PIXELSCORE 1.0: A principled approach to global NFT rarity estimation.

Dmitry Storcheus
@dstorcheus

Adi Kancherla
@nneverlander

Abstract

This paper introduces PIXELSCORE 1.0, a novel algorithm for estimating the global rarity of NFTs. In contrast to existing *local* trait-based rarity ranking methods, which compare the rarities of tokens only within a given collection, our algorithm is *global*. It compares the rarity of any pair of NFTs from any collection and ranks all existing NFTs by pixel rarity. The algorithm is based on a statistical analysis of raw pixels of NFTs and a novel probabilistic concept of *pixel set rarity* that we introduce in this paper. Our algorithm is *objective* because it relies purely on a mathematical formula. We provide an extensive empirical study of the algorithm on a sample of more than 55,000 collections with over 10 million NFTs, revealing the most and least rare NFTs according to our analysis.

1 Motivation

Traditionally, an NFT's rarity is computed based on the average rarity of its traits [1, 2], and the rarity of a given NFT is the total probability of occurrence of its current traits. Another method of computing rarity is the Jaccard index [3–5] used, for example, by NFTGo. A major drawback of this approach is its quadratic complexity. Using it to compute rarities for all existing NFTs quickly becomes infeasible as they grow in number. As we show in Section 3, our algorithm scales linearly using a distributed histogram estimation method and can be used to estimate rarities for any number of NFTs.

Our method is a fundamentally different yet extremely simple approach to estimating NFT rarity. It is based not on traits but rather on the raw pixels that make up the NFT. The benefits of the novel theoretical framework and the PIXELSCORE algorithm proposed in this paper include but are not limited to:

- *Objectivity*. The rarity score is based on a clear mathematical formula and bare-bones NFT data, which makes it free from biases, market trends and community sentiment created by influencers, forums, discord groups, and creators.
- *Generality*. Because our algorithm is based on raw pixels, it applies to abstract NFTs, other types of art and any digitally represented visual objects. Unlike trait-based methods, our algorithm does not require NFTs to have distinguishable object traits (hats, glasses, etc.). Furthermore, unlike asset-based methods, our algorithm does not require the tokens to be traded.
- *Cross-Collection Compatibility*. A significant benefit of our approach is that unlike trait-based rarity scores, which are tied to a particular collection, we can use pixel-based scores to estimate *global* NFT rarity independent of specific collection traits. This fundamentally allows us to compare any pair of NFTs in existence by rarity and rank all existing NFTs on all blockchains.
- *Scalability*. PIXELSCORE allows obtaining rarity scores and rankings for all existing NFTs. The algorithm scales linearly with growing input set size.

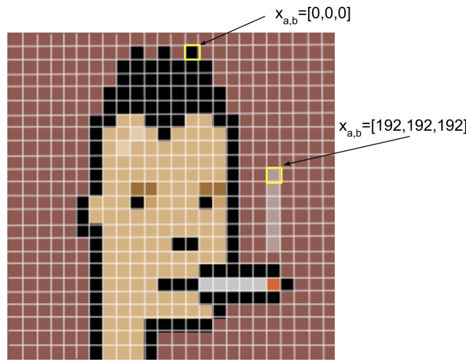

Figure 1: An illustration of the matrix representation of CryptoPunk 7124 [6]

- *Stability*. As we prove in Section 2, the empirical rarity estimate used by PIXELSCORE is close to the theoretical rarity with high probability. Moreover, the empirical estimate converges to the theoretical one at a rate of $1/m$, where m is the sample size, which guarantees the robustness of our approach.

While this paper analyzes the rarity of NFTs specifically, the same mathematical model can be applied to any visual objects, including cat pictures, Instagram posts and digital representations of physical art. PIXELSCORE refers to both the probabilistic framework of rarity analysis presented in this paper and the algorithm that uses such a framework to compute a rarity score and rank for a given NFT, presented in Section 3. The rarity scores and ranks produced by the PIXELSCORE algorithm for a single NFT or collection of NFTs are referred to as *pixelscores* and *pixelranks*.

The rest of the paper is structured as follows: in Section 2, we introduce the probabilistic concept of rarity behind our algorithm. In Section 3, we formulate the PIXELSCORE algorithm and show how it computes the rarity scoring function from Section 2 on real data. Section 4 presents the results of an extensive empirical study where we used PIXELSCORE to rank over 10 million NFTs. Additionally, we reveal the most and least rare NFTs according to our analysis and present several insights about the distribution of individual pixels in the collections that we analyzed.

2 Pixel Set Rarity: Theoretical Background

In this section, we introduce the mathematical model behind the PIXELSCORE algorithm. Our model is based on novel probabilistic concepts of rarity: *Set Rarity*, *Pixel Rarity* and *NFT Rarity*. We build upon these concepts to derive the theoretical rarity scoring function $\mathcal{R}(\mathbf{z})$ for an NFT $\mathbf{z}$ and its empirical estimate $\widehat{R}(\mathbf{z})$. Further, we demonstrate the stability guarantees of such a rarity measure by showing that the estimate $\widehat{R}(\mathbf{z})$ obtained from a random sample is close to $\mathcal{R}(\mathbf{z})$ with high probability, under weak assumptions imposed on the true pixel data distribution.

2.1 Pixel distribution

Pixel rarity is inversely proportional to the frequency of occurrence of that pixel with respect to the population distribution. Therefore, we start this section with a probabilistic model of the pixel space that leads to the formulation of rarity.

Let $x \in \mathbb{R}^3$ be a 3-dimensional RGB pixel vector with its $[\text{red}, \text{green}, \text{blue}]$ components. Because pixel values are bounded by 255, the actual pixel space $\mathbb{X} \subset \mathbb{R}^3$ is the hypercube $\mathbb{X} = [0, 255]^3$. For a square image with maximum dimensions $d \times d$, let $x_{a,b}$ be its pixel located in the a-th row and the b-th column. Any NFT is represented as a $d \times d$ matrix of pixels $\mathbf{x} = \{x_{a,b} : 1 \leq a, b \leq d\}$, which is illustrated in Figure 1. An NFT *collection* is a set of m NFTs $C_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. We use a separate definition for a sample of NFTs $S_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, which can include tokens from different collections. We use $i \in [1, m]$ to index sampled NFTs $\mathbf{x}_i$ and $a, b \in [1, d]$ to index individual pixels $[\mathbf{x}_i]_{a,b}$ of these NFTs located at row a and column b. To estimate its rarity, an NFT token is denoted as $\mathbf{z}$. Note that $\mathbf{z}$ can be either in the sample S_m or out of the sample. By separating $\mathbf{z}$ from

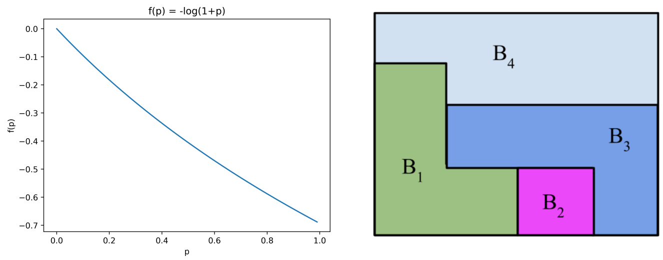

Figure 2: Left: Plot of the $f(p) = -\log(1+p)$ for $p \in [0, 1]$. Right: Decomposition of a set $\mathbb{X}$ into disjoint subsets $\mathbb{X} = B_1 \cup B_2 \cup B_3 \cup B_4$.

$\mathbf{x}$, we make it clear to the reader that a sample $S_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is explored to estimate the rarity distribution of the entire pixel space, which is then used to provide the rarity score for $\mathbf{z}$.

We denote the population and empirical distributions of an individual pixel $x_{a,b}$ by $\mathcal{D}_{a,b}$ and $\widehat{\mathcal{D}}_{a,b}$. Analyzing such distributions plays a critical role in rarity scoring of NFTs by the PIXELSCORE algorithm introduced in the following sections. Let $B \in \mathbb{X}$ be a bounded set (i.e., hypercube $[0, 1]^3$). The empirical probability measure of B, denoted as $\widehat{P}_{a,b}(B) = \widehat{P}_{a,b}\big(x_{a,b} \in B\big)$ for pixel $x_{a,b}$ under empirical distribution $\widehat{\mathcal{D}}_{a,b}$, is estimated using a sample $S_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ as follows:

$$\widehat{P}_{a,b}(B) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{[\mathbf{x}_i]_{a,b} \in B}, \tag{1}$$

where $\mathbb{1}_{(\cdot)}$ is an indicator function of the underlying event. $\widehat{P}_{a,b}(B)$ is simply the fraction of NFTs in the sample S_m in which the individual (a, b)-th pixel falls within the set B.

2.2 Rarity

We start by introducing the concepts of the rarity of a single set in Definition 1 and the rarity of an individual pixel in Definition 2. We build upon the concepts above to introduce the rarity of an NFT in Definition 3, which can be easily extended to sets of an arbitrary number of NFTs (collections, wallets, etc.). We start with *frequency*, which can be understood as being inversely proportional to *rarity*. Recall that the population distribution $\mathcal{D}_{i,j}$ introduced in the previous section serves as a frequency measure—it can tell us the probability that a given pixel falls within some interval.

Now, we argue that the *rarity* of the set B is inversely proportional to its probability mass, per the following definition:

Definition 1 (Set Rarity)**.** *Let $x_{a,b} \sim \mathcal{D}_{a,b}$ be an individual pixel for any set $B \in \mathbb{X}$. Then the rarity of B w.r.t pixel $x_{a,b}$ is*

$$\mathcal{R}_B(x_{a,b}) = -\log\left(1 + P_{a,b}(B)\right), \tag{2}$$

where $P_{a,b}(B)$ is the probability mass of B under $\mathcal{D}_{a,b}$.

The function $f(p) = -\log(1+p)$ is strictly decreasing and bounded for $p \in [0, 1]$, which provides a convenient and intuitive description of rarity: the smaller the probability mass p, the higher the rarity (see Figure 2). The concept of $\mathcal{R}_B(x_{a,b})$ describes the rarity of a particular subset B of the pixel space $\mathbb{X}$ in terms of how *unlikely* a random pixel $x_{a,b}$ drawn from the population distribution $\mathcal{D}_{a,b}$ falls within B. The rarity of the same set B can be different with respect to distinct individual (a, b)-th and (a', b')-th pixels because, in general, $\mathcal{D}_{a,b} \neq \mathcal{D}_{a',b'}$.

Given a decomposition (bucketization) of the space $\mathbb{X}$ into a union of p discount subsets $\mathbb{X} = \cup_{k=1}^{p} B_k$, we can describe the rarity of an individual pixel $[\mathbf{z}]_{a,b}$ that belongs to some NFT token $\mathbf{z}$ in terms of set rarities of $B_1, \ldots, B_p$.

Definition 2 (Pixel Rarity)**.** *Let $[\mathbf{z}]_{a,b} \in \mathbb{X}$ be an individual pixel of an NFT token $\mathbf{z}$. For any sequence of p disjoint subsets $\{B_k : 1 \leq k \leq p\}$ s.t. $\mathbb{X} = \cup_{k=1}^{p} B_k$, the rarity of pixel $[\mathbf{z}]_{a,b}$ is*

$$\mathcal{R}([\mathbf{z}]_{a,b}) = \sum_{k=1}^{p} \mathcal{R}_{B_k}(x_{a,b}) \mathbb{1}_{[\mathbf{z}]_{a,b} \in B_k} = -\sum_{k=1}^{p} \log\big(1 + P_{a,b}(B_k)\big) \mathbb{1}_{[\mathbf{z}]_{a,b} \in B_k}, \tag{3}$$

where $P_{a,b}(B_k)$ is the probability mass of B_k under $\mathcal{D}_{a,b}$. The indicator function $\mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k}$ is 1 if $[\mathbf{z}]_{a,b}$ belongs to B_k and 0 otherwise.

Because the subsets B_k are disjoint, the pixel rarity $\mathcal{R}([\mathbf{z}]_{a,b})$ can be interpreted as the set rarity of that $B_{k'}$, which includes $[\mathbf{z}]_{a,b}$.

Definition 3 (NFT Rarity)**.** *Let $\mathbf{z} = \{[\mathbf{z}]_{a,b} \in \mathbb{X} : 1 \le a, b \le d\}$ be an NFT. For any sequence of p disjoint subsets $\{B_k : 1 \le k \le p\}$ s.t. $\mathbb{X} = \cup_{k=1}^{p} B_k$, the rarity of $\mathbf{z}$ is*

$$\mathcal{R}(\mathbf{z}) = \frac{1}{d^2} \sum_{a,b=1}^{d} \mathcal{R}([\mathbf{z}]_{a,b}) = -\frac{1}{d^2} \sum_{a,b=1}^{d} \sum_{k=1}^{p} \log\left(1 + P_{a,b}(B_k)\right) \mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k}, \tag{4}$$

where $P_{a,b}(B_k)$ is the probability mass of B_k under $\mathcal{D}_{a,b}$. The indicator function $\mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k}$ is 1 if $[\mathbf{z}]_{a,b}$ belongs to B_k and 0 otherwise.

Definitions 1–3 above provide a rigorous and yet extremely flexible statistical framework to rank all existing NFTs and collections on any blockchain. They allow, for example, computing the rarity of any NFT collection $C_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ as the average rarity of its NFTs: $\mathcal{R}(C_m) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{R}(\mathbf{x}_i)$. In a similar way, one can evaluate the rarity of an NFT portfolio, wallet or any other set of distinct tokens.

2.3 Empirical rarity estimation

The NFT rarity $\mathcal{R}(\mathbf{z})$ in Definition 3 is based on the population distributions of pixels $\mathcal{D}_{a,b}$ for all $1 \le a, b \le d$. When such distributions are unknown, we provide an empirical rarity estimate $\widehat{\mathcal{R}}(\mathbf{z})$ based on a random sample of NFTs S_m of size m.

$$\widehat{\mathcal{R}}(\mathbf{z}) = -\frac{1}{d^2} \sum_{a,b=1}^{d} \sum_{k=1}^{p} \log\left(1 + \widehat{P}_{a,b}(B_k)\right) \mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k}, \tag{5}$$

where $\widehat{P}_{a,b}(B_k) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{[\mathbf{x}_i]_{a,b}\in B_k}$.

2.4 Stability

The rarity concepts introduced in Definitions 1–3 satisfy asymptotic stability guarantees, which may not hold for other rarity notions, such as trait-based or asset-based rarity. In other words, the empirical rarity estimate $\widehat{\mathcal{R}}(\mathbf{z})$ is close to the theoretical rarity $\mathcal{R}(\mathbf{z})$ with high probability. This means that for sufficiently large samples of NFTs, empirical rarity is a reliable estimate of the true population rarity in the global space of all NFTs.

Theorem 1 (Stability)**.** *Let $\mathbf{z} = \{[\mathbf{z}]_{a,b} \in \mathbb{X} : 1 \le a, b \le d\}$ be an NFT and $S_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be an i.i.d. sample used to compute $\widehat{\mathcal{R}}(\mathbf{z})$. Then, for any $\delta > 0$ with probability at least $1 - \delta$, the following inequality holds:*

$$\left|\widehat{\mathcal{R}}(\mathbf{z}) - \mathcal{R}(\mathbf{z})\right| \le \frac{1}{m} + \frac{\log 2}{d}\sqrt{\frac{\log \frac{\delta}{2}}{2}} \tag{6}$$

Proof. Using triangle inequality

$$\left|\widehat{\mathcal{R}}(\mathbf{z}) - \mathcal{R}(\mathbf{z})\right| \le \left|\widehat{\mathcal{R}}(\mathbf{z}) - \mathbb{E}\left[\widehat{\mathcal{R}}(\mathbf{z})\right]\right| + \left|\mathbb{E}\left[\widehat{\mathcal{R}}(\mathbf{z})\right] - \mathcal{R}(\mathbf{z})\right|,$$

we will bound each of the terms $\left|\widehat{\mathcal{R}}(\mathbf{z}) - E\left[\widehat{\mathcal{R}}(\mathbf{z})\right]\right|$ and $\left|E\left[\widehat{\mathcal{R}}(\mathbf{z})\right] - \mathcal{R}(\mathbf{z})\right|$ separately.

$$\mathbb{E}(\widehat{\mathcal{R}}(\mathbf{z})) = -\frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k=1}^{p}\mathbb{E}\Big[\log\Big(1+\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{[\mathbf{x}_i]_{a,b}\in B_k}\Big)\Big]\mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k} \tag{7}$$

$$= -\frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k:[\mathbf{z}]_{a,b}\in B_k}\mathbb{E}\Big[\log\Big(1+\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{[\mathbf{x}_i]_{a,b}\in B_k}\Big)\Big] \tag{8}$$

$$= -\frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k:[\mathbf{z}]_{a,b}\in B_k}\mathbb{E}\Big[\log\Big(1+\widehat{P}_{a,b}(B_k)\Big)\Big] \tag{9}$$

$$\approx \frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k:[\mathbf{z}]_{a,b}\in B_k}\left[\log(1+P_{a,b}(B_k))) - \frac{\mathbb{V}\big(\widehat{P}_{a,b}(B_k)\big)}{2(1+\mathbb{E}\big(\widehat{P}_{a,b}(B_k)\big))^2}\right] \tag{10}$$

$$= \frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k:[\mathbf{z}]_{a,b}\in B_k}\left[\log(1+P_{a,b}(B_k))) - \frac{1}{m}\frac{P_{a,b}(B_k)(1-P_{a,b}(B_k))}{2(1+P_{a,b}(B_k))^2}\right] \tag{11}$$

$$= \mathcal{R}(\mathbf{z}) - \frac{1}{m}\sum_{i=1}^{m}\frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k:[\mathbf{z}]_{a,b}\in B_k}\frac{P_{a,b}(B_k)(1-P_{a,b}(B_k))}{2(1+P_{a,b}(B_k))^2}. \tag{12}$$

Step (10) above follows from a second-order Taylor expansion [7] of $\log(1+x)$ around $x_0 = \mathbb{E}[x]$, which yields $\mathbb{E}\big[\log(1+x)\big] \approx \log\big(1+\mathbb{E}[x]\big) - \frac{\mathbb{V}[\mathrm{x}]}{2(1+\mathbb{E}[\mathrm{x}])^2}$. Because $P_{a,b}(B_k) \in [0,1]$, $P_{a,b}(B_k)(1-P_{a,b}(B_k)) \leq 1$ and $2(1+P_{a,b}(B_k))^2 \geq 1$, we conclude that the fraction $\frac{P_{a,b}(B_k)(1-P_{a,b}(B_k))}{2(1+P_{a,b}(B_k))^2} \leq 1$, which leads us to

$$\big|\widehat{\mathcal{R}}(\mathbf{z}) - \mathbb{E}\big[\widehat{\mathcal{R}}(\mathbf{z})\big]\big| \leq \frac{1}{m}.$$

$$\widehat{\mathcal{R}}(\mathbf{z}) = -\frac{1}{d^2}\sum_{a,b=1}^{d}\sum_{k=1}^{p}\log\big(1+\widehat{P}_{a,b}(B_k)\big)\mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k}$$
$$= \sum_{a,b=1}^{d} Y_{a,b}$$

$Y_{a,b} = -\frac{1}{d^2}\sum_{k=1}^{p}\log\big(1+\widehat{P}_{a,b}(B_k)\big)\mathbb{1}_{[\mathbf{z}]_{a,b}\in B_k}$ Because $Y_{a,b} \in [0, \log 2/d^2]$, by Hoeffding's [8, 9] inequality

$$P\big(\big|\widehat{\mathcal{R}}(\mathbf{z}) - E[\mathcal{R}(\mathbf{z})]\big| \geq \epsilon\big) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{a,b}(\log 2/d)^2}\right) = \exp\left(\frac{-2\epsilon^2 d^2}{(\log 2)^2}\right)$$

Letting $\delta = \exp\left(\frac{-2\epsilon^2 d^2}{(\log 2)^2}\right)$ and solving for ϵ, we conclude that the following inequality applies for any $\delta > 0$ with probability at least $1-\delta$:

$$\big|\widehat{\mathcal{R}}(\mathbf{z}) - E[\mathcal{R}(\mathbf{z})]\big| \leq \frac{\log 2}{d}\sqrt{\frac{\log\frac{\delta}{2}}{2}}$$

Combining with the previous result, we get the following concentration inequality, which completes the proof:

$$\big|\widehat{\mathcal{R}}(\mathbf{z}) - \mathcal{R}(\mathbf{z})\big| \leq \big|\widehat{\mathcal{R}}(\mathbf{z}) - \mathbb{E}\big[\widehat{\mathcal{R}}(\mathbf{z})\big]\big| + \big|\mathbb{E}\big[\widehat{\mathcal{R}}(\mathbf{z})\big] - \mathcal{R}(\mathbf{z})\big| \leq \frac{1}{m} + \frac{\log 2}{d}\sqrt{\frac{\log\frac{\delta}{2}}{2}}.$$

$\square$

2.5 Embeddings

In some cases, statistical analysis in the original pixel space, endowed with standard Euclidean distance metric $\|\mathbf{x} - \mathbf{x}'\|$, may not be sufficient to describe the similarity of shapes in an image. Thus, we map raw pixels to the *embedding space* Φ, endowed with an inner product $\langle \cdot\ ,\ \cdot \rangle_\Phi$ and the distance metric $\| \cdot \|_\Phi$ induced by such inner product. We denote the population and empirical distributions of the embedding maps $\Phi(x_{a,b})$ of individual pixels $x_{a,b}$ by $\Phi(\mathcal{D}_{i,j})$ and $\widehat{\Phi}(\mathcal{D}_{i,j})$.

More specifically, let $\Phi(B)$ be any bounded subset of the embedding space Φ with its preimage B, that is, $B = \{x_{a,b} \in \mathbb{X} : \Phi(x_{a,b}) \in \Phi(B)\}$ The probability mass of $\Phi(B)$ under the population distribution $\mathcal{D}_{a,b}$ is $P(\Phi(B))$. $P(\Phi(B))$ is interpreted as the probability that an embedding of a randomly chosen pixel falls within the set $\Phi(B)$. The PIXELSCORE algorithm can be extended to the embedding space, which requires a decomposition of such space into disjoint subsets $\Phi = \cup_{k=1}^{p} \Phi(B_k)$.

2.6 How to interpret pixelscores

The PIXELSCORE algorithm can be used to determine the rarity of:

- an NFT within a collection
- an NFT globally
- an entire collection
- a portfolio or wallet

We refer to the rarity estimates $\mathcal{R}(\mathbf{z})$ produced by the algorithm for sets of NFT tokens listed above as *pixelscores*. For each item (an NFT or a collection), we also derive a *binned pixelscore* and a *pixelrank* based on its pixelscore value. While pixelscores are continuous within a bounded range and can be normalized to [0, 1], binned pixelscores and pixelranks are easy to interpret natural numbers. The binned pixelscores are obtained by binning $\mathcal{R}(\mathbf{z})$ into 10 buckets, where the binned pixelscore equal to 10 is assigned to the rarest tokens, and the binned pixelscore equal to 1 is assigned to the least rare ones. While pixelscores and binned pixelscores are directly proportional (i.e., rare items have higher scores), the pixelscores and pixelranks are inversely proportional. The higher the pixelscore, the lower the pixelrank (i.e., rare items have lower pixelranks). For clarity and consistency, the rarest NFT will have a pixelrank of 1.

PIXELSCORE($S_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}, \mathbf{z}$)

1 $\cup_{k=1}^{p} B_k = \mathbb{X} \leftarrow$ Partition pixel space
2 **for** $k \leftarrow 1$ **to** p **do**
3   **for** $a, b \leftarrow 1$ **to** d **do**
4         $\widehat{P}_{a,b}(B_k) \leftarrow \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{[\mathbf{x}_i]_{a,b} \in B_k}$
5     $\widehat{\mathcal{R}}_{B_k}(x_{a,b}) \leftarrow -\log(1 + \widehat{P}_{a,b}(B_k))$
6 **for** $a, b \leftarrow 1$ **to** d **do**
7     $\widehat{\mathcal{R}}([\mathbf{z}]_{a,b}) \leftarrow \sum_{k=1}^{p} \widehat{\mathcal{R}}_{B_k}(x_{a,b}) \mathbb{1}_{[\mathbf{z}]_{a,b} \in B_k}$
8 $\widehat{\mathcal{R}}(\mathbf{z}) \leftarrow \frac{1}{d^2} \sum_{a,b=1}^{d} \widehat{\mathcal{R}}([\mathbf{z}]_{a,b})$
9 **return** $\widehat{\mathcal{R}}(\mathbf{z})$

Figure 3: PIXELSCORE pseudocode. Inputs: S_m, a sample of NFTs of size m and an out-of-sample NFT $\mathbf{z}$, the rarity of which needs to be calculated. Returns: $\widehat{\mathcal{R}}(\mathbf{z})$, the rarity estimate of $\mathbf{z}$.

3 PIXELSCORE 1.0 Algorithm

The PIXELSCORE algorithm is presented in Figure 3. It relies on the theoretical concepts of rarity introduced in Section 2 and uses the empirical rarity estimate

$$\widehat{\mathcal{R}}(\mathbf{z}) = -\frac{1}{d^2} \sum_{a,b=1}^{d} \sum_{k=1}^{p} \log\left(1 + \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{[\mathbf{x}_i]_{a,b} \in B_k}\right) \mathbb{1}_{[\mathbf{z}]_{a,b} \in B_k} \tag{13}$$

obtained on an i.i.d. sample of NFTs of size m.

The PIXELSCORE algorithm takes two inputs: S_m, an i.i.d. sample of NFTs of size m and an out-of-sample NFT $\mathbf{x}$, the rarity of which needs to be calculated. The algorithm returns the rarity estimate of $\mathbf{x}$. The algorithm can be trivially extended to return the rarities of an arbitrary set of NFTs in a scalable way. In Step 1, the algorithm partitions the pixel space $\mathbb{X}$ into a union of p disjoint subsets $B_1, \ldots, B_p$. In Steps 2–5, it computes the probability mass estimate $\widehat{P}_{a,b}(B_k)$ for each individual (a, b)-th pixel and each set B_k. Steps 6–8 compute individual pixel rarity estimates for $[\mathbf{z}]_{a,b}$ and aggregate them to obtain the NFT rarity estimate $\widehat{\mathcal{R}}(\mathbf{z})$. To compute the rarity of another NFT $\mathbf{z}'$ or a new set of NFTs, the saved results of Steps 1–5 can be reused; only Steps 6–8 need to be executed in that case. Step 4 can be carried out in an asynchronous distributed way over batches of the input sample S_m.

For the sake of simplicity, we present the pseudocode in Figure 3 in the original pixel space $\mathbb{X}$; however, it can be extended in a straightforward way to the embedding space, which will require decomposing $\Phi(\mathbb{X})$, the image of $\mathbb{X}$ under embedding map Φ into disjoint subsets B_k, instead of the original space $\mathbb{X}$.

4 Empirical Study: Mysterious Green Square and Other Rare NFTs

We have conducted an extensive empirical study on more than 55,000 collections with a total of roughly 10 million NFTs. We have executed the PIXELSCORE algorithm on these NFTs and identified the most and least rare collections and individual tokens.

The image space embedding was learned using a custom convolutional neural network (CNN) [10] architecture initialized from ImageNet. Each NFT was resized to 224×224 pixels and converted to RGB when necessary. Additional preprocessing was conducted, including missing data cleaning and removing outliers. The embedding space was broken down into 1000 disjoint subsets (bins) $B_1, \ldots, B_{1000}$ using scalable k-means++ clustering [11]. The empirical probability distribution of each pixel was computed using the image data of 10 million NFTs with the help of a distributed histogram learning algorithm [12]. Executing the partition of the space into subsets for larger samples

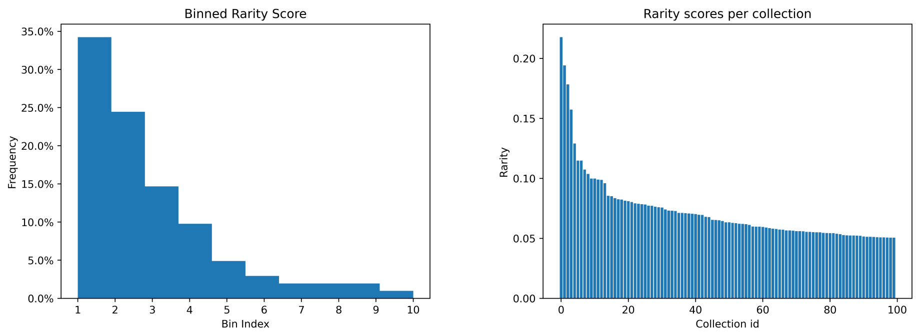

Figure 4: Left: Histogram of the binned rarity index I computed on a sample of 10 million tokens, where $I(\mathbf{x}) = 10$ is the most rare, and $I(\mathbf{x}) = 1$ is the least rare. **Right**: Average rarity scores for the top 100 collections.

Figure 5: Random selection of the rarest 1% of the 10 million NFTs sampled.

of NFTs (i.e., more than 250 million) would require either performing the clustering on a smaller sub-sample, then generalizing it to a full data set or applying a random partition of the full data.

Using the empirical probability distribution as computed above, we executed the PIXELSCORE algorithm and computed the rarity score for each NFT from the 10-million set and ranked each NFT by rarity from Pixelrank 1 to Pixelrank 10 million. This provides a global cross-collection rarity ranking, which is a key feature of PIXELSCORE compared to other rarity estimation methods.

In addition to the real-valued rarity estimate $\mathcal{R}(\mathbf{z})$, we provide an integer-valued rarity index $I(\mathbf{z}) \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, which is obtained by binning $\mathcal{R}(\mathbf{z})$ into 10 buckets, where $I(\mathbf{z}) = 10$ corresponds to the most rare tokens and $I(\mathbf{z}) = 1$ to the least rare. Such an index is easy to interpret and visualize. The histogram or the rarity index based on our data sample is provided in Figure 4.

Our empirical study shows exciting results and deep insight into the global rarity structure of the NFT space and reveals the most and least rare NFTs in this space. Figure 5 presents a random selection of the top 1% rarest NFTs in the 10-million set. We have also included the ranking results for one of the most popular collections, Bored Apes Yacht Club [13], in Figure 6.

As a reward to readers who were patient enough to read the paper all the way to the end of this section, we are happy to reveal the *rarest* NFT from all data that we analyzed. The result is simple and yet remarkably genius—lo and behold—Green Square! See Figure 7.

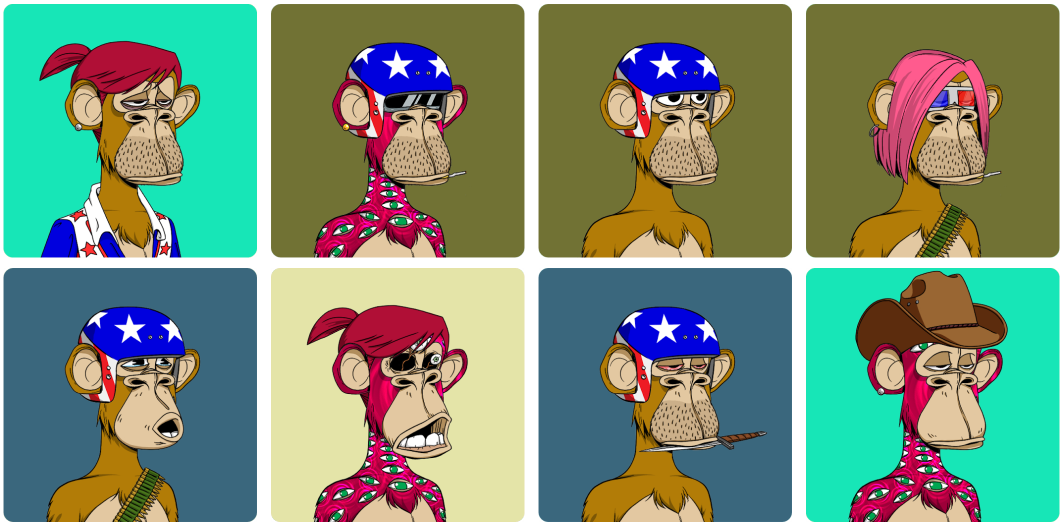

Figure 6: Rarest Eight Bored Apes by PIXELSCORE.

Figure 7: Green Square—the rarest NFT in our data set.

5 Conclusion

The PIXELSCORE algorithm introduced in this paper is the first *global* rarity estimation method in the NFT space. Unlike existing methods, it allows consistent comparison of the rarities of NFTs from different collections and thus ranks the entire NFT space with *one objective method*. Moreover, as we showed in Sections 3 and 4, the algorithm is scalable and can be executed on all existing NFTs with a reasonable amount of computation. We believe that our algorithm offers significant value to NFT creators, collectors and traders by providing consistent rarity benchmarks. We hope it will greatly benefit NFT buyers and holders by allowing them to objectively evaluate the rarity of their portfolio. Finally, PIXELSCORE is able to dynamically update and adjust to the shifts in distribution caused by the exponentially growing number of newly minted NFTs.

Disclaimer: Rarity scores are not the only factor determining the price and desirability of an NFT or a collection. Other factors, such as trends, community, personal preference, and business cycle, should be considered.

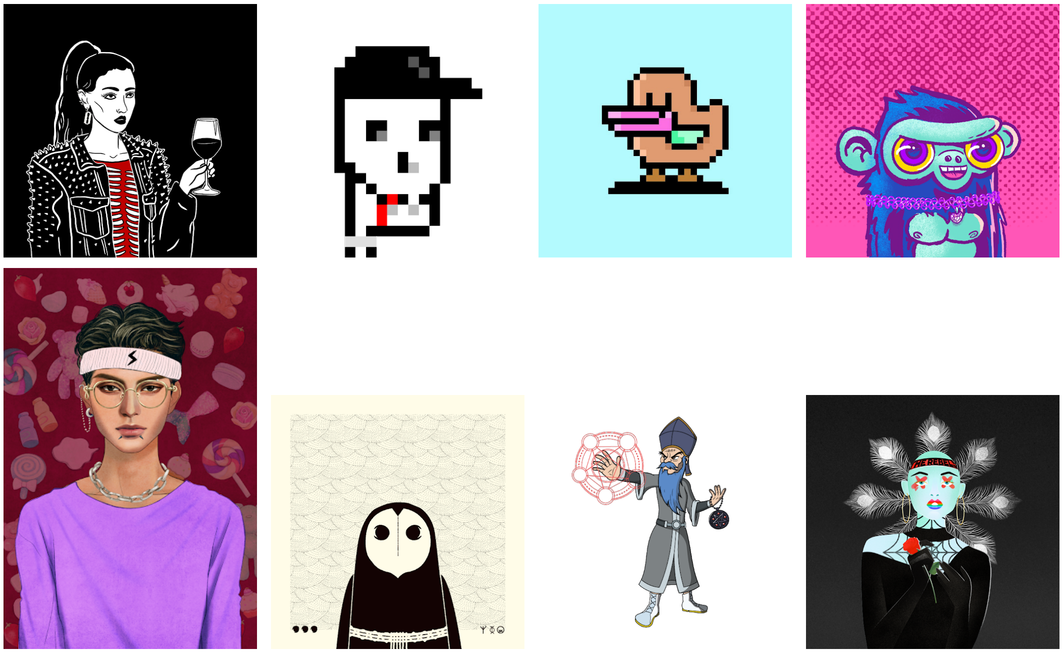

Figure 8: Random selection of the least rare tokens of the 10 million NFTs sampled.

Acknowledgments

Colleagues

References

[1] Ranking rarity: Understanding rarity calculation methods. https://raritytools.medium.com/ranking-rarity-understanding-rarity-calculation-methods-86ceaeb9b98c.

[2] Explaining nft rarity score models. https://www.publish0x.com/navid-ladani/explaining-nft-rarity-score-models-xmmorpr.

[3] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

[4] Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index.

[5] Qazal. The ultimate guide to nftgo's new rarity model. https://nftgo.medium.com/the-ultimate-guide-to-nftgos-new-rarity-model-3f2265dd0e23.

[6] Cryptopunks. https://opensea.io/collection/cryptopunks.

[7] Yee Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Advances in neural information processing systems*, 19, 2006.

[8] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[9] J Kahane. Propriétés locales des fonctions à séries de fourier aléatoires. *Studia Mathematica*, 19(1):1–25, 1960.

[10] Convolutional neural network. https://en.wikipedia.org/wiki/Convolutional_neural_network.

[11] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *arXiv preprint arXiv:1203.6402*, 2012.

[12] Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Fast and sample near-optimal algorithms for learning multidimensional histograms. In *Conference On Learning Theory*, pages 819–842. PMLR, 2018.

[13] Bored ape yacht club. https://opensea.io/collection/boredapeyachtclub.