A. Theorem 1: Low Stretch Spanning Trees

Observe that in the construction of $RRCF(S)$ where the sum of the lengths of the sides of $B(S)$ is denoted by $P(S)$, the probability that we separated two points $\mathbf{x}(1), \mathbf{x}(2)$ by the first cut is proportional to the L_1 distance measure:

$$\frac{1}{P(S)} \sum_i |\mathbf{x}(1)_i - \mathbf{x}(2)_i|$$

Also observe that $P(S)$ decreases by at least a factor of $(1-\frac{1}{2d})$. To observe this, note that if we start with a rectangle where side i has length ℓ_i (and $\sum_i \ell_i = P(S)$ then the new expected perimeter of either side decreases by (the first term corresponds to choosing the dimension and the second term is the expected decrease in that dimension),

$$\sum_i \frac{\ell_i}{P(S)} \frac{\ell_i}{2} \geq \frac{\sum_i \ell_i^2}{2 \sum_i \ell_i} \geq \frac{\sum_i \ell_i}{2d}$$

Lemma 7 *Given a tree T in $RRCF(S)$ suppose we measure the distance between two points as follows: we find the first level where the points are separated and let the set be S. We assign the distance $P(S)$, i.e., the sum of the edge lengths of the minimum bounding box $B(S)$. Then the expected length of pair of points* $\mathbf{x}(1), \mathbf{x}(2)$ *is* $L_1(\mathbf{x}(1), \mathbf{x}(2))$ *times the expected number of steps to separate* $\mathbf{x}(1), \mathbf{x}(2)$. *Observe that we never assign a distance less than* $L_1(\mathbf{x}(1), \mathbf{x}(2))$.

Proof: Follows from the fact that the distance assigned at a level corresponding to S' is $P(S')$ and the probability of that distance assignment is $L_1(\mathbf{x}(1), \mathbf{x}(2))/P(S')$. The expected distance therefore is $L_1(\mathbf{x}(1), \mathbf{x}(2))$ times the expected number of steps that separate the two points! □

Remark: Note that the expected number of steps can be bounded by $O(d \log P(S)/L_1)$ since $P(S)$ decreases by a $(1 - \frac{1}{2d})$ factor in expectation. Other bounds can also be used – for example logarithm of the ratio of the total volume to the volume of the smallest box that contains $\mathbf{x}(1), \mathbf{x}(2)$ since in each step we divide the volume by $\frac{1}{2}$ in expectation.

B. Omitted Proofs

B.1. Proof of Lemma 1

We restate the lemma.

Lemma 8 *The expected displacement caused by a point x is the expected number of points in the sibling node of the leaf node containing x, when the partitioning is done according to the algorithm in Definition 1.*

Proof: In the absence of x, (in Figure 2b) the representation would be $q_0, \ldots, q_r, 0, \ldots$, in other words we would need 1 fewer bit to represent the point p. Therefore:

$$f(y,Z,T) - f(y, Z - \{x\}, T) = \begin{cases} 1 & y \in \text{sibling } c \text{ of } x \\ 0 & \text{otherwise} \end{cases}$$

The lemma follows. □

B.2. Proof of Lemma 2

We restate the lemma.

Lemma 9 $\textsc{CoDisp}(x, Z, |S|)$ *can be estimated efficiently.*

Proof: Observe that following the logic of Lemma 1, the difference

$$f(y,S,T) - f(y, S - C, T)$$

is nonzero for $y \in Z - C$ if and only iff we delete all the elements in a sibling subtree containing y. For example in Figure 2a, if $|b|$ is large, then the collusive displacement will be large only if we delete all the nodes in c along with x. Moreover to achieve any nonzero displacement we have to *simultaneously delete* all copies of x; and the result will scale down based on the number of duplicates. Observe that a consequence, given a T, we can compute

$$\max_{x \in C \subseteq S} \frac{1}{|C|} \sum_{y \in S - C} \Big(f(y,S,T) - f(y, S - C, T) \Big)$$

optimally by considering only the subtrees in the leaf to root path defined by x. □

B.3. Proof of Lemma 3

We restate the lemma.

Lemma 10 *Given point p and set of points S with an axis parallel minimal bounding box $B(S)$ such that $p \notin B$:*

(i) For any dimension i, the probability of choosing an axis parallel cut in a dimension i that splits S using the weighted isolation forest algorithm is exactly the same as the conditional probability of choosing an axis parallel cut that splits $S \cup \{p\}$ in dimension i, conditioned on not isolating p from all points of S.

(ii) Given a random tree of $RRCF(S \cup \{p\})$, conditioned on the fact the first cut isolates p from all points of S, the remainder of the tree is a random tree in $RRCF(S)$.

Proof: Consider the first part. Let the length of the minimum bounding box of S in dimension i be ℓ_i. Let the length of the minimum bounding box of $S \cup \{p\}$ in dimension i be ℓ'_i. Thus the probability (density) of choosing a cut C in dimension i that splits S is

$$\frac{1}{\ell_i} \frac{\ell_i}{\sum_i \ell_i}$$

where the first term is the probability density conditioned on the dimension and the second term is the probability of choosing dimension i. The probability density of achieving the same cut in constructing a weighted isolation forest of $S \cup \{p\}$ conditioned on not isolating p and S is

$$\frac{1}{\ell_i}\mathbb{P}r\left[\text{Choosing dimension } i | \text{not isolating } p \text{ and } S\right]$$

Probability of choosing dimension i **and** not isolating p and S is $\ell_i / \sum_i \ell'_i$. Therefore

$$\begin{aligned}&\mathbb{P}r\left[\text{Choosing dimension } i | \text{not isolating } p \text{ and } S\right]\\ =&\frac{\mathbb{P}r\left[\text{Choosing dimension } i \textbf{ and } \text{not isolating } p \text{ and } S\right]}{\mathbb{P}r\left[\text{not isolating } p \text{ and } S\right]}\\ =&\frac{\ell_i / \sum_i \ell'_i}{\sum_i \ell_i / \sum_i \ell'_i} = \frac{\ell_i}{\sum_i \ell_i}\end{aligned}$$

The last part follows from the observation that the probability of not isolating p and S is $\sum_i \ell_i / \sum_i \ell'_i$. This proves part (i). Note that part (ii) follows from construction. □

B.4. Proof of Lemma 4

We restate the lemma.

Lemma 11 *If T were drawn from the distribution $RRCF(S)$ then Algorithm 1 produces a tree T' which is drawn at random from the probability distribution $RRCF(S - \{p\})$.*

Proof: Given T was drawn from $RRCF(S)$ consider the random forest algorithm that would have produced this tree. Consider also the random forest algorithm as it produces T'. We will stochastically couple the decisions of the split operation – mirror the same split in T' as in T (Lindvall, 1992). Even though the splits across the two trees are correlated, if we consider only T or T', it would appear that the respective tree was produced with the right distribution. Of course the mirroring will not always be obvious, but we address that below. Initially we have a set $S' = S$. Consider the cases (a)–(b) below:

(a) Suppose that we choose the dimension i in splitting T and the point p does not lie on the bounding box of S' in dimension i. In this case the presence or absence of the point p does not affect the distribution of cuts are the same irrespective of the point set S' or $S' - \{p\}$. The construction of T' therefore can choose the same dimension i and the same cut as in T, and that could correspond to be a valid step with the correct probability. Note that after the cut, p can belong to only one side ,say S''. We will set $S' = S''$ and recurse. Note that we can use the same subtree (as in T) for the subset $S' - S''$ since there were no change to the point set. The construction of T' can completely mirror T for these subsets, and the construction will preserve the correct probabilities.

(b) Otherwise, we choose dimension i in splitting T and point p lies on the bounding box of S' in dimension i. We now have two cases.

   (i) Point p is separated from rest of S'. In this case T produces a sibling tree $T(u)$ starting at node u which is the sibling node of node v containing the isolated point p. But then in T' we do not have p, and $T(u)$ is a random tree from $RRCF(S' - \{p\})$ and the construction is correct using part (ii) of Lemma 3.

   (ii) Point p is not separated from S'. By Lemma 3, conditioned on the fact that p is not separated from S' we are choosing a random cut which separates $S' - \{p\}$ along a chosen dimension i. This is therefore an appropriate choice for T' using part (i) of Lemma 3 and we choose the same cut in T'. Again we have two subsets, and in T, p belongs to only one side. We recurse on that side – for the other side the construction of T, T' can be identical since they have the same set of points.

The above cases are mutually exclusive and exhaustive. This proves the Lemma. □

B.5. Proof of Lemma 6

We restate the lemma.

Lemma 12 *If T were drawn from the distribution $RRCF(S)$ then Algorithm 1 produces a tree T' which is drawn at random from the probability distribution $RRCF(S \cup \{p\})$.*

Proof: We proceed in a manner similar to the proof of Lemma 4 — *however instead of using tree T to define the splits for T', we will first make a decision about T' and then mirror T.* Suppose that we have currently the set of nodes S'. Note that the case of $S' = \emptyset$ is trivial. Therefore assume $S' \neq \emptyset$ and we are given a tree $T(S')$ from $RRCF(S')$.

(a) If we decide to separate p and S' then Step 6 in Algorithm 2 generates such a cut. Now after the cut, we observe that $T(S')$ is already a random tree in $RRCF(S')$ and we can simply use $T(S')$ to define T'.

(b) If we decide to **not** separate p and S' then using part (i) of Lemma 3, we can choose any cut that splits the bounding box $B(S')$. *Note that the first cut in $T(S')$ is exactly such a cut chosen with the correct distribution.* Therefore we can use the same cut in T' in this case. Note that we now have two sides and p only affects one side – we can use the same subtree as in $T(S')$ for the side that does not contain p. We the side S'' that contains p. Note that we recursively maintain the property that we have a random tree $T(S'')$ from $RRCF(S'')$.

The above steps are mutually exclusive and exhaustive. This proves the Lemma. □

C. Depth and Co-Displacement

As remarked in Section 4.1 the introduction, it is unclear why the (shallowness of) depth in a random cut forest is the best possible predictor of anomalies. We show another example to illustrate that the depth provides us a very noisy signal. Consider the data in Figure 6 which shows an empty donut – the exterior has 2500 points (generated from a Gaussian with mean as the origin and standard deviation $\sqrt{2}$ and we remove all points within distance 3 from origin till we get 2500 points). The interior circle has 2500 points from a Gaussian around the origin with standard deviation $\sqrt{0.2}$ and we remove all points which are at a distance larger than 1. We then add the single outlier point $(1.5, 0)$.

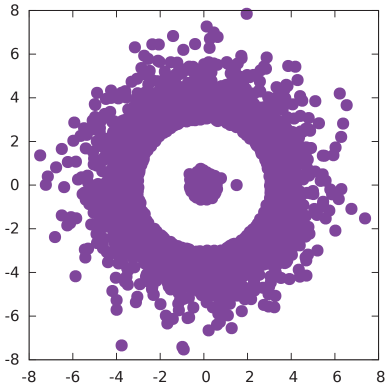

Figure 6. Another simple example dataset

Figure 7 shows the two other points which have larger CoDisp() than the $(1.5, 0)$ point. The legend shows the amount of the collusive displacement.

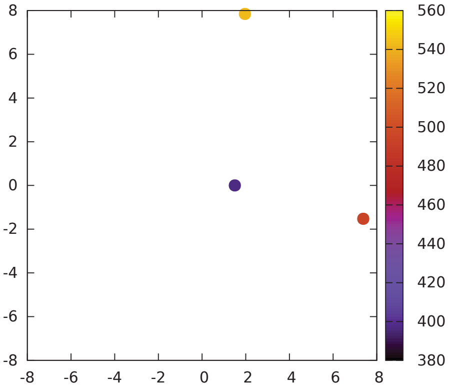

Figure 7. CoDisp() above the point $(1.5, 0)$

Figure 8 shows the 87 points (in a particular run) with expected depth below that of $(1.5, 0)$ when we constructed the trees with the recursive bias. The points are distributed as we would expect in the periphery – any distribution there will be points which are far from the central point, but they are not necessarily outliers.

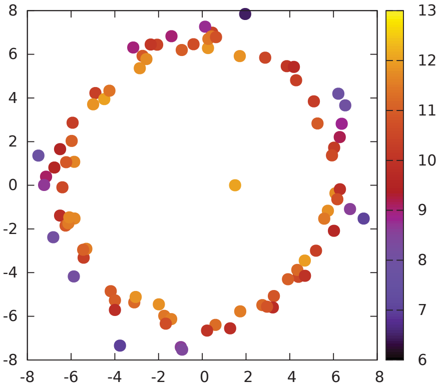

Figure 8. $g()$ above the point $(1.5, 0)$

The notion of "surprise" which is captured by (large) CoDisp() is not the same as the (shallow) expected depth. In fact the sum of these two quantities (almost, given the difference between Disp() and CoDisp() – but the Gaussian generation ensures no duplicates) correspond to the total increase in model complexity as discussed in Definition 2. Depth alone does not appear to be a sufficient statistic. Of course this leaves open the possibility of a non-parametric model-based definition that incorporate both the expected depth and the collusive displacement. If we were to assign costs to the bits representations then a case can be argued for CoDisp()$/g()$, however pinning that quantity to a more fundamental notion likely requires substantial work which is left for the future.