(j,k)$ is set to be $M^*_P(j,k) \leq \min_{i' \in S}\{M(i',j)\} \leq M(i,j)$. Moreover, as j,k are in the same part, $M(i',j) = M(i',k)$ for all $i' \in S$, so in particular $M(i,j) = M(i,k)$. By same reason as above, $\tilde{M}(i,j) = M(i,j)$ and $\tilde{M}(i,k) = M(i,k)$. In summary, $\tilde{M}(j,k) \leq \tilde{M}(i,j) = \tilde{M}(i,k)$, so the triple $\{i,j,k\}$ is not violating in $\tilde{M}$.

- Suppose all $i,j,k \in [n] \setminus S$. We consider three sub-cases, according to how many different parts they were in.

– If i,j,k belonged to the same part P, we go into a case analysis on the category of P.

∗ Suppose P is an Easy Part or an Active Part in M. Then $\tilde{M}(i,j), \tilde{M}(j,k), \tilde{M}(i,k)$ are all set to be x. Thus, triple $\{j,k,l\}$ is not violating in $\tilde{M}$.

∗ Suppose P is a Versatile Part. $\tilde{M}(i,j), \tilde{M}(j,k), \tilde{M}(i,k)$ are set to be $M^*_P(i,j), M^*_P(j,k), M^*_P(i,k)$ respectively. By definition of M^*_P, this square matrix encodes an ultrametric, so the triple $\{j,k,l\}$ is non-violating in $\tilde{M}$.

– Suppose two of the indices i,j belonged to the same part P and k belonged to a different part P', then there exists a separator $v \in S$ such that $M(v,k) \neq M(v,i) = M(v,j)$. Moreover, as i,j,k are consistent with S, we must have $\tilde{M}(v,k) = M(v,k)$, $\tilde{M}(v,i) = M(v,i)$, and $\tilde{M}(v,j) = M(v,j)$. Now consider the triple $\{v,k,i\}$: it is not violating in $\tilde{M}$ because, by the argument above, any triple containing an index in S is not violating in $\tilde{M}$. Thus, because of the inequality $\tilde{M}(v,k) = M(v,k) \neq M(v,i) = \tilde{M}(v,i)$, we must have $\tilde{M}(i,k) = \max\{\tilde{M}(v,k), \tilde{M}(v,i)\}$. By the same argument, $\tilde{M}(j,k) = \max\{\tilde{M}(v,j), \tilde{M}(v,k)\}$. Thus, the chain of equality holds:

$$\tilde{M}(i,k) = \max\{\tilde{M}(v,k), \tilde{M}(v,i)\} = \max\{\tilde{M}(v,k), \tilde{M}(v,j)\} = \tilde{M}(j,k).$$

Now, we go into a case analysis on the category of P.

∗ Suppose P is an Easy Part or an Active Part. Then $\tilde{M}(i,j)$ is set to be x. Therefore, $\tilde{M}(i,k) = \tilde{M}(j,k) \geq \tilde{M}(v,k)$ by the chain of equality above, while $\tilde{M}(v,k) = M(v,k) \geq x$ as x is the minimum positive entry in M. Thus, $\tilde{M}(i,k) = \tilde{M}(j,k) \geq \tilde{M}(i,j)$.

∗ Suppose P is a Versatile Part. Then $\tilde{M}(i,j)$ is set to be $M^*_P(i,j)$, which is at most $\min_{i' \in S, j \in P}(M(i',j)) \leq M(v,j) = \tilde{M}(v,j)$. Therefore, $\tilde{M}(i,k) \leq \tilde{M}(v,j) \leq \tilde{M}(j,k) = \tilde{M}(i,k)$ by the chain of equality above.

All of the above show that $\{i,j,k\}$ are not-violating in $\tilde{M}$.

– Suppose i,j,k belonged to three different parts P_1, P_2, P_3 respectively. Then there exists a separator $v \in \mathrm{SEP}(i,j) \subset S$ such that $M(v,i) \neq M(v,j)$. We will now assume, without loss of generality, that $M(v,i) < M(v,j)$ (by re-naming i and j). By the above argument, $\{v,i,j\}$ contains one index in S and so is not violating in $\tilde{M}$. Moreover, $\tilde{M}(v,i) = M(v,i)$, $\tilde{M}(v,j) = M(v,j)$, and $\tilde{M}(v,k) = M(v,k)$ as i,j,k are consistent with S. Thus, we must have $\tilde{M}(i,j) = \max\{\tilde{M}(v,i), \tilde{M}(v,j)\} = M(v,j)$. Now, we go into a case analysis on how $M(v,k)$ compares with $M(v,i)$ and $M(v,j)$.

∗ Suppose $M(v,k) \neq M(v,i)$ and $M(v,k) \neq M(v,j)$. This implies $\tilde{M}(v,k) \neq \tilde{M}(v,i)$ and $\tilde{M}(v,k) \neq \tilde{M}(v,j)$ Then because triple $\{v,i,k\}$ contains an index in S, it is not violating in $\tilde{M}$ so we must have $\tilde{M}(i,k) = \max\{\tilde{M}(v,k), \tilde{M}(v,i)\} = \max\{M(v,k), M(v,i)\}$; similarly, triple $\{v,j,k\}$ is not violating in $\tilde{M}$ so we must have $\tilde{M}(k,j) = \max\{\tilde{M}(v,k), \tilde{M}(v,j)\} = \max\{M(v,k), M(v,j)\}$. If $M(v,k) > M(v,j)$, then $M(v,k) > M(v,i)$ as well by the

assumption $M(v,j) > M(v,i)$, so $\tilde{M}(i,k) = M(v,k)$ and $\tilde{M}(k,j) = M(v,k)$. Since $\tilde{M}(i,j) = M(v,j) < M(v,k)$, the triple $\{i,j,k\}$ is not violating in $\tilde{M}$. Now if $M(v,k) < M(v,j)$, then $\tilde{M}(k,j) = M(v,j)$. Since $\tilde{M}(i,k) = \max\{M(v,k), M(v,i)\} < M(v,j)$ and $\tilde{M}(i,j) = M(v,j)$, the triple $\{i,j,k\}$ is not violating in $\tilde{M}$.

* Suppose $M(v,k) = M(v,i)$. As $M(v,i) < M(v,j)$, $M(v,k)$ is less than $M(v,j)$ as well. These imply $\tilde{M}(v,k) = \tilde{M}(v,i) < \tilde{M}(v,j)$. Therefore, as the triple $\{v,k,j\}$ is not violating in $\tilde{M}$, we must have $\tilde{M}(k,j) = \tilde{M}(v,j) = M(v,j)$. Moreover, since the triple $\{v,i,k\}$ is not violating in $\tilde{M}$, we must have $i,k \leq \tilde{M}(v,k) = \tilde{M}(v,i) < \tilde{M}(v,j) = M(v,j)$. Therefore, the triple $\{i,j,k\}$ is not violating in $\tilde{M}$.
* Suppose $M(v,k) = M(v,j)$. As $M(v,i) < M(v,j)$, $M(v,k)$ is larger than $M(v,i)$ as well. These imply $\tilde{M}(v,i) < \tilde{M}(v,j) = \tilde{M}(v,k)$. Since the triple $\{v,i,k\}$ is not violating in $\tilde{M}$, we must have $\tilde{M}(i,k) = \tilde{M}(v,k) = \tilde{M}(v,j) = M(v,j)$. Since the triple $\{v,j,k\}$ is not violating in $\tilde{M}$, we must have $\tilde{M}(j,k) \leq \tilde{M}(v,j) = M(v,j)$. Together with $\tilde{M}(i,j) = M(v,j)$, the triple $\{i,j,k\}$ is not violating in $\tilde{M}$.

In summary, $\|M - \tilde{M}\|_0 < \varepsilon n^2$ and $\tilde{M} \in \mathcal{P}^U$, which implies that the contra-positive of the lemma statement is true. ∎