Non-Rarity Metrics for Non-Fungible Tokens

Carol Alexander[a,c] Xi Chen[b]

Abstract

Within the new asset class of non-fungible tokens, personal profile picture (PFP) collections are avatars-with-benefits. Like Pokémon cards, the value of a token should depend on its rarity, indeed some very rare tokens have sold for several millions of dollars. However, each Pokémon card carries symbols from which its rarity can be determined at-a-glance, whereas the rarity of a PFP token must be measured from the characteristics defined in the metadata of the collection when it is minted to a blockchain. Unfortunately, there is nothing yet in the public domain that measures rarity correctly, except for a few special collections that have been designed to have independent traits. Although numerous metrics are now used by the PFP industry, different metrics can give vastly different results. This obscures any fundamental relationship between rarity and price and leads to great inefficiency in the PFP market. Our invariance results clarify the state-of-the-art, allowing classification of the numerous supposedly-different metrics into just four distinct cases, each a special case of weighted power mean. Importantly, neither the NFTGo Jaccard distance nor the OpenRarity Shannon entropy differ from the most commonly used metrics, i.e. the Pythagorean means, so they are not new, as claimed. We derive tests for trait independence, showing that most of the ∼200 PFP collections analysed have dependent traits, so none of these metrics is mathematically correct. For traders in PFPs we present two novel visualization tools, with code, which identify how discordant different rarity rankings can be, depending on the marketplace used.

Keywords: Personal profile picture, Pythagorean mean, Weighted power mean, Trait normalization

JEL codes: C60, G10, O30, O35, O36

Acknowledgements: This research was funded by a grant from the DFINITY Foundation for developing rarity metrics for non-fungible tokens. We are grateful and very much indebted to Dr. Peter M. Williams for numerous delightful and enlightening discussions about the mathematical concepts used in this paper.

[a]Professor of Finance, Department of Accounting and Finance, Jubilee Building, University of Sussex Business School, Falmer, Brighton BN1 9SL. Tel: +44 (0)1273 873950; Email: c.alexander@sussex.ac.uk.

[b]Assistant professor, Department of Accounting and Finance, Jubilee Building, University of Sussex Business School, Falmer, Brighton BN1 9SL. Tel: +44 (0)1273 872703; Email: xi.chen@sussex.ac.uk.

[c]Corresponding author.

Non-Rarity Metrics for Non-Fungible Tokens

Abstract

Within the new asset class of non-fungible tokens, personal profile picture (PFP) collections are avatars-with-benefits. Like Pokémon cards, the value of a token should depend on its rarity, indeed some very rare tokens have sold for several millions of dollars. However, each Pokémon card carries symbols from which its rarity can be determined at-a-glance, whereas the rarity of a PFP token must be measured from the characteristics defined in the metadata of the collection when it is minted to a blockchain. Unfortunately, there is nothing yet in the public domain that measures rarity correctly, except for a few special collections that have been designed to have independent traits. Although numerous metrics are now used by the PFP industry, different metrics can give vastly different results. This obscures any fundamental relationship between rarity and price and leads to great inefficiency in the PFP market. Our invariance results clarify the state-of-the-art, allowing classification of the numerous supposedly-different metrics into just four distinct cases, each a special case of weighted power mean. Importantly, neither the NFTGo Jaccard distance nor the OpenRarity Shannon entropy differ from the most commonly used metrics, i.e. the Pythagorean means, so they are not new, as claimed. We derive tests for trait independence, showing that most of the ∼200 PFP collections analysed have dependent traits, so none of these metrics is mathematically correct. For traders in PFPs we present two novel visualization tools, with code, which identify how discordant different rarity rankings can be, depending on the marketplace used.

Keywords: Personal profile picture, Pythagorean mean, Weighted power mean, Trait normalization

JEL codes: C60, G10, O30, O35, O36

1 Introduction

A non-fungible token (NFT) is an immutable certificate of ownership of a unique asset – typically, but not always, of a digital asset. For example, real works of art may be too costly to purchase for small investors, but 'fractionalization' divides the art work into smaller pieces which cost less, and each unique piece can be owned via an NFT. However, most NFTs refer to ownership of digital files such as photos, videos, audio recordings and pixel maps – and then one refers to both the digital asset and the certificate of ownership as 'the NFT'.[1] Such NFTs are commonly generated as a large collection of similar but non-identical tokens, the only limit on the total quantity supplied being that each must be unique.

There are numerous different types of digital-asset NFTs and within these there is a category called generative art NFTs (Driotcour, 2021; Oh et al., 2022). We focus on a particular type of generative art NFT called a profile picture (PFP) collection. Two famous examples of PFP collections are CryptoPunks and Bored Ape Yacht Club (BAYC). PFP collections are created by randomly generating a very large number of tokens, each representing a unique image derived from a set of original designs. They first set a fixed number of traits like background, mouth, eyes, ears, hat and so on. Each trait has a fixed number of attributes, e.g. background could be orange, blue, green, red etc. Then images are generated in layers to ensure that all the attributes sit correctly in the final image, and after a few trial runs a finite number (typically 10,000 or less) of different computer-generated avatars are minted as NFTs, either all-at-once or in stages, onto a blockchain. Typically the number of tokens is fixed at the start of the project. This is one of the features that distinguishes PFPs from other generative art NFTs such as Cryptokitties, which can breed and are therefore not in fixed supply – indeed by now there are well over 2 million unique CryptoKitties.

The global NFT market cap is currently 6.64 billion USD.[2] Well over 700 NFT generative collections are actively traded, with floor prices listed here, and the vast majority of these collections are PFPs.[3] But the tokens that are perceived as especially rare trade well above these floor prices. For instance, in August 2022 a particular Bored Ape with solid gold fur and several other rare attributes sold for $1.5 million. Prices of some CryptoPunks seem even more ridiculous, especially given that they were originally distributed free in June 2017, and the highest price ever paid (for Punk #5822) was over $23 million. These prices may seem crazy, and indeed could well be the result of wash trading, which is very often observed in NFT markets (Bonifazi et al., 2023; Serneels, 2023; Tariq and

[1] NFT certificates are smart contracts, and the vast majority are on the Ethereum chain using the Ethereum Request for Comments (ERC) protocols. The first ERC protocol designed specifically for NFTs was ERC-721, later followed by ERC-1155.

[2] For latest figures see CoinCodex.

[3] The floor price is the last traded price of any NFT in that collection, regardless of rarity. Blockchain data providers monitor these prices as indicators of general trends in the PFP market.

Sifat, 2022; Wachter et al., 2022; Mukhopadhyay and Ghosh, 2021). Nevertheless, tokens that are perceived to be especially uncommon can command very high prices, and such tokens have now become the digital profile of choice. The musician Madonna once paid over half a million US dollars for a Bored Ape.[4]

Ownership of a PFP usually provides membership to an exclusive community in which a particularly rare avatar conveys status whilst also maintaining anonymity.[5] Even if there is speculation by some that metaverse developments will not proceed as planned (Vidal-Tomás, 2023) the currently-perceived value of PFP tokens is directly linked to the *prospect* of status and power if or when the PFP eventually becomes a 3D avatar in the metaverse. The more exclusive the collection, and the rarer your token within the collection, the greater your status within the members club, and in the virtual reality becoming linked to this club.

Both scarcity and rarity are built into the original design of a PFP collection. For example, the oldest NFT collection CryptoPunks could have minted millions of different tokens instead of merely 10,000; with 6039 males, 3840 females, 88 Zombies, 24 Apes and 9 Aliens. Further attributes refer to background, skin color, head-wear, teeth, hairstyles, hair color, and so forth. Hence, for example, the combination of zombie and cigarette attributes is very rare whereas male, pipe and black hair are fairly commonly found together. The role of scarcity in price formation is a fundamental tenet of economic theory, and is by now very well understood. However, previous to this paper there has been no academic research on how to measure rarity in PFP collections correctly – and in the absence of this research it is not possible to even begin to analyse the relationship between rarity and price. Cho et al. (2023) demonstrate that there is very week correlation between price and rarity in both CryptoPunks and BAYC collections, although the relationship is stronger when analysing one particular trait, if it is rare. This could be due, at least in part, to wash trading, which is prevalent in many collections as previously noted. But it is also driven by the ecosystem-wide confusion surrounding the correct metric for measuring rarity. Indeed, the results of Cho et al. (2023) indicate that a better and universally adopted measure for rarity could greatly improve PFP market efficiency.[6]

[4]See Bloomberg, 25 March 2022.

[5]NFT collections usually provide owners with a club having several member-only benefits. For instance, owning certain Apes also gives access to the Yacht Club, which provides private parties, permission for restaurant franchises and access to merchandise collections. Yuga labs has led the way in this development, by launching many different Ape collections, by acquiring the Meebits and CryptoPunks from Larva Labs, and now by promoting interoperability between the various micro-metaverses so that one avatar can be used in all. Yuga Labs even have a token called ApeCoin which can now be used a a currency in their 'Otherside' metaverse, for example to purchase land via NFT's called 'Otherdeeds', and Yuga lab's PFPs are being converted to 3D versions that are designed to be compatible with all digital world platforms, not only Yuga Labs own game Otherside.

[6]A few papers on NFTs make references to rarity, but these are tangential. For instance, Schaar and Kampakis (2022) conclude that the rarity of a PFP is an important price determinant alongside other factors such as ETH volatility; Kong and Lin (2021) argue that the value of a specific token should

Of all the various NFTs designed for the metaverse – for fashion, land purchase, gaming accessories and so on – PFPs are particularly easy to analyse; theoretically because each token is utterly unique (because it is a digital identity) and empirically because hundreds of different collections have been minted and enough are actively traded to have data on which to attempt to establish an empirical relationship between rarity and price. However, this line of research has had little success so far, as demonstrated by Cho et al. (2023) among others. We maintain that developing a unique, universally accepted rarity metric is the most fundamentally important research to undertake on NFT markets now, and that many other challenges associated with data analytics on NFT transactions will become easier once the industry knows how to measure rarity properly.

PFPs are traded on relatively new online platforms (like Opensea, Magic Eden, Raribles and many others) that provide little or no information about the way they measure rarity. And even if there is a description of the method employed it is based on some *ad hoc* mathematical model that has no theoretical foundation whatsoever. Indeed, before this paper – which aims to offer resolutions to this problem – one could not tell whether any given PFP is relatively rare or common, because different platforms yield completely different results. For example, Ape token #2794 depicted in Figure 1 is ranked as the second rarest according to some platforms, as rank number 360 according to other platforms, and some others even rank it as 5040 which is relatively common.

Figure 1: **Bored Ape #2974.**

This Ape is ranked as number 2 (the second rarest) according to some platforms, as rank number 360 according to other platforms, and some others even rank it as 5040.

Our paper presents the first academically rigorous analysis of the rarity metrics currently employed by NFT market places. We shall classify each (purportedly-different) metric as a particular type of weighted power mean. As well as the much-needed standardization of terminology, we present theoretical results on invariance of rarity scores and ranks which allow the classification of *all* metrics in the pubic domain – including those supposedly-new approaches like NFTGo's Jaccard distance and OpenRarity's

depend on the preferences of the particular investor; and Lee (2022) attempts to infer the desirability of a token from its price, on the assumption that prices accurately reflect the rarity of the token.

Shannon entropy – into just four distinct types. Then we show that only one of these can ever be a mathematically correct measure of "rarity", according to the Oxford English Dictionary definition of: "*... a thing that is unusual and is therefore often valuable or interesting.*" A condition to apply this metric is that trait distributions are independent. Our statistical tests reveal that less than 20 of the ∼200 collections analysed here have independent traits. Finally we introduce new visualization tools for characterising the potential for highly-discordant ranking to result from applying the four fundamentally-different metrics – both for a particular token and in the PFP collection as a whole.

In the following: Section 2 provides an overview of the NFT ecosystem and surveys the rarity metrics that are currently in the public domain; Section 3 presents our mathematical results, which are then applied to data from some well-known PFP collections; Section 4 explains the importance of independent traits for current rarity metrics used, and applies statistical tests for trait independence, then we introduce some new visualization tools that allow users to analyse collections as a whole as well as individual tokens within a collection; and Section 5 concludes. The Appendix uses the data set provided by Nadini et al. (2021) to illustrate some of the novel descriptive aids that we introduce, for almost 200 different PFP collections.

2 The Market for Non-Fungible Tokens

PFP collections are traded in numerous marketplaces, most of which rank every token listed in a collection by a single number which purports to represent the rarity of the attributes that the token possesses. Unfortunately, descriptions of the various ranking methods currently in the public domain are mired in confusion because they are poorly presented. Sometimes terms are used without any definition at all, and when terms *are* defined the terminology often differs between ranking sites. For some terms, their use on one site means something very different to their use elsewhere.[7] Given the market cap of NFTs is now almost 7 billion USD such confusion is really quite astonishing.

So we begin by clarifying the terminology used in this paper, and some fundamental blockchain concepts too, so as not to lose readership. The term 'token' applies because an NFT is a special type of 'smart contract', i.e. a self-executing code which produces actions that are stored as transactions on a public blockchain. The term 'non-fungible' applies because, unlike other types of crypto assets, not all NFTs are created equal. An NFT is also 'immutable' because the consensus protocol of a public blockchain is specifically designed so that transactions cannot be changed. The first blockchain record of an NFT is a process called 'minting'. Once minted onto a particular blockchain the NFT remains on the same chain. After minting, further smart contract transactions can be recorded on the blockchain which track the history of any ownership transfers of the NFT.

[7] For example, 'statistical rarity' can mean a variant of the product or the sum or the harmonic mean of the attribute frequencies – compare HowRare.is, Nansen.ai God Mode and Rarity.tools for instance.

Investor attention on the NFT ecosystem has been expanding continually (Mukhopadhyay and Ghosh, 2021). Even after the global fall in crypto markets in 2022, major players were not deterred. For example, shortly after the collapse of the FTX crypto exchange in November 2022 Nike launched its own digital sportswear NFT platform; and on 15 November a pair of virtual Steve Jobs' sandals sold for over $218,750 (Katte, 2022). The numerous marketplaces for creating and trading NFT collections are especially popular among the millennial generation. New collections are minted daily and many of them use a business model which endows owners with various opportunities, such as networking events in real and virtual worlds. Both Davidsson et al. (2020) and Chandra (2022) identify other novel forms of business offered by NFTs, and discuss their potential for disrupting established models of entrepreneurship. Chalmers et al. (2022) argue that companies like Yuga Labs and Larva Labs that generate PFP collections are successful because they define algorithmic, social and goal coordination mechanisms for user activities.

NFTs are traded using a peer-to-peer transaction that is recorded on a blockchain, but most of these transactions are made and settled off-chain on an auction marketplace not unlike eBay. The first NFT marketplace, OpenSea, was founded in 2017. In January 2022 it was valued at $13.3 billion for its Series C funding round. Following OpenSea, the Nifty Gateway and Super Rare marketplaces were launched in 2018 and the former was acquired by Gemini in November 2019. During 2021 the largest centralized and decentralised crypto exchanges such as Coinbase, Binance, FTX and Uniswap also began to serve the NFT market. In 2022 several more marketplaces were launched, including Magic Eden, Looksrare, X2Y2, Sudoswap, and Blur and specialised new blockchains like Flow were developed specifically for carrying NFTs.[8] With the establishment of so many NFT marketplaces, buyers flooded in, and by January 2022 the monthly trading volume of NFTs reached a record high of 5.59 billion USD.[9] By now, PFPs dominate the NFT marketplace both in terms of trading volume and consumer interest, since some are fetching such enormously high prices. But investors quickly realised that it was difficult to model the fair price for PFPs, and consequently numerous analytics platforms sprung up extremely rapidly, most purporting to offer new and better products. These are very popular on social media.[10] By now at least twenty-five 'aggregator' platforms offer tracking services on PFP trades across different marketplaces and/or provide rarity ranks and other analytics on individual tokens from hundreds of different PFP collections.

Thus, PFPs rapidly became the dominant asset in the NFT ecosystem. But the market for PFPs is highly inefficient, even for the major collections such as BAYC and

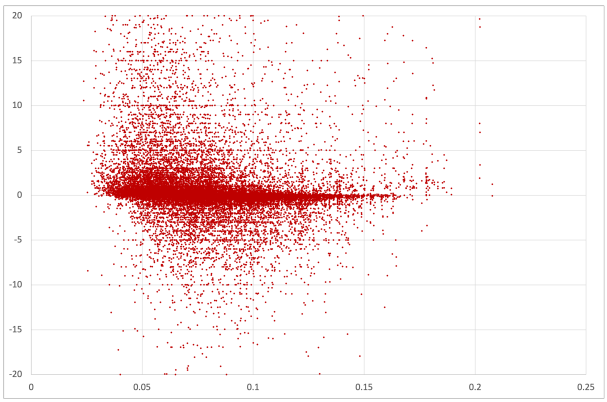
[8] Magic Eden raised $1.6 billion in its Series B funding round!

[9] See NFT Marketplace Monthly Volume.

[10] For instance, just in terms of Twitter followers at the time of writing: Rarity sniper has 443,000; Nansen.io has 228,000; Blur has 235,600; Trait sniper has 200,700; Rarity.tools has 89,600; Curio has 6,012; and Flow has 2,472.

CryptoPunks. To demonstrate this, Figure 2 displays the price-rarity relationship for BAYC. Each point is a token trade during the period 1 May 2021 and 30 April 2022, which corresponds to the height of the PFP market in general. The vertical axis shows the deviation of the traded price from the median price, this median being calculated from all trades of BAYC tokens on the same day as the trade. The unit of measurement is the native token for the Ethereum chain ether (ETH).[11] The horizontal axis is the geometric rarity score described later in this paper but we could have used any other rarity metric in the public domain to make the same point. We have truncated some extreme outliers to focus better on the vast majority of trades. Figure 2 exhibits no clear relationship between traded price and rarity score at all. If such a relationship existed then low rarity scores, which correspond to the trades of more rare tokens, would command above average prices. But this is not apparent in Figure 2 from visual inspection.[12] So despite its rapid expansion the PFP market remains immature and inefficient, as already shown by Cho et al. (2023) and demonstrated for NFTs in general by Ante (2022). This could contribute to wild speculations around prices, making them extremely difficult to predict, which harms all NFTs as a new tool of value creation. The confusion between rarity metrics obviously creates search frictions and information asymmetry for investors, and this also leads to poor market efficiency (Hou and Moskowitz, 2005; DeGennaro and Robotti, 2007; Wilson, 2012).

Figure 2: **Median Deviation Prices of BAYC.**



Each point represents a trade on a BAYC token. The horizontal axis is the geometric rarity score and vertical axis is the deviation of ETH price from the median price of all trades of BAYC tokens on the same day. Based on OpenSea trade data between 1 May 2021 and 30 April 2022.

[11] Not measured in USD, because the extreme fluctuations of the ETH/USD rate would contaminate results. Also, because of the dominance of Ethereum, ETH has become the numéraire for decentralised finance.

[12] The majority of very rare tokens do transact above the median price, but many tokens with low rarity scores were also traded at prices well below the median, and moving to the right along the horizontal scale one sees absolutely no relationship between price and rarity score, even for tokens with relatively low scores.

Figure 3: **Ten Most Rare Apes in the BAYC Collection**
The ranking of these Apes is highly dependent on the rarity metric used. There is very little similarity between the top ten PFPs appearing in each row. The top row uses the 'trait rarity', the second row uses 'rarity score', the third row uses 'statistical rarity' and the fourth row uses 'average rarity'.
Source: Rarity.tools BAYC Collection

Rarity.Tools survey a few of the ranking methods used by the industry at the time.[13] The same company also provides a ranking site with a user interface having an assortment of different ranking metrics,[14] and on playing with this it quickly becomes clear that different metrics produce very different rankings. For example, Figure 3 shows that the 'top ten' most rare Apes in the BAYC collection very much depend on the rarity metric selected. The top row uses the 'trait rarity', defined as the smallest of all its attribute frequencies; the second row uses the 'rarity score' defined as the sum of the reciprocals of the PFP's attribute frequencies;[15] the third row uses 'statistical rarity' which is the product of the attribute frequencies; and the fourth row uses the arithmetic average of the attribute frequencies, which Rarity.tools calls the 'average rarity'.

The two equal top-ranked Apes in the first row have only one attribute in common, their background color – but the background trait has no association with rarity because all background colors are more-or-less equally likely, as we shall see later on. But what is most striking about Figure 3 is that the top ten PFPs are very different in each row. And, the Ape #2974 displayed in Figure 1 only appears once in Figure 3 – it is ranked fifth based on the 'rarity score' metric.

This paper analyses almost 200 different PFP collections, but our main empirical study

[13] See also the HowRare.is FAQ, the Sensorium blog Best NFT Rarity Tools Lapuschin (2022) and many other copies of the Rarity.tools blog that have appeared elsewhere.

[14] Plus a toggle for switching 'trait normalization' on or off, plus some ad-hoc weightings that can be customized for users that value specific traits more than others.

[15] Also see this Reddit discussion.

focuses on the three most well-known and actively-traded PFP collections.[16] The oldest and most famous collection is CryptoPunks, which was minted in 2017 and airdropped free to the Reddit community at that time. These early entrepreneurs may have made huge investment returns, depending on their luck of the draw in the airdrop. The average sale price per token remains fairly constant at 200 ETH (about $370,000 at the time of writing) and many sell for even higher prices. Total sales to date exceed $2 billion. All Punks have at least three attributes: background, gender, skin color. There are eight Punks with only three attributes and one with ten attributes but none have ever been sold. Most CryptoPunks have five or six attributes – details may be found here. Another very well-known PFP is Bored Ape Yacht Club (BAYC) which is a collection of 10,000 unique bored apes images. Minted in April 2021, it sold out within a week. At that time each image, however rare, had a mint price of 0.08 ETH. As with Punks, some Apes are very rare and others are quite common, and the highest price ever paid (for Ape #8817) was $3.4 million. BAYC tokens have seven different traits: the color of the background for the head-shot, clothes, earring, eyes, fur, hat and mouth. W also analyse Doodles, a collection of 10,000 tokens with six traits, minted in October 2021. These hand-drawn Doodles are categorized as skellys, cats, aliens, apes and mascots. The collection also includes dozens of rare heads, costumes, and colorways of the lead artist's palette.

3 The Mathematics of Rarity

In the following, subsection 3.1 defines the terminology and mathematical notation required for understanding rarity. It also shows how all PFP rarity metrics are equivalent to a special case of *generalised mean* – in fact, all but one are equivalent one of the three well-known *Pythagorean means*. A toy example illustrates the differences between them. Subsection 3.2 introduces *rarity scores* and proves some general results about scores, which allow us to further prove that both the NFTGo (supposedly combinatorial) and the OpenRarity (supposedly information-theoretic) models are not actually new as claimed by the developers, because each is a different Pythagorean mean in disguise. Subsection 3.3 introduces the *weighted power mean* as a universal metric which encompasses all variations of the rarity metrics in use today (including the *ad hoc* weightings for different traits that are allowed by Rarity.tools) and then explains how weights may be chosen in the power mean to accommodate *trait normalization* in a mathematically coherent manner.

[16] See Cryptoslam sales volume dashboard.

3.1 A Framework for Measuring Rarity

We characterize a PFP collection firstly by its *traits* of which there are a finite number of types. Then, by assigning the attribute "none" when the trait is absent from the image,[17] an individual token can be defined by a unique set of *attributes* and each attribute refers to one specific value for every trait. For example, a collection may be defined as possessing five traits: background, gender, hair, eyes and clothes. Each trait has a fixed number of attributes,[18] for instance there could be ten different background colors, three different gender types, four hair types, five eye types and eight clothes types. And while there may be many PFPs that are on a blue background, male with short hair, at most one PFP can have the specific attributes {background = blue, gender = male, hair = short, eyes = brown, clothes = pirate}. The maximum number of unique attribute strings in this toy example is $10 \times 3 \times 4 \times 5 \times 8 = 4,800$. Since each PFP must be unique, there can be no more than 4,800 PFPs but there could be less than this number. Typical collections mint only a very small fraction of possible tokens. For instance, there might be around two million potential PFPs in the collection but only 10,000 are minted.

In its basic form, *i.e.* without apply different weights to different traits (which we deal with later in Section 3.3) every rarity metric in the public domain is either the rarest trait (the minimum of all attribute frequencies) or a monotonic transformation of one of the well-known Pythagorean means, *i.e.* the arithmetic, geometric or harmonic mean of the attribute frequencies. For instance, Rarity.tool's 'statistical rarity' metric is the product of the attribute frequencies – they do not take the n^{th} root of this product to obtain the geometric mean. Likewise their 'rarity score' metric is the *reciprocal* of the harmonic mean, divided by n and not the harmonic mean itself. Before considering the effect of such monotonic transformations on rarity metrics, Table 1 illustrates the effect of using different metrics using a toy example of a PFP collection with just three traits (gender, hair and eyes) and 60 tokens. The upper section shows the number of tokens having each attribute and the frequency of each trait, and the lower section lists the attributes of just three of the 60 tokens in the collection, followed by the frequency of their rarest trait, the harmonic mean, the geometric mean, and the average of the attribute frequencies.

The ranking of the three tokens (ID1, ID2, ID3) depends on the method used. The first token with attributes {gender = neither, hair = none, eyes = green} is the rarest according to the geometric and arithmetic means; the second token with attributes {gender =

[17] For instance, many Bored Apes have no earring. Some collections also allow more than one value for a trait, e.g. a token might have zero, one or two earrings. In this case the earring trait must be replicated to earring 1 and earring 2, so that those with no earring have the attribute "none" for both, those with one earring have a value such as "cross" or "stud" for earring 1 and "none" for earring 2, and those with two earrings might have, say, "cross" for earring 1 and "stud" for earring 2.

[18] The attributes of every token are stored as metadata with the original smart contract used for minting the PFP collection. For most collections these data are publicly available from marketplaces. We have used the OpenSea API to obtain trait data for CryptoPunks, BAYC and Doodles collections to analyse in this paper.

Table 1: **Illustration of Existing Rarity Metrics**
The upper section shows the number of tokens having each attribute. For each trait (gender, hair and eyes) the row sum in column "Total" is 60, because there are 60 tokens in the collection. Then it shows the attribute frequencies (in red) calculated as the number having that attribute divided by 60. For example, 24 of them are male, *i.e.* 24/60 = 40% of them. The lower section of the table shows the attributes of just three tokens, followed by the four different rarity measures derives using the metric defined in the text (in blue). Bold type denotes the minimum value in each column, i.e. the "rarest" token according to that particular metric.

Gender	**Male**	**Female**	**Neither**			**Total**	
Number	24	21	15			60	
Frequency	0.4	0.35	0.25			1	
Hair	**Short**	**Long**	**Mohican**	**None**		**Total**	
Number	24	18	9	9		60	
Frequency	0.4	0.3	0.15	0.15		1	
Eyes	**Brown**	**Blue**	**Green**	**Bloody**	**X-ray**	**Total**	
Number	18	15	12	9	6	60	
Frequency	0.3	0.25	0.2	0.15	0.1	1	
ID	**Gender**	**Hair**	**Eyes**	Minimum	Harmonic	Geometric	Arithmetic
1	Neither	None	Green	0.150	0.191	**0.196**	**0.200**
2	Female	Mohican	Bloody	0.150	**0.150**	0.199	0.217
3	Male	Short	X-ray	**0.100**	0.200	0.252	0.300
Rarest ID				ID 3	ID 2	ID 1	ID 1

female, hair = mohican, eyes = bloody} is the rarest according to the harmonic mean; and the third token with attributes {gender = male, hair = short, eyes = x-ray} is the rarest according to the minimum (*i.e.* the rarest trait).

Having fixed ideas with this simple example, we now introduce some notion. Suppose a PFP collection has n traits. For instance, the BAYC has traits: *background, clothes, earring, eyes, fur, hat, mouth* so $n = 7$ for this collection. In general, we can label traits using the random variables $X_1, \ldots, X_n$. Each trait has a predefined number of attributes so let trait X_i have θ_i different values, including the possible value 'none'. Thus, for $i = 1, \ldots, n$ each random variable X_i can take values x_{ij} for $j = 1, \ldots, \theta_i$,[19] and the total number of attributes in the collection is $\theta_1 + \theta_2 + \ldots + \theta_n$.

Next, suppose there are m tokens in the collection, with IDs $A^1, \ldots, A^m$. For each $k = 1, \ldots, m$ we can represent the token with ID A^k by its unique set of attributes. This is because its metadata, when minted on the blockchain, includes an assignment from traits to specific attributes which is *unique*, that is no other token has the same set of attributes. In other words, there is a bijective map between the token ID and it's set of attributes:

$$A^k \longleftrightarrow \left\{X_1 = x_1^k, \ldots, X_n = x_n^k\right\}. \tag{1}$$

For instance, we may have a BAYC token with {background = blue, clothes = black suit, earring = cross, fur = brown, hat = none, mouth = dagger}. Note that 'none' here means that no hat is present, and 'none' is included as a possible value for the hat trait.

[19] For example X_1 could take 9 different values $x_{11}, \ldots x_{19}$ and X_2 could take 6 different values $x_{21}, \ldots x_{26}$, and so on.

The *count* of a particular value x_{ij} for trait X_i is the number of tokens having an attribute with that value. For example, if 850 bored apes have a hoop earring, the count of the hoop earring attribute is 850.[20] In mathematical terms, the count of x_{ij} is the number of tokens for which $x_i^k = x_{ij}$, which may be written as:

$$m_{ij} = |k \ni x_i^k = x_{ij}| = \sum_{k=1}^{m} \mathbb{1}_{x_i^k = x_{ij}}, \tag{2}$$

where the 'cardinality' $|B|$ of a set B is the number of elements in it, and the indicator function $\mathbb{1}_{a=b}$ takes the value one when a equals b and zero otherwise. Since every token must have exactly one attribute per trait the sum of the counts is the same for every trait, that is:

$$\sum_{j=1}^{\theta_i} m_{ij} = m, \qquad \text{for } i = 1, \ldots, n. \tag{3}$$

The *frequency* p_{ij} is calculated by dividing each trait count by m, the total number of tokens in the collection, that is:

$$p_{ij} = P\left(X_i = x_{ij}\right) = \frac{m_{ij}}{m}, \tag{4}$$

where P denotes 'probability' and by (3) and (4) we must have:

$$\sum_{j=1}^{\theta_i} p_{ij} = 1, \qquad \text{for } i = 1, \ldots, n.$$

We also assume that the number of tokens m is fixed, so that the counts and frequencies (4) can be computed as soon as the entire collection has been minted. For example, if 2,000 tokens have a blue background and there are 10,000 tokens in total, the blue background attribute has count 2,000 and frequency 20%.

Now consider a specific token A^k, which may be characterized by its unique set of attributes $\left\{x_1^k, \ldots, x_n^k\right\}$. Having calculated all the counts and frequencies using (2) and (4), we may now map the attributes of token A^k to their corresponding counts and frequencies. This gives an alternative representation of the token: instead of the set of attributes (1) we map them into a set of corresponding attribute counts $\left\{m_1^k, \ldots, m_n^k\right\}$ and frequencies $\left\{p_1^k, \ldots, p_n^k\right\}$, where:

$$m_i^k = m_{ij} | x_{ij} = x_i^k \quad \text{and} \quad p_i^k = \frac{m_i^k}{m}, \quad \text{for } i = 1, \ldots, n. \tag{5}$$

[20] We could also call this the *absolute frequency* instead of the count, but that is longer especially because then, the attribute frequency defined below would have to be termed the *relative frequency* of the attribute.

Unlike the unique set of attributes $\{x_1^k, \ldots, x_n^k\}$, the attribute count and frequency representations, $\{m_1^k, \ldots, m_n^k\}$ and $\{p_1^k, \ldots, p_n^k\}$ need not be a unique representation of the token. Two tokens may share identical attribute counts (and so also frequencies) even though the set of attributes cannot be the same.

3.2 Invariance under Monotonic Transformations

Armed with these definitions and notations we may now provide rigorous definitions of the metrics currently employed by the various rarity analytics platforms. We begin by formalizing the definition of the rarest trait, the most common trait and by recalling the definitions of the three Pythagorean means. Ranking sites present their metrics in different forms, but we can show that every metric in the public domain produces an identical rarity score to one of the metrics in Table 2 below.

Even the supposedly-novel rarity analytics platform NFTGo and the open-source code provided by OpenRarity are actually just employing a Pythagorean mean in disguise. Yet their terminology – *Jaccard distance* and *Shannon entropy* respectively – suggests each metric is new. Moreover, both NFTGo and OpenRarity make abundant comments that aim to impress upon NFT investors both the novelty and superiority of their own rarity metric.[21] However, this is not the case.

Table 2: **Pythagorean Mean Rarity Metrics**

For a given token A^k let us denote its rank by r_*^k with the subscript $*$ indicating which metric is applied. Well-known algebra proves that we always have:

$$r_{\text{min}}^k \leq r_{\text{harmonic}}^k \leq r_{\text{geometric}}^k \leq r_{\text{arithmetic}}^k \leq r_{\text{max}}^k \qquad k = 1, \ldots, n.$$

Metric	Definition	Mathematical Formulation
Minimum	Also called 'rarest trait' this is the smallest of a token's attribute frequencies.	$r_{\text{min}}^k = \min_i p_i^k$
Harmonic Mean	This is the reciprocal of the arithmetic average of the reciprocals of the attribute frequencies.	$r_{\text{harmonic}}^k = \dfrac{n}{\sum_{i=1}^n \left(p_i^k\right)^{-1}}$
Geometric Mean	This is the n^{th} root of the product of all the attribute frequencies.	$r_{\text{geometric}}^k = \prod_{i=1}^n \left(p_i^k\right)^{1/n}$
Arithmetic Mean	Also called 'average trait' is just the average of all the attribute frequencies.	$r_{\text{arithmetic}}^k = n^{-1} \sum_{i=1}^n p_i^k.$
Maximum	This is the largest of a token's attribute frequencies.	$r_{\text{max}}^k = \max_i p_i^k$

To prove this assertion we must first understand the effect of a strict monotonic transformation on a rarity measurement r. For example, the transformation $r \mapsto r^{\frac{1}{n}}$ is strictly increasing (as is its inverse $r \mapsto r^n$) because n (the number of traits) is positive.

[21] See NFTGo's medium article here and OpenRarity's claims here. Their code is open-source and the methodology is describe in some detail here

This implies that the final ranks produced by taking the product of trait frequencies must be identical to those obtained using the geometric mean. Similarly, the transformation $r \mapsto nr$ and its inverse $r \mapsto \frac{r}{n}$ are both strictly increasing, so the ranks produced by summing the trait frequencies are identical to those obtained from the arithmetic mean. In short, ranks are invariant under strictly increasing transformations.

We now prove that every rarity metric currently in the public domain is a strict monotonic transformation of one of the Pythagorean means.[22] Some are monotonic *increasing* transformations – for example, those metrics based on the product of the attribute frequencies that do not take the n^{th} root; and some are monotonic *decreasing* transformations – for example, the OpenRarity code produces a strictly decreasing transformation of the geometric mean, as we demonstrate using a few lines of algebra below. For another example, the 'rarity score' in Rarity.tools is a decreasing monotonic transformation of the harmonic mean. It is the sum of the reciprocals of the attribute frequencies. It does not take its reciprocal and then multiply by n, which would convert the sum of the reciprocals to the harmonic mean. And the transformation $r \mapsto nr^{-1}$ is strictly decreasing.

Any strictly increasing transformation affects the size but not the order of the rarity measures, so that ranks are invariant under these transformations and the smaller the rarity measure the more rare the token. However, any monotonic decreasing transformation not only affects the size, but it also reverses the order of the rarity measurements, so the *greater* the rarity measure the more rare the token. Therefore, any marketplace that employs a metric that is a decreasing transformation of a Pythagorean mean must be *reversing* the rank order produced by their metric before publishing the rarity ranks on their sites.

Rarity Scores

We can avoid the complexity arising from all such transformations by converting the raw rarity measurement to a *rarity score*, which we shall do as follows.[23] Given any rarity metric r the resulting measurements r^k for all tokens A^k can be standardized to lie in the same range $[0, 1]$ by finding the minimum $r_{\min}$ and maximum $r_{\max}$ measurements for rarity metric r, over all the tokens in the collection. Then for each token A^k we map $r^k \mapsto \bar{r}^k$ where the image is the rarity score, calculated as:

$$\bar{r}^k = \frac{r^k - r_{\min}}{r_{\max} - r_{\min}}. \tag{6}$$

For example, suppose that the metric r is the geometric mean and, according to this metric, the collection has a most rare token with the minimum geometric mean value of

[22] Except the rarest trait $r^k_{\min}$, which is a limit of the generalised mean introduced in the next section, but this seldom used for ranking PFPs.

[23] Given the confusions and misunderstanding in the industry at present, it will come as no surprise that token analytics platforms employ the term 'score' in a variety of different ways.

0.1, and a least rare token with the geometric mean 0.5. Suppose another token A^k has rarity $r^k = 0.4$. Then the rarity score for this token, according to (6) is

$$\bar{r}^k = \frac{0.4 - 0.1}{0.5 - 0.1} = 0.75.$$

In the following, Remarks 1 and 2 follow trivially from the definition (6) of the rarity score and Remark 3 remains a conjecture:[24]

1. Given a rarity measure r, and real numbers a, b with $a \neq 0$, the linear map $r \mapsto ar+b$ is strictly increasing when $a > 0$ and strictly decreasing when $a < 0$. The rarity score (6) is unchanged by this map when $a > 0$. But when $a < 0$, then scores are reversed under this map so the most rare token has the highest score and the least rare has the lowest score;
2. Any non-linear, continuously differentiable, strictly increasing transformation $r \mapsto f(r)$ with $f'(r) > 0$ everywhere, changes the value of rarity scores but does not affect their ordering so ranks are unaffected by the transformation. However, a similar but strictly decreasing transformation $r \mapsto f(r)$ with $f'(r) < 0$ changes the rarity scores and *reverses* their ordering so ranks must be assigned from 1 (most rare and *highest* score) to m (least rare and *lowest* score);
3. We conjecture that when the transformation is convex, *i.e.* $f''(r) > 0$, it increases the skewness of the rarity score distribution; and when it is concave, *i.e.* $f''(r) < 0$, the transformation decreases the skewness of the rarity score distribution.

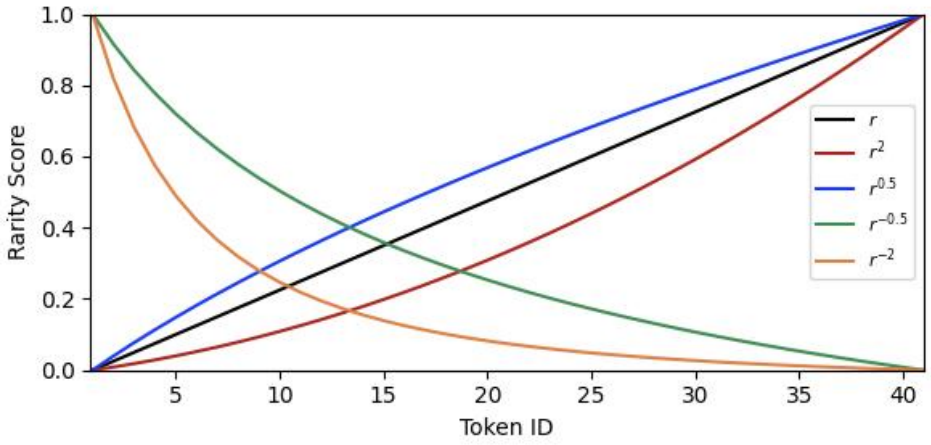
Recall that a positive (negative) skew occurs when the mass of the density leans to the left (right).[25] The skewness of the score density influences the tendency for scores to be unevenly distributed in the right and left tails, but this type of information is lost once scores are converted to ranks, which are uniformly distributed by definition.

To illustrate Remarks 1 and 2, Figure 4 depicts the rarity scores for 41 of the tokens in the toy example above. The token IDs are selected so that raw rarity measures r^k are evenly spaced between 0.1 and 0.5, and increasing with the ID number k, so that the scores for these tokens are a linear increasing function of the token ID, depicted by the black line. The other lines represent the scores for four simple functions of r, viz. $f_1(r) = r^2$ (convex increasing and depicted in red), $f_2(r) = r^{1/2}$ (concave increasing and depicted in blue), $f_3(r) = r^{-1/2}$ (concave decreasing and depicted in green), $f_4(r) = r^{-2}$ (convex decreasing and depicted in orange). The reversal of scores resulting from a decreasing transformation is clear.

[24] A theoretical proof would require knowledge of the moment generating function of a non-linear transformation of an arbitrary random variable, which is an unsolved statistical problem.

[25] Typically, but not always, positively skewed densities have mode > median > mean and the opposite ordering occurs for negatively skewed densities.

Figure 4: **Effect of Different Monotonic Transformations on Rarity Scores.** The rarity scores for 41 tokens in the example (vertical scale) versus token ID (horizontal scale). Tokens are selected so that their scores $\bar{r}^k$ are a linear function of token ID k (depicted by the black line). Then we apply four simple transformations of the raw measures r^k: $f_1(r) = r^2$ (convex increasing transformation, score depicted in red), $f_2(r) = r^{1/2}$ (concave increasing and score depicted in blue), $f_3(r) = r^{-1/2}$ (concave decreasing and depicted in green), $f_4(r) = r^{-2}$ (convex decreasing, orange).



Now consider Remark 3, which depends on the curvatures of the transformed rarity measures. To this end, we analyse the impacts of convex and concave transformations of a rarity metric r by assuming the base case, before the transformation, has a parametric rarity score distribution. Since scores have range $[0, 1]$, we suppose that $\bar{r}$ has a beta distribution which has density:

$$f(\bar{r}; \alpha, \beta) = c\, \bar{r}^{\,\alpha-1}(1 - \bar{r}^{\,\beta-1}), \tag{7}$$

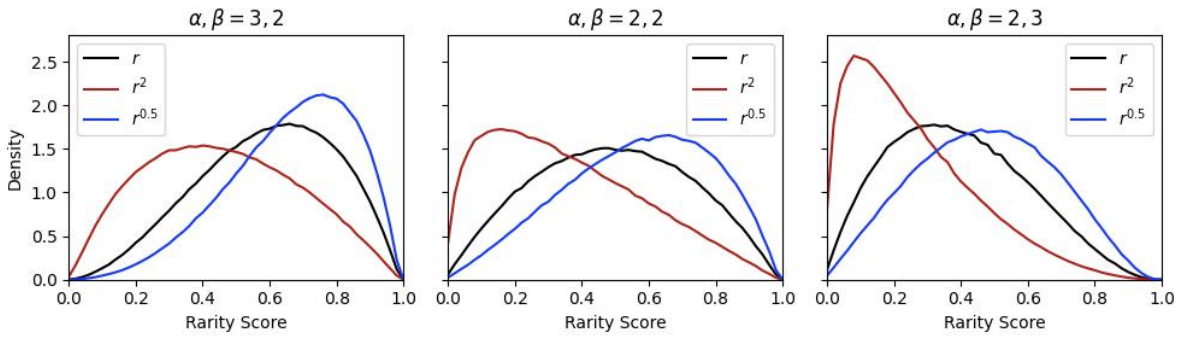
where the normalizing constant c is the reciprocal of the beta function $B(\alpha, \beta)$. The beta density can take a variety of shapes (even bath-tub) depending on the values of α and β. For instance, the simple case $\alpha = \beta = 2$ yields a parabola-like score density.

As previously, we suppose the minimum rarity measure is $r_{\min} = 0.1$ and the maximum is $r_{\max} = 0.5$ but instead of picking out 41 tokens with evenly-spaced measures r^k between 0.1 and 0.5, for $k = 1, \ldots, 41$, here we make the more realistic assumption that there are 10,000 tokens, as there are in most collections. To back-out the rarity measures r^k from the scores $\bar{r}^k$ we just invert the linear transformation (6). Then we apply two transformations to these r^k, one convex ($r \mapsto r^2$) and one concave ($r \mapsto r^{1/2}$), and generate the corresponding rarity scores. Figure 5 plots the score densities so obtained, with three different base case parameters (α, β) in (7) set to $(3, 2)$, $(2, 2)$ and $(2, 3)$ respectively.[26] Clearly, the convex transformation $r \mapsto r^2$ increases the skewness of the rarity score density, i.e. in each plot the red curves have modes lying to the left of the modes in the

[26]These densities are simulated, and so for the purpose of smoothing we actually sample 1 million rather than 10,000 tokens from the original score density. Even then, the curves are not entirely smooth, but they are sufficient to illustrate our points.

black curves; likewise, the concave transformation $r \mapsto r^{1/2}$ decreases the skewness of the rarity score density, i.e. the blue curves have modes lying to the right of the modes of the black curves in each plot.

Figure 5: **Effect of Rarity Transformations on the Score Distribution.**
Transformed rarity score densities when the base case follow a beta distribution with density (7), with (α, β) set to $(3, 2)$, $(2, 2)$ or $(2, 3)$ respectively , depicted in black in each graph. The rarity measures r are transformed using the convex function $r \mapsto r^2$ (dark red) and the concave function $r \mapsto r^{1/2}$ (dark blue).

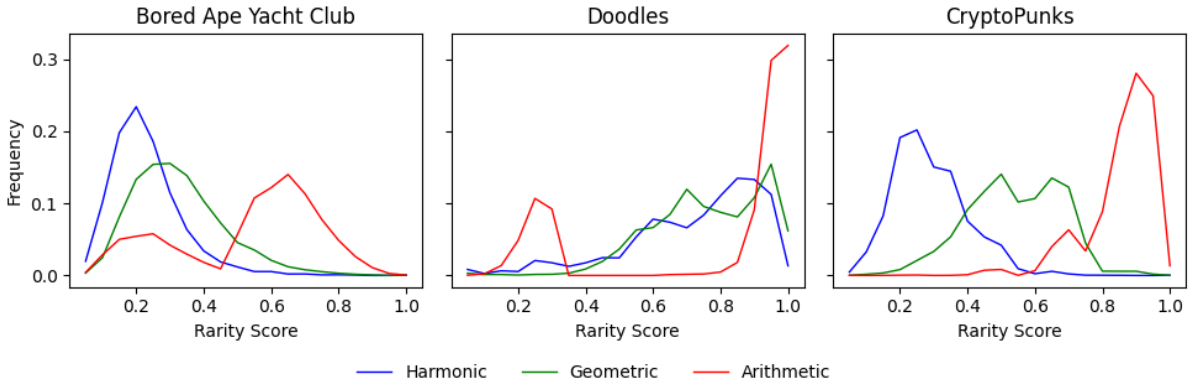


The rarity score transforms all metrics to the same standardized range of $[0, 1]$ without affecting the final rank order of tokens under different metrics. But in doing so it throws away valuable information, which makes an efficient pricing mechanism for PFPs even more difficult to achieve. The significant advantage of reporting rarity scores rather than ranks to the public, is that one may discern whether tokens with consecutive ranks are actually quite different, or very similar, in their rarity characteristics. By definition, rank distributions are always uniform, but score distributions carry much extra information on the relative rarity of tokens within a collection. The distribution of scores can often reveal that one can hardly distinguish between the rarity of various clusters of tokens, but this information is completely lost once scores are converted to ranks. Unfortunately, since only ranks and not scores are output to the public on the platforms' interfaces, all the insights that can be derived from the shapes of different rarity score densities are unavailable to the traders and investors of PFPs.

To reinforce our case that scores rather than (or at least as well as) ranks should be reported, we present an empirical study of our three classic collections viz. BAYC, Doodles and CryptoPunks. For each collection, Figure 6 compares the relative frequencies derived from the rarity scores of the 10,000 tokens in each collection, when scores are based on either the harmonic, or geometric or arithmetic rarity metrics. These frequencies are depicted in blue, green and red, respectively. First consider the BAYC collection (on the left). The harmonic mean score has a high, positive skew, the geometric mean also has positive skew and the arithmetic mean is bimodal, with negative skew. Now, the majority of platforms publish ranks based on a monotonic transformation of the harmonic mean, see Table 3. For instance, the 'rarity score' metric of Rarity.tools applies

the strictly decreasing convex transformation $r \mapsto nr^{-1}$ to the harmonic mean. The harmonic mean (depicted in blue) allocates a score greater than 0.75 to very few tokens. After a convex transformation, such as Rarity.tools, Remark 3 suggests that this positive skew is increased even further, so that virtually no tokens would have a score exceeding 0.75. Next consider the arithmetic mean frequency, depicted in red, where about 13% of tokens have rarity score greater than 0.75. The 'average trait' metric on Rarity.tools and the 'statistical rarity' advocated by Nansen.io, are both increasing linear transformations of the arithmetic mean and so, by Remark 1, their score frequencies are identical to the red line shown.[27] Similar remarks apply to the CryptoPunks score frequencies, displayed on the right. However, the score frequencies for Doodles are rather strange. Perhaps not all collections are designed using coherent algorithms.

Figure 6: **Rarity Scores from Harmonic, Geometric and Arithmetic Means**
The relative frequency plots of rarity scores derived from harmonic, geometric and arithmetic mean rarity metrics. These are computed using each collection's trait data which are downloaded from OpenSea.



NFTGo's Jaccard Distance

Next consider the methodology proposed by NFTGo.[28] We translate their 'Ultimate Guide' into our notation. Suppose two tokens A^k and A^l in the same collection having n traits have attributes $\{x_1^k, \ldots, x_n^k\}$ and $\{x_1^l, \ldots, x_n^l\}$ respectively. Then the proportion of attributes they have in common is called the *Jaccard similarity*, which may be written using the indicator function as

$$n^{-1} \sum_{i=1}^{n} \mathbb{1}_{x_i^k = x_i^l}.$$

The *Jaccard distance* between these two tokens is one minus the Jaccard similarity, written as:

$$JD(A^k, A^l) = 1 - n^{-1} \sum_{i=1}^{n} \mathbb{1}_{x_i^k = x_i^l}. \tag{8}$$

[27] By contrast, the NFTGo 'Jaccard Distance' metric is a decreasing linear transformation of the arithmetic mean, as we prove below. Hence, again by Remark 1, the NFTGo score density is a 'flipped' version of the red line.

[28] Full details are provided in NFTGo (2021).

NFTGo's rarity measure for token A^k is the average of its Jaccard distances from all the other tokens. That is:[29]

$$JD^k = m^{-1} \sum_{l=1}^{m} JD(A^k, A^l). \tag{9}$$

NFTGo then apply the equation (6) to obtain the final rarity score for token A^k, which we denote $\bar{r}^k_{\text{Jaccard}}$.[30] Combining (8) and (9) yields:

$$\begin{aligned} JD^k &= m^{-1} \sum_{l=1}^{m} \left(1 - n^{-1} \sum_{i=1}^{n} \mathbb{1}_{x_i^k = x_i^l} \right) \\ &= 1 - n^{-1} \sum_{i=1}^{n} \sum_{l=1}^{m} m^{-1} \mathbb{1}_{x_i^k = x_i^l} \\ &= 1 - n^{-1} \sum_{i=1}^{n} p_i^k = 1 - r^k_{\text{arithmetic}}. \end{aligned}$$

Hence, the NFTGo Jaccard distance metric is nothing more than a decreasing linear transformation of the arithmetic mean. Hence, from Remark 1, the Jaccard distance 'flips' the arithmetic mean rarity score density in the vertical axis. Put another way, now setting $r_{\text{arithmetic}} = r$ for short, the reversal in rarity scores under the two different metrics is clear from:

$$\bar{r}^k_{\text{Jaccard}} = \frac{JD^k - JD_{\min}}{JD_{\max} - JD_{\min}} = \frac{r_{\max} - r^k}{r_{\max} - r_{\min}} = 1 - \frac{r^k - r_{\min}}{r_{\max} - r_{\min}} = 1 - \bar{r}^k.$$

Therefore, NFTGo must apply ranks in the opposite direction to sites that employ the arithmetic mean, or any strictly increasing transformation thereof, like Rarity.tools and Nansen.io. This way, all three sites yield identical rankings for all tokens. That is, the NFTGo method is not new at all, as claimed. It is also mathematically incorrect, as we shall demonstrate in subsection 4.1 below.

OpenRarity's Shannon Entropy

Returning to the example of applying the transformations $r \mapsto \pm n \log_2(r)$ to metrics r, let us dwell a little more on the OpenRarity methodology. Their description contains just one formula,[31] which may be written in our previously-defined notation as:

$$IC^k = a \sum_{i=1}^{n} \log_2(p_i^k), \tag{10}$$

[29] NFTGo use $m-1$ rather than m in the denominator of (9) and exclude $k = l$ from the sum. But we prefer to use m here because there is no need to exclude $k = l$ since $JD(A^k, A^k) = 0$. Also, the division by m makes the equivalence between JD^k and the arithmetic mean (in Table 2) immediately obvious.

[30] The factor of 100 used by NFTGo makes no difference to the final rank and can therefore be ignored.

[31] See OpenRarity Methodology.

where a is a negative constant. For the purpose of ranking tokens it is simple to prove that the geometric mean in Table 2 is equivalent to (10), even though they produce different scores, which also have their order reversed. We only need to state the transformation between the two metrics, which is strictly decreasing and convex, as follows:

$$r \mapsto na \log_2 r \qquad \text{yields the map} \qquad \prod_{i=1}^{n} \left(p_i^k\right)^{1/n} \mapsto a \sum_{i=1}^{n} \log_2(p_i^k).$$

The domain of the map is the geometric mean and its image is IC^k as defined by (10). Like other convex transformations of the geometric mean (e.g. the 'statistical rarity' metric of Rarity.tools and HowRare.is and Icy.tools) the skew of its score density will be greater than the skew of the geometric mean score density (Remark 3). That is, more tokens are allocated lower scores, compared with the geometric mean metric. However, in this case lower scores indicates fewer rare tokens, not a greater numbers of rare tokens, since the transformation is strictly decreasing. By Remark 2, OpenRarity must allocate ranks in reverse order, i.e. lowest score is least rare and highest score is most rare. Since neither raw measures nor standardized scores are published, this all this information is lost. In fact, the OpenRarity metric is not new at all, as claimed. The ranks provided by OpenRarity are identical to those provided by all other the sites that employ the geometric mean or some monotonic transformation thereof.

Classification of Rarity Metrics

The value of a rarity metric depends on the numbers of tokens m, traits n and also, implicitly on the number of attributes per trait, because this affects the trait counts and frequencies. For this reason, rarity measures derived from two different collections cannot be compared and neither can their standardized versions, i.e. two rarity scores from different collections cannot be compared. However, to each token A^k we may assign a rank R^k by ordering all rarity scores $\bar{r}^k$, $k = 1, ..., m$ in either increasing or decreasing order of magnitude.[32] Metrics that are a Pythagorean mean or an increasing transformation thereof sort the rarity measures or scores by increasing order of magnitude, so tokens are ranked from 1 to m with 1 being most rare (smallest rarity measurement, score 0) to m being the most common (greatest rarity measurement, score 1). But metric that are a decreasing transformation of a Pythagorean mean sort measures or scores by decreasing order of magnitude, so that the token with the greatest measurement, i.e. with score 1 has rank 1 and the token with the lowest measurement, i.e. score 0 has rank m.

This way, rank 1 always denotes the most rare token and rank m is the most common. Since any strictly increasing transformation does not alter the ordering of the scores it

[32]Ties should be dealt with in the usual way: when more than one token has the same rarity score they are assigned equal ranks but then one or ranks are skipped before assigning a rank to the token with the next highest measurement.

will not affect the rank of any token. The same comment applies for strictly decreasing transforms, since ranking now proceeds from high to low. In other words, rarity ranks are invariant under both strictly increasing and strictly decreasing continuous monotonic transformations.

Table 3: **Survey of Metrics used by Ranking Analytics Platforms**
When the model described on their website yields an identical ranking to a Pythagorean mean, we indicate this using ✓. Some sites, like Rarity.tools, provide more than one ranking model. The * indicates the ability to add/remove trait normalization. RaritySniffer, NFTStats and RarityMon are marked † because no details of their rarity models are available, but we have reverse engineered the model type by comparing their rankings with those provided by Rarity.tools. However, we were unable to match the rankings provided by CryptoSlam, MomentRanks, NFTinit and NFTEXP with any known model, and neither do these providers describe their methodology.

Platform	Minimum	Harmonic	Geometric	Arithmetic	Unknown
CryptoSlam					✓
HowRare.is		✓*	✓		
Icy.tools			✓		
LuckyTrader		✓*			
MomentRanks					✓
Nansen				✓	
NFTEXP					✓
NFTgo				✓	
NFTinit					✓*
NFTonchained		✓			
NFTSniff		✓			
NFTStats†		✓			
OpenRarity			✓		
RankNFT		✓			
Rarity.tools	✓*	✓*	✓*	✓*	
RarityMon†		✓			
RaritySniffer†		✓*			
Rarity Sniper		✓			
Traitsniper		✓			

For this reason, and using Remarks 1 and 2 above, we have shown that every known ranking metric in the public domain is a strict monotonic decreasing or increasing transformation of a Pythagorean mean. Hence, there are only four distinct ranking methodologies – the 'rarest trait' (used only by Rarity.tools) and the three Pythagorean means.[33] Table 3 categorizes the rarity models used by different platforms. Rarity.tools is the most flexible, allowing users to select the ranking model and, when the *trait normalization* mode is switched on, to apply different weights to attribute frequencies – see Section 3.3 for further investigation of this option. Most other platforms use a model that yields identical ranks to the harmonic mean and three of these (HowRare.is, LuckyTrader and RaritySniffer) also allow trait normalization to be turned on or off. HowRare.is also has a model that yields a ranking that is equivalent to the geometric mean ranking, as do

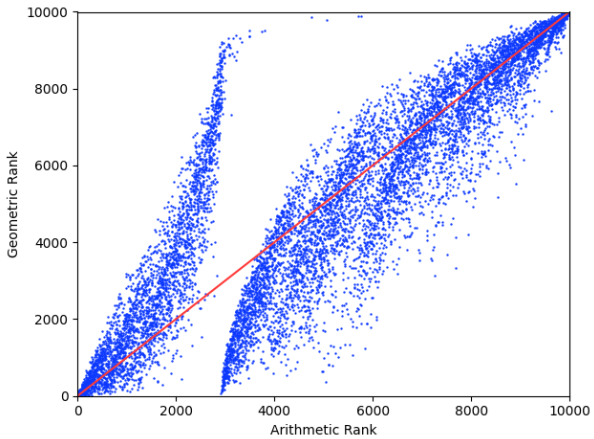
[33]For instance, many NFT analytics platforms use a monotonic transformation of the harmonic mean, including Rarity Sniffer, Trait Sniper, NFTSniff, RankNFT and NFT Stats. While true, this is not immediately obvious, because none of these analytics platforms use proper mathematical notation.

Icy.tools and OpenRarity. A few other platforms apply models that produce rankings that are equivalent to the arithmetic mean, and there are four platforms, namely CryptoSlam, Moment Ranks, NFTinit and NFTEXP, that publish rankings that we are unable to match with known model. Unfortunately, none of them disclose any details of their methodology, and the results diverge considerably from platform to platform.

Next we examine how divergent rarity ranks could be when derived from different Pythagorean means. To this end, in Figure 7 each point represents a BAYC token, and for each token we show the rarity rank ascribed by NFTGo (and Nansen) on the horizontal axis versus, on the vertical axis, the rank ascribed by OpenRarity (and IcyTools and HowRareIs). Each ranking ranges from 1 (most rare) to 10,000 (most common). The tokens falling on or near the black line are those for which the two rankings agree. It is apparent that the BAYC generative algorithm produced two different batches of tokens. Those in the left-hand cluster are typically ranked relatively rare by NFTGo but rather common by OpenRarity. The larger cluster on the right contains tokens that are relatively common according to NFTGo but OpenRarity only agrees with NFTGo about the tokens in the top far right of the scatter plot, for which both models agree are very common. There are many tokens lying well below the black line in this second cluster, meaning that OpenRarity classifies them as far more rare than NFTGo does.

Figure 7: **Arithmetic versus Geometric Rarity Ranks for the BAYC Collection.**

Each point represent a BAYC token. The horizontal axis is the arithmetic mean rarity measure applied by NFTGo (and Nansen and Rarity.tools) and the vertical axis is the geometric mean rarity measure applied by OpenRarity (and IcyTools and HowRareIs). Rank 1 = most rare; Rank 10000 = most common.

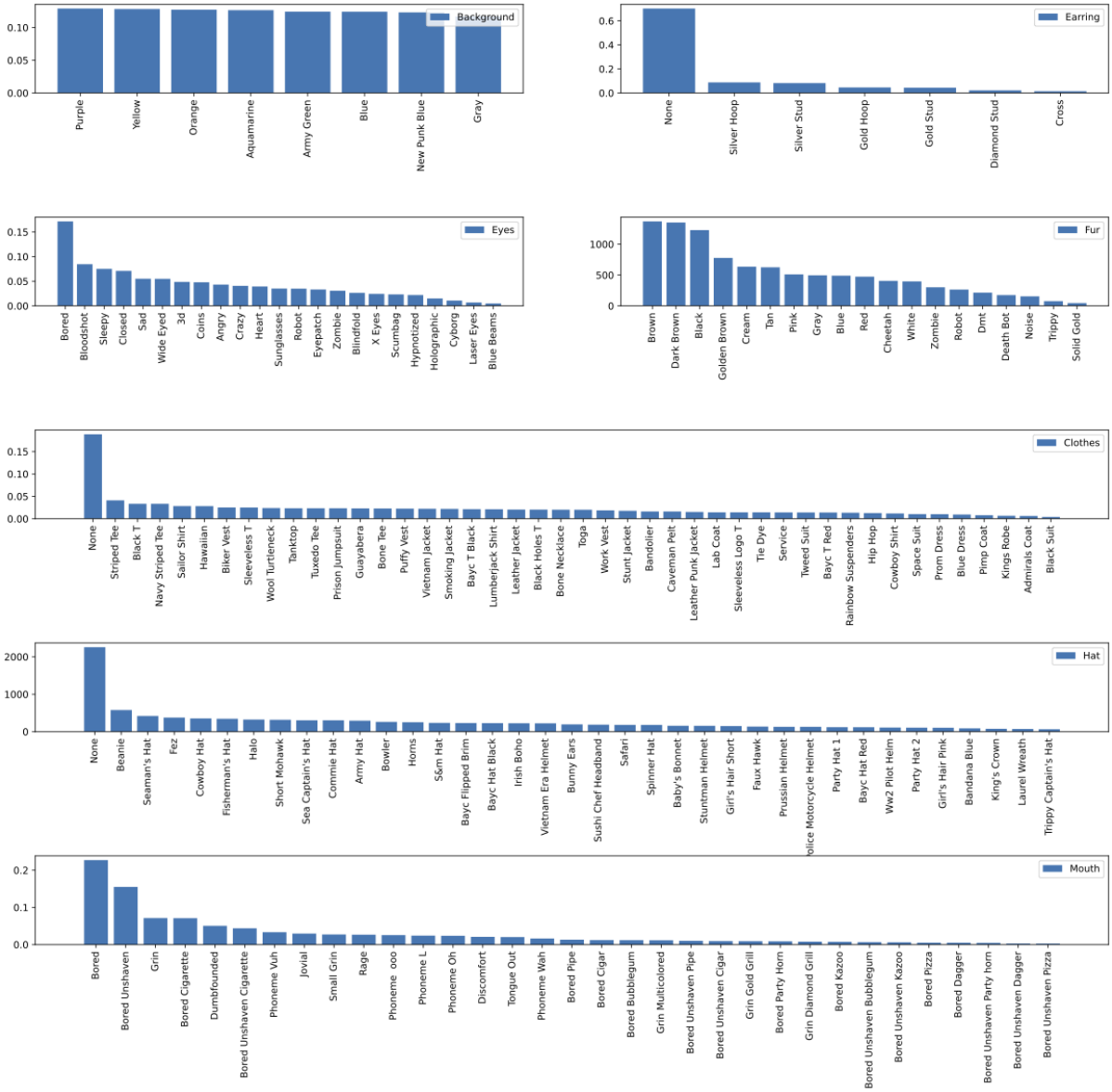


Finally, considering all three of our classic collections and all three Pythagorean means, Table 6 in the Appendix shows how divergent both scores and ranks can be depending on the metric used. The results shows that both rank correlations and score correlations can be surprisingly low.

3.3 Trait Normalization

Once the entire collection has been minted we have all the data required to compute rarity metrics, namely the *trait distributions* which define the allocations of attributes to traits over the entire collection. For example, Figure 8 depicts the trait distributions for BAYC. The background trait adds nothing to the rarity score, since all backgrounds are (more or less) equally likely; there are eight types of background, and each has a frequency of approximately one-eighth, i.e. 0.125. By contrast, having any sort of earring can add much to the final rarity score for an Ape in BAYC. That is, an earring type with frequency 0.125 is somehow more *special* than a background with frequency 0.125.

Figure 8: **Trait Frequencies in Bored Ape Yacht Club**
For each of the seven BAYC traits, we plot the empirical frequencies of different attributes.



For another (toy) example, suppose a collection has 20 different attributes for the mouth trait, such as smile, grin, sulk, big teeth, cigarette, etc. and suppose all are

equally likely so each type of mouth has a frequency of 5%. And suppose there are only three different types of skin, where 60% of the tokens have red skin, 35% have yellow skin, and only 5% of the tokens have blue skin. Having blue skin is very rare, and each mouth type is equally rare (actually, equally common). But a common frown has the same attribute frequency as the rare blue skin, *i.e.* 5%.

A few rarity analytics platforms allow users to weight the traits in such a way that some traits are more special than others, e.g. blue skin is more special than a frown. Rarity.tools even allows weights to be determined by personal preferences alone, e.g. when an investor has a penchant for X-Ray eyes, and so desires that tokens possessing this attribute be ranked higher than they would normally be. In this case, the user can simply select *ad hoc* weights according to their own preferences. But the 'trait normalization' offered by several sites (see Table 3) is just an on/off toggle. There is no opportunity to adjust weights. And the method used for trait normalization is not explained anywhere – not even on Rarity.tools.[34]

In this subsection we explain how to implement trait normalization by selecting the appropriate weights in a *power mean* rarity metric. This resolves another open issue, because exactly how traits should be normalized is one more area of obfuscation by rarity metric providers, and so also a source of widespread misunderstanding among practitioners. Using different weights in the power mean allows users to standardize the results when some traits have many attributes and others only a few.

Let p be any non-zero real number and $\{\omega_1, \ldots, \omega_n\}$ be a set of positive weights. Then the weighted power mean with exponent p of a set of n positive real numbers $\{a_1, \ldots, a_n\}$ is defined as:

$$M_p\left(a_1, \ldots, a_n | \omega_1, \ldots, \omega_n\right) = \left(\frac{\sum_{i=1}^{n} w_i a_i^p}{\sum_{i=1}^{n} w_i}\right)^{1/p}. \tag{11}$$

This single formula provides a unified rarity metric that encompasses all the rarity metrics currently in the public domain. The exponent parameter p in (11) can be any non-zero real number, with special definitions for limiting cases as $p \to \pm\infty$ and for $p = 0$. Table 4 exhibits the weighted power means that are of specific interest because they are equivalent to the rarity metrics currently used.

When $\{\omega_1, \ldots, \omega_n\} = \{1, \ldots, 1\}$ the formula (11) reduces to the ordinary, non-weighted power mean which is also called the *generalized mean*:

$$M_p(a_1, \ldots, a_n) = \left(n^{-1} \sum_{i=1}^{n} a_i^p\right)^{1/p}, \tag{12}$$

We have already encountered some special cases of the generalized mean as the Pythagorean means, but there are infinitely many more because p can be any real number, not neces-

[34] For which the only public information we can find is the utter confusion evident from this Reddit discussion ending in a request to "share the math".

Table 4: **Special Cases of the Weighted Power Mean** (11)

Rarity Metric	p	Definition
Weighted Minimum	$-\infty$	$M_{-\infty}(a_1,\ldots,a_n\|\omega_1,\ldots,\omega_n) = \min\{\omega_1 a_1,\ldots,\omega_n a_n\}$
Weighted Harmonic Mean	-1	$M_{-1}(a_1,\ldots,a_n\|\omega_1,\ldots,\omega_n) = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{a_i}}$
Weighted Geometric Mean	0	$M_0(a_1,\ldots,a_n\|\omega_1,\ldots,\omega_n) = \left(\prod_{i=1}^n a_i^{\omega_i}\right)^{1/\sum_{i=1}^n \omega_i}$
Weighted Arithmetic Mean	1	$M_1(a_1,\ldots,a_n\|\omega_1,\ldots,\omega_n) = \frac{\sum_{i=1}^n w_i a_i}{\sum_{i=1}^n w_i}$
Weighted Maximum	$+\infty$	$M_{\infty}(a_1,\ldots,a_n\|\omega_1,\ldots,\omega_n) = \max\{\omega_1 a_1,\ldots,\omega_n a_n\}$

sarily an integer. Some common examples are included in Table 5. The metrics displayed in that table use frequencies, e.g. to obtain the geometric mean for the token A^k we use $\{a_1,\ldots,a_n\} = \{p_1^k,\ldots,p_n^k\}$ in the formula, which is clearly then identical to the geometric mean in Table 2. Similar comments apply to all of the metrics in Table 2.

Table 5: **Special Cases of the Generalized Mean** (12)

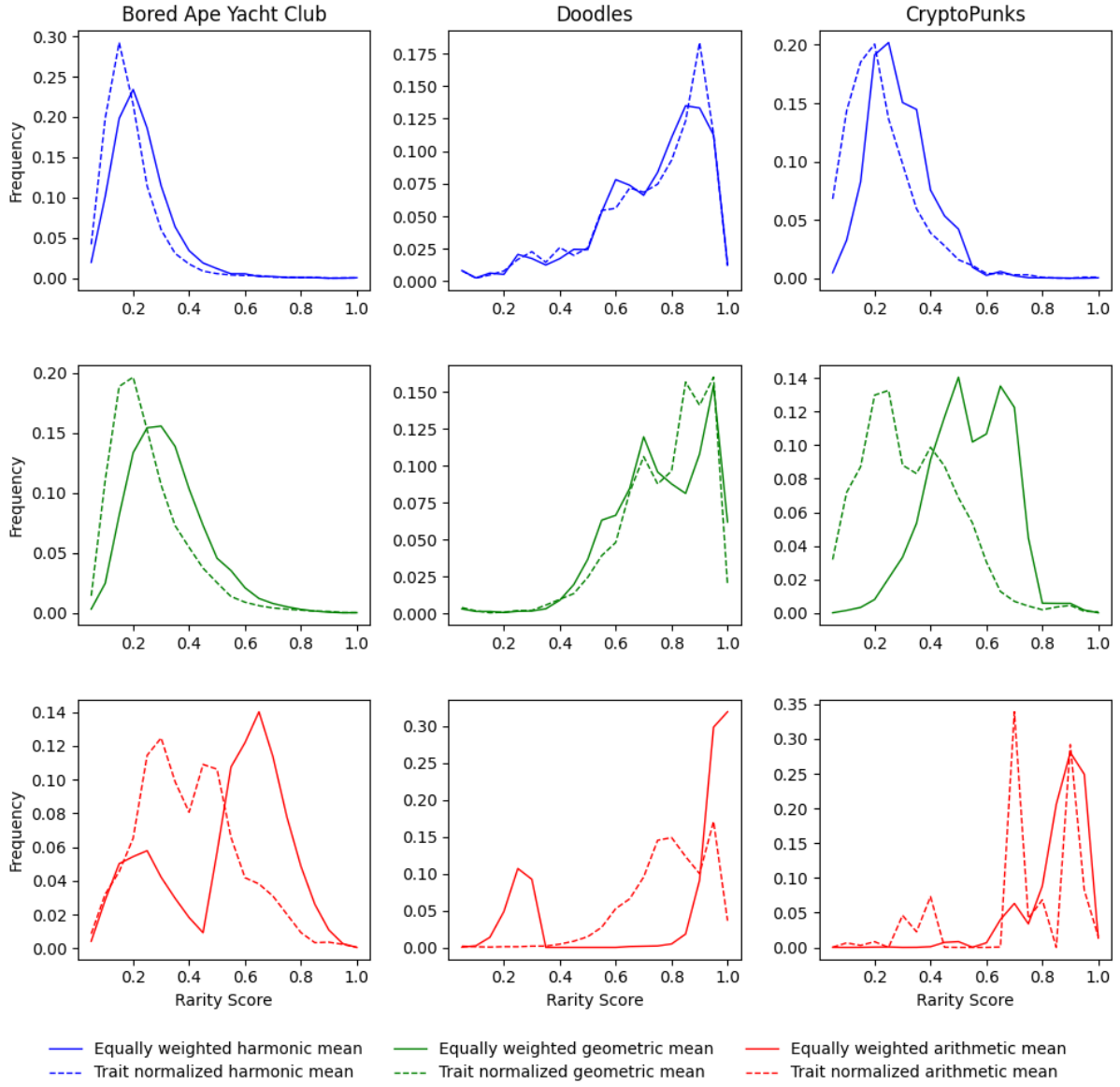
Rarity Metric	Definition
Minimum	$\lim_{p\to-\infty} M_p(a_1,\ldots,a_n) = \min\{a_1,\ldots,a_n\}$
Harmonic Mean	$M_{-1}(a_1,\ldots,a_n) = \frac{n}{\sum_{i=1}^n \frac{1}{a_i}}$
Geometric Mean	$\lim_{p\to 0} M_p(a_1,\ldots,a_n) = M_0(a_1,\ldots,a_n) = \left(\prod_{i=1}^n a_i\right)^{1/n}$
Arithmetic Mean	$M_1(a_1,\ldots,a_n) = n^{-1}\sum_{i=1}^n a_i$
Maximum	$\lim_{p\to\infty} M_p(a_1,\ldots,a_n) = \max\{a_1,\ldots,a_n\}$

In light of our results in the previous section, we could just as well have defined the metrics in Table 2 using counts rather than frequencies because they are simply related via (5). In other words, the same equations may be written using counts instead of frequencies, then dividing each sum or product thus obtained by m. And being a linear transformation, division by m does not even affect the scores defined by (6). In other words, since scores rather than ranks are the important variable, we can apply the Pythagorean mean formulae to counts or frequencies, it makes no difference. In fact, for reasons that become clear later we shall use counts rather than frequencies, thus setting $\{a_1,\ldots,a_n\} = \{m_1^k,\ldots,m_n^k\}$ in the general formula (11) for the weighted power mean.

Now we consider how the weights $\{\omega_1,\ldots,\omega_n\}$ can be chosen for trait normalization.

To motivate the problem consider a simple example of two traits in a collection with 10,000 tokens where 'mouth' has 100 different types, all equally likely, so 100 tokens have a mouth of type 1 and so on. And 'eyes' have only two different types, 99% of tokens have blue eyes and only 1% have brown eyes so 9,900 have blue eyes and only 100 have brown eyes. Even though any given token has a 1% chance of brown eyes and a 1% chance of any type of mouth, having brown eyes is clearly much more special than any sort of mouth. We want to weight these two 100 counts, or 1% frequencies differently, and we may do so by assigning a *smaller* weight to the eyes frequencies than to the mouth frequencies (recall, smaller frequencies are rarer). One simple way to do this is to let the weights for each trait $i = 1, \ldots, n$ be equal to number of attributes, *i.e.* to set $\omega_i = \theta_i$.

Figure 9: **Rarity Score Densities Before and After Trait Normalization**
Rarity score densities for unweighted (solid) and weighted (dashed) Pythagorean means: harmonic (blue, upper), geometric (green, middle) and arithmetic (red, lower)



To see the effect of trait normalization we compare the harmonic, geometric and arithmetic mean rarity scores before and after the normalization, for the same three collections as previously analysed. The trait normalised (weighted) means scores are depicted using the dashed lines, and the unweighted scores, which are identical to those shown in Figure 6, are shown in the same color as that figure, *i.e.* blue for the harmonic mean, green for the geometric and red for the arithmetic mean. The results are depicted in Figure 9. Clearly, the effect of trait normalization depends on both the collection and the rarity metric applied.

We conclude this section by noting that using counts rather than frequencies and setting $\omega_i = \theta_i$ can be justified because, in this way, the weights for normalization need not depend on p. By using counts it does not matter whether the weight is a multiplicative or exponent factor in the weighted power mean: setting $\omega_i = \theta_i$ for trait normalization is independent of p when using counts. On the other hand, a little trivial algebra shows that using frequencies not counts for $\{a_1, \ldots, a_n\}$ *would* require setting an exception, in that $\omega_i = \theta_i$ would only reduce a measurement where smaller means more rare when $p \neq 0$. In the case that $p = 0$ we would need to set $\omega_i = \theta_i^{-1}$ not $\omega_i = \theta_i$, as it is for all other values of p. Our recommendation for trait normalization being based on counts rather than frequencies is not the only possibility for assigning more weight to traits with fewer attributes, but we have promoted this method because it is the only consistent rule that allows weights to be assigned independently of p.

4 Tools for Analysing Personal Profile Picture Collections

In the following, subsection 4.1 examines the crucial role that trait independence plays in identifying the existence of a mathematically correct metric for rarity (i.e. the geometric mean, or monotonic transformations thereof) and explains how to test for independent traits with illustrative examples. And since the previous section has emphasized that rarity scores are very much more informative than ranks, but currently most platforms only report ranks, subsection 4.2 examines what we can learn from differences in ranks. It introduces the concept of a unique bar code plus a set of QR codes which provide a visual summary of the overall characteristics of a PFP collection. Then we analyse the bar and QR codes for our three well-known collections, and explain how the QR codes may be used to identify the potential for ranks based on different metrics to be highly discordant.

4.1 Importance of Independent Traits

When considering the mathematical correctness of the different ranking methodologies, it is important to know whether a PFP collection generation protocol has allocated at-

tributes using independent or dependent traits.[35] For example, is the color of the eyes affected by the background that is allocated or not? And likewise, does the hair style depend on the gender? In other words, can the allocation of attributes be thought of as independent events? If the answer is yes, in other words the traits are independent random variables, then there is only one mathematically correct rarity metric – which is the geometric mean, and any strictly transformation thereof.

The allocation of attributes during the process of minting NFTs can be thought of as a probabilistic event. Let us consider three independent events A, B and C. For instance, A could be the event "allocate gender = male", B could be the event "allocate hair = long" and C could be the event "allocate eyes = green". The probability of *intersection* of all three events, that is of finding a token possessing all three attributes may then be written Prob(A and B and C), and when A, B and C are independent then the rules of probability state that

$$\text{Prob}(A \text{ and } B \text{ and } C) = \text{Prob}(A) \times \text{Prob}(B) \times \text{Prob}(C).$$

This rule also applies to any number of independent events. That is, to find the probability that a token has a given set of attributes selected from independent traits, we need to multiply the trait frequencies. Using our results in Section 3, the product of trait frequencies yields ranks that are identical to the geometric mean. Hence, the geometric mean of the frequencies, or of the counts, and any metric that is a monotonic transform of this mean, is mathematically correct but only when traits are independent. Table 7 in the Appendix tests for independence of traits in the BAYC and Doodles collections. From these results it appears that only BAYC was designed to have independent traits.

The arithmetic mean *adds* these counts or frequencies. But there is only one context in which one should add probabilities, and that is when we want to find the much larger *union* of all events, Prob(A or B or C). In the special case that A and B and C are *mutually exclusive* events, that is when Prob(A and B and C) = Prob(A and B) = Prob(A and C) = Prob(B and C) = 0 we have

$$\text{Prob}(A \text{ or } B \text{ or } C) = \text{Prob}(A) + \text{Prob}(B) + \text{Prob}(C).$$

But the entire point of token collections is that traits are not mutually exclusive. For instance, using the example of Table 1 again, there is a unique token that is male and has long hair and has green eyes, so labelling the events male, long hair, green eyes as A, B and C accordingly, we have Prob(A and B and C) = 1/60, not zero. Hence adding the trait frequencies makes no statistical sense at all. A similar comment applies to the

[35] In probabilistic language, using '|' to denote 'given', or being conditional on, if: Prob($C|A$) = Prob($C|B$) = Prob(C) then event C is independent of event A and independent of event B. When all three events are independent similar equations apply on interchanging A, B and C.

harmonic mean. Indeed, the harmonic mean metric has no probabilistic interpretation whatsoever. Nevertheless, it has become the most common metric of all, as seen from Table 3.

When traits are independent the geometric mean is a statistically-correct rarity metric. It is possible (though laborious) to test for trait independence using a simple chi-squared test.[36] The results for our three classic collections are reported in Table 7 of the Appendix, and a summary of independence test for the other collections is also provided there. We conclude that less than 10% of the collection have independent traits, one of these being BAYC. But the traits in Doodles and CryptoPunks are clearly highly dependent.

4.2 Codes for Summarizing Collection Characteristics

For any token A^k the generalized mean frequencies (or counts) have the ordering:

$$r^k_{\text{min}} \leq r^k_{\text{harmonic}} \leq r^k_{\text{geometric}} \leq r^k_{\text{arithmetic}} \leq r^k_{\text{max}} \qquad k = 1, \ldots, n. \tag{13}$$

However, this ordering becomes lost once raw measurements are converted to scores or ranks, R^k. The only thing we know is that the larger the interval $[r^k_{\text{min}}, r^k_{\text{max}}]$ the greater the potential for divergence between the harmonic, geometric and arithmetic rarity scores or ranks. With this intuition, we now introduce our first collection-wide characteristic, which we call it's *bar code* for reasons that should be obvious from its display. The bar code is a visualization of the map:

$$k \mapsto r^k_{\text{max}} - r^k_{\text{min}}, \qquad k = 1, \ldots, n. \tag{14}$$

Each point in the bar code corresponds to a specific token in a collection. The horizontal axis is the ID given to the token when it was minted onto the blockchain, and on the vertical axis we plot the difference between the maximum attribute frequency and the minimum attribute frequency for that token, as described in (14). The top line in the bar code indicates how divergent the rarity scores derived from different metrics can be: the higher the top line, the greater the potential difference between rarity scores. For instance, the tokens with points along a line close to the maximum value of one must have at least one very common attribute and at least one very rare attribute. The lower the top bar the more evenly distributed are the attributes across different tokens.

Since the industry reports rarity scores as ranks we need a way to compare the ranks produced by different rarity metrics, rather than their scores. For this we introduce another novel concept, that of the collection's *QR codes*. Unlike the bar code, which is unique, there can be many QR codes. A QR code is a scatter plot of any pair of rarity

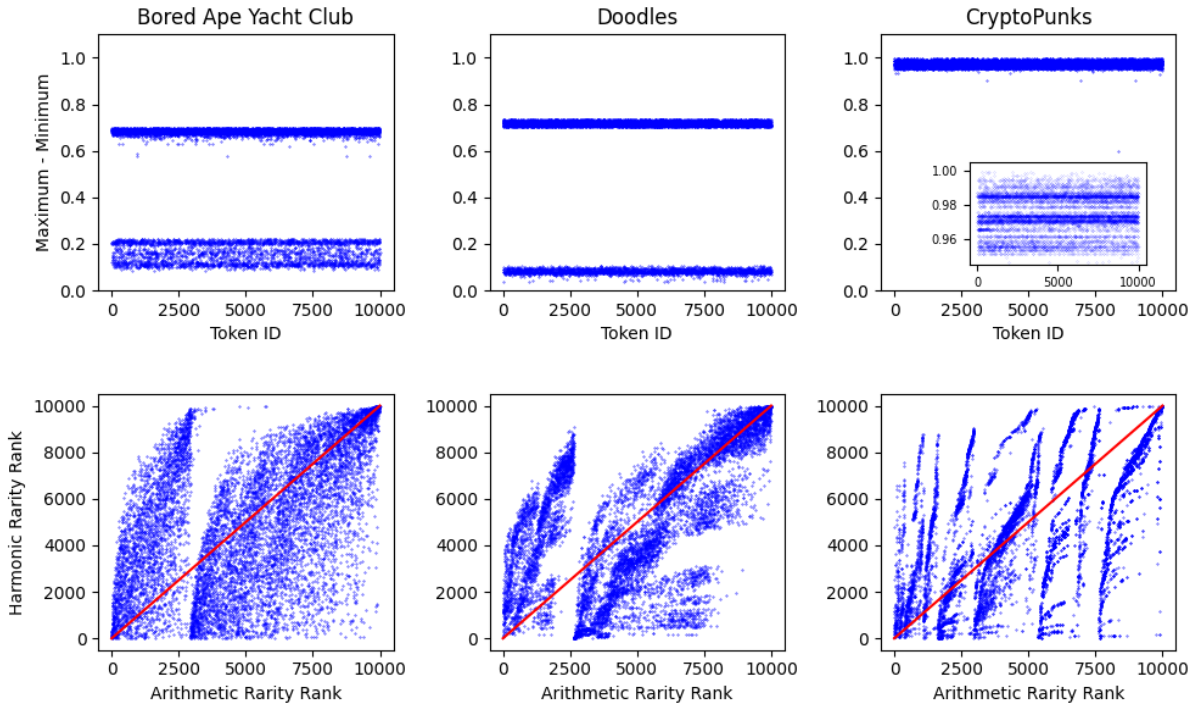
[36] See the chi-squared test for categorical data. However, when there are many traits types the contingency tables can become rather sparse, and chi-squared tests then lack robustness. For this reason our code also outputs the Cramer's V statistics for different pairs.

ranks, (R_1, R_2) where, again, each point corresponds to a specific token in the collection:

$$k \mapsto (R_1^k, R_2^k), \qquad k = 1, \ldots, n. \tag{15}$$

The 45° line on a QR plot represents those A^k for which $R_1^k = R_2^k$, *i.e.* those tokens which have identical ranks under both rarity metrics. The further a token's point lies from this line, the greater the difference between its rarity ranks, according to these two metrics.

Figure 10: **Bar and QR Codes for BAYC, Doodles and CryptoPunks**
Bar codes (above) and harmonic-arithmetic QR codes (below) for our three collections.



We present the bar codes and harmonic-arithmetic QR codes for 196 PFP collections in the Appendix, using the data provided by Nadini et al. (2021). To help readers interpret these, Figure 10 compares the bar and harmonic-arithmetic QR codes for our three classic PFP collections, in the upper and lower panels of this figure respectively. All three collections have a large number of tokens for which the bar codes (max-min plots) have a high-level line, the highest line being near one for CryptoPunks. This mean that most Punks have a max score close to one (one quite common trait) and a min score close to zero (one quite rare trait).[37] As a result, due to the inequality (13), there is considerable scope for divergence between different Pythagorean (or other) metrics, so the collection as a whole is not very robust to changes in the ranking methodology. For instance, CryptoPunks will be ranked very differently, depending on the rarity metric

[37] A notable exception is the Punk with token ID 8348, which is the only Punk having all ten attributes. The max-min value for this punk is a little less than 0.6 and you can see its isolated point just above the insert to the bar code.

used. For this collection we have also zoomed in on the line near one in the bar code, to see more of its structure (see the inset to the top right plot). This reveals a stratification of several lines at similar but slightly different levels. This pattern affects the QR code, as we shall discuss below. The upper lines for BAYC (left) and Doodles (middle) are lower, nearer to 0.7 than to 1, but still there is much scope for discordant rankings to be generated by different metrics. BAYC and Doodles have another definite line in their bar codes, between 0.1 and 0.2 for BAYC and between 0.5 and 0.9 for Doodles. The tokens lying along these lines have much less divergence in their rankings according to different methods. In fact, these are the tokens which lie near the 45° line in the QR code.

To see this, the lower panel in Figure 10 presents the QR Code, defined by (15) for the two most commonly used rarity metrics in the industry (at the time of writing) *i.e.* the harmonic mean and the arithmetic mean. Again, each point corresponds to one token, and we simply plot the harmonic mean of its attribute frequencies on the vertical axis and the arithmetic mean on the horizontal axis. There are two distinct clusters of tokens for BAYC and for Doodles, but not for CryptoPunks. For instance, in BAYC there is a clear break at around arithmetic rank 2900: almost all the BAYC tokens that are arithmetically ranked between 0 and about 2500 are much less rare according the harmonic rank; and a (slightly less) large proportion of the tokens that are arithmetically ranked between about 2500 and 10,000 are much more rare according the harmonic rank.[38]

The clusters observed in BAYC and Doodles result from the bimodal feature in their arithmetic mean scores, whereas their harmonic mean scores densities are unimodal, as evident from Figure 6.[39] The red line at 45° superimposed on the QR codes shows where the two metrics yield equal ranks; that is, any point lying along the red line corresponds to a token that has the same rank under both arithmetic and harmonic rarity metrics. The further from this line the point lies, the greater the discrepancy between the rankings reported on different marketplaces or analytic sites for that token.

It is *not* the presence of two distinct lines in the bar code that translates into two distinct clusters in the QR code, as can be verified from the bar and QR codes displayed in the Appendix 5. There we see numerous collections (e.g. Bones and Bananas, Bonsai, Bored Mummy Waking Up, and many more) that have two clusters in the QR code but only one line, or they can have many lines in their bar code. The two distinct clusters in QR codes are instead related to the bimodal nature of just one of the two score densities.

[38] A similar discordant ranking anomaly occurs for Doodles: most of the tokens that are arithmetically ranked between about 0 and 2500 are much less rare according to the harmonic rank; and many of the more common that are arithmetically ranked between about 2500 and 10,000 are more rare according to the harmonic rank. By contrast, the QR code for CryptoPunks is quite different – there are still many tokens that have very different ranks depending on the metric, but no ordering in the discordance of ranks is evident here.

[39] Although there is a smaller mode for the arithmetic score at about 70 for CryptoPunks, the other score densities are also bimodel, and as a result the CryptoPunk QR code has several strata rather than two definite clusters.

QR plots can be used to identify whether one specific token has similar or very different ranks, according to the metric used. This can also help investors decide which platform to use. For instance, if token A^k has a high rank from platform A which uses the geometric mean and a low rank from platform B which employs the harmonic mean, then it is better to sell on platform A and buy on platform B. QR plots also provide insight to the efficiency of the market for that PFP collection. If most points in the QR plot lie close to the 45° line, the two metrics rank most tokens similarly, and so the relationship between rarity and price will not be blurred by different platforms providing totally different ranks. Similarly, the more dispersed the QR scatter plots, the less the confidence associated with any price-rarity relationship.

5 Summary and Conclusions

A universally accepted definition of rarity which can be measured using a mathematically correct metric has become a major barrier to establishing greater efficiency in PFP markets. It is fundamental problem which requires attention before other forms of data analysis on NFTs can be effectively researched. We start with a necessary clarification of terminology for discussions about rarity metrics for PFP collections. This paper is also the first in the field to even use proper mathematical notion.

First we show that rarity scores are invariant (reversed) under increasing (decreasing) linear transformations of the rarity metric, and that non-linear, strict monotonic transformations change the shape of the rarity score density so that its skewness increases (decreases) when the function is convex (concave). Then we show that ranks are invariant under any strict monotonic transformation of the rarity metric. Such a transformation typically changes both the raw measure and its standardized version, the score, but these effects disappear once we convert the score to a rank. Scores contain far more information than ranks do about the distribution of rarity for all tokens in the collection.

Then we prove that every known PFP rarity metric is a strict monotonic transformation of a weighted power mean. Notably, we have proved that the Jaccard distance metric proposed by NFTGo is a decreasing linear transformation of the unweighted arithmetic mean, so it produces the same rarity scores and ranks as the arithmetic mean does, for each token in the collection. We also prove that the open-source code provided by OpenRarity (a consortium of OpenSea, Coinbase, PROOF, X2Y2, icy.tools, LooksRare, Rarible, Curio and others) is a convex, monotonic decreasing transformation of the geometric mean, so its ranking of tokens is identical to the ranking produced by the unweighted geometric mean.[40] Their claims to have developed a new combinatorial or information-theoretic approach are, therefore, not true.

All known methods provide ranks that are identical to those produced by a generalized

[40] Although the scores do change because the OpenRarity score distribution has greater skew than the geometric mean score distribution.

mean with four possible values of the exponent p. Indeed, apart from the 'rarest trait', which is the limit of the power mean as $p \to -\infty$, every code in the public domain uses (a possibly weighted version) of either the arithmetic or the geometric or the harmonic mean. It does not matter whether we take the mean of attribute counts or frequencies. We prefer to employ counts because this gives a consistent method to assign weights in the mean for trait normalization that is independent of p.

Most of the rarity metrics currently in use are mathematically incorrect. More specifically, at most one of the power mean metrics, the geometric one corresponding to $p = 0$, could be correct, but even then it can only be applied to very special collections where traits are independent. We have provided a method to test for this, and only a few of the $\sim$200 collections that we analyze have independent traits.

Next we summarize the characteristics of an entire PFP collection using two novel concepts: a unique bar code which is a graphical representation of the difference between the rarity scores corresponding to the two extreme exponent values $p \in \{-\infty, +\infty\}$; and a set of QR codes for the collection, corresponding to scatter plots of rarity ranks for two different values of p. These codes are also a simple way to identify exactly which tokens suffer from a discordant ranking anomaly whereby the token is classed as rare according to one metric and common according to another. Our theoretical results, code and data visualization tools may be useful for NFT traders; to know which marketplaces giving a particular token the highest ranking – sell there; and which platforms give a token the lowest ranking – buy there.

The generalized power means that are the focus of this paper can always be regarded as a *distance* metric – but then there are literally hundreds of other distance metrics which could serve as alternatives (Deza and Deza, 2009). The NFT market cannot evolve if any subjectively-chosen distance metric could be chosen to represent 'rarity'. Yet, while so much disagreement on rarity ranks remains, the relationship between price and rarity will be impossible to model, making the entire PFP market highly inefficient. The development of a universal metric for rarity, which allows traits to be dependent and applies to all NFTs (such as breedable collections which have an ever-expanding supply) is a highly complex problem which is beyond the scope of this paper. While such a metric is being developed and accepted, our paper brings much-needed clarity to the concept of rarity in PFP collections.

References

Ante, L. (2022), 'Non-fungible token (NFT) markets on the Ethereum blockchain: Temporal development, cointegration and interrelations', *Economics of Innovation and New Technology* **1**, 1–19.

Bonifazi, G., Cauteruccio, F., Corradini, E., Marchetti, M., Montella, D., Scarponi, S., Ursino, D. and Virgili, L. (2023), 'Performing wash trading on NFTs: Is the game worth the candle?', *Big Data and Cognitative Computing* **38**(7).

Chalmers, D., Fisch, C., Matthews, R., Quinn, W. and Recker, J. (2022), 'Beyond the bubble: Will NFTs and digital proof of ownership empower creative industry entrepreneurs?', *Journal of Business Venturing Insights* **17**, Article No. 00309.

Chandra, Y. (2022), 'Non-fungible token-enabled entrepreneurship: A conceptual framework', *Journal of Business Venturing Insights* **18**, Article No. 00323.

Cho, J., Serneels, S. and Matteson, D. (2023), 'Non-fungible token transactions: Data and challenges', *Data Science in Science* **2**(1), Article No. 2151950.

Davidsson, P., Recker, J. and von Briel, F. (2020), 'External enablement of new venture creation: A framework', *Academy of Management Perspectives* **34**(3), 311–332.

DeGennaro, R. P. and Robotti, C. (2007), 'Financial market frictions', *Economic Reviews* **92**, 1–16.

Deza, M. and Deza, E. (2009), *Handbook of Distances*, Springer, Berlin.

Driotcour, B. (2021), 'Generative art and nfts', *Art in America* **http://bitly.ws/SEED**.

Hou, K. and Moskowitz, T. J. (2005), 'Market frictions, price delay, and the cross-section of expected returns', *Review of Financial Studies* **18**(3), 981–1020.

Katte, S. (2022), 'Nifty news: Nike unveils NFT platform, Steve Jobs' sandals sell for $200k and more', *Cointelegraph, http://bitly.ws/Sg92* .

Kong, D. R. and Lin, T. C. (2021), 'Alternative investments in the FinTech era: The risk and return of non-fungible tokens', *ssrn: 3914085* .

Lapuschin, M. (2022), 'Best NFT rarity tools', *Sensorium* .

Lee, Y. (2022), 'Measuring the impact of rarity on price: Evidence from NBA top shot', *Marketing Letters* **33**(3), 485–498.

Mukhopadhyay, M. and Ghosh, K. (2021), 'Market microstructure of non fungible tokens', *ArXiv: 2112.03172* .

Nadini, M., Alessandretti, L., Di Giacinto, F., Martino, M., Aiello, L. M. and Baronchelli, A. (2021), 'Mapping the NFT revolution: Market trends, trade networks, and visual features', *Scientific Reports* **11**(1), 1–11.

NFTGo (2021), 'The ultimate guide to NFTGo's new rarity model', *Medium, http://bitly.ws/Sg7C* .

Oh, S., Rosen, S. and Zhang, A. L. (2022), 'Investor experience matters: Evidence from generative art collections on the blockchain', *ssrn: 4042901* .

Schaar, L. and Kampakis, S. (2022), 'Non-fungible tokens as an alternative investment: Evidence from CryptoPunks', *Journal of the British Blockchain Association* **5**(1), Article No. 31949.

Serneels, S. (2023), 'Detecting wash trading for non-fungible tokens', *Finance Research Letters* **52**, Article No. 103374.

Tariq, S. and Sifat, I. (2022), 'Suspicious trading in non-fungible tokens: Evidence from wash trading', *ssrn: 4097642* .

Vidal-Tomás, D. (2023), 'The illusion of the metaverse and meta-economy', *International Review of Financial Analysis* **86**, Article No. 102560.

Wachter, V., Jensen, J., Regner, F. and Ross, O. (2022), 'NFT wash trading: Quantifying suspicious behaviour in NFT markets', *ArXiv: 2022.03866* .

Wilson, C. M. (2012), 'Market frictions: A unified model of search costs and switching costs', *European Economic Review* **56**(6), 1070–1086.

Appendix

Correlations Between Rarity Metrics

Table 6: **Rarity Score and Rank Correlations**
For the three Pythagorean means, we report (a) Pearson's correlations between rarity scores (above); (b) Spearman's ρ between rarity ranks (middle) and Kendall's τ between rarity ranks (below) for Bored Ape Yacht Club (left), Doodles (middle) and CryptoPunks (right). Abbreviations are 'Har.' for the harmonic mean, 'Geo.' for the geometric mean and 'Arith.' for the arithmetic mean.

	BoredApeYachtClub			Doodles			CryptoPunks		
Panel A: Pearson's ρ on Rarity Scores									
	Har.	Geo.	Arith.	Har.	Geo.	Arith.	Har.	Geo.	Arith.
Har.	1	0.872	0.463	1	0.792	0.14	1	0.805	0.461
Geo.	0.872	1	0.780	0.792	1	0.67	0.805	1	0.850
Arith.	0.463	0.780	1	0.14	0.67	1	0.461	0.850	1
Panel B: Spearman's ρ on Rarity Ranks									
	Har.	Geo.	Arith.	Har.	Geo.	Arith.	Har.	Geo.	Arith.
Har.	1	0.827	0.514	1	0.805	0.605	1	0.809	0.593
Geo.	0.827	1	0.866	0.805	1	0.935	0.809	1	0.917
Arith.	0.514	0.866	1	0.605	0.935	1	0.593	0.917	1
Panel C: Kendall τ on Rarity Ranks									
	Har.	Geo.	Arith.	Har.	Geo.	Arith.	Har.	Geo.	Arith.
Har.	1	0.641	0.374	1	0.654	0.453	1	0.635	0.442
Geo.	0.641	1	0.716	0.654	1	0.798	0.635	1	0.773
Arith.	0.374	0.716	1	0.453	0.798	1	0.442	0.773	1

Table 6 reports the Pearson correlation between different rarity scores; and the Spearman and Kendall rank correlations between different rarity ranks. Spearman's ρ and Kendall's τ are non-parametric measures of rank correlation between two variables that are related via a monotonic function. They are more general than Pearson's correlation, which only applies to linear functions of normal variables. Each correlation coefficient is based on a sample size of 10,000 paired observations where each point corresponds to a token.

As expected from the Pythagorean mean orders and from the score density plots in the text, correlations are lowest between harmonic and arithmetic means, since these are furthest apart. Yet, they are also the two predominant metrics used by the industry, adding further weight to the many confusions that traders and investors in NFTs have surrounding rarity measurement.

Tests for Trait Independence

Table 7 reports the Chi-squared and Cramer's V results for BAYC and Doodles. For BAYC most chi-squared tests cannot reject the null of independence at the 5% significance level. That is, the entries above the diagonal are lower than the entries below the diagonal – with only three exceptions picked out in bold. Background and clothes show some

Table 7: **Independence Test Results for Bored Ape Yacht Club and Doodles**
For each panel, we present two sets of results. In the first set, the upper triangle contains the test statistics of each pair of traits, where the test for each trait has a different contingency table (not reported here for brevity). The lower triangle for the independence tests sates the 95% critical values of the chi-squared test statistic for the corresponding contingency table. The second set of results in each panel reports the Cramer's V value of each pair of traits.

Panel A: Bored Ape Yacht Club

	Background	Clothes	Earring	Eyes	Fur	Hat	Mouth
A.1: Independence test							
Background		**359**	40	149	109	282	219
Clothes	342		271	995	760	**1961**	1372
Earring	58	296		126	125	269	174
Eyes	184	1019	160		377	**1163**	675
Fur	153	840	133	443		628	529
Hat	290	1641	251	859	708		1332
Mouth	260	1463	225	767	633	1232	
A.2: Cramer's V							
Background		0.0716	0.0257	0.0461	0.0395	0.0635	0.0560
Clothes			0.0672	0.0673	0.0650	0.0738	0.0655
Earring				0.0459	0.0456	0.0670	0.0538
Eyes					0.0458	0.0727	0.0554
Fur						0.0591	0.0542
Hat							0.0645

Panel B: Doodles

	Face	Hair	Body	Background	Head	Piercing
B.1: Independence test						
Face		116677	127257	65707	145995	169
Hair	4164		121055	100919	212548	229
Body	3154	4164		78390	142748	205
Background	1859	2450	1859		109054	53
Head	2536	3345	2536	1496		126
Piercing	196	255	196	120	160	
B.2: Cramer's V						
Face		0.4607	0.4811	0.4533	0.5762	0.0750
Hair			0.4693	0.5617	0.6952	0.0873
Body				0.4951	0.5697	0.0826
Background					0.5839	0.0419
Head						0.0649

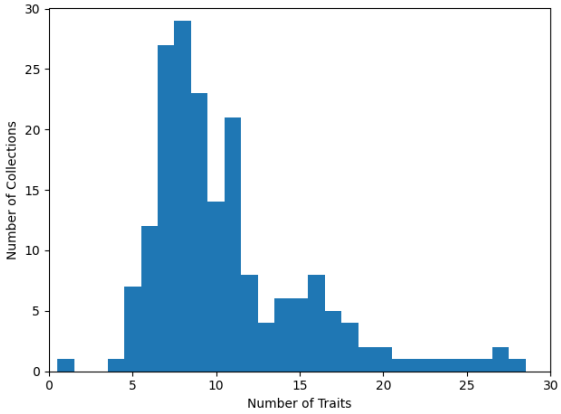
weak dependence, as does hat with clothes and eyes. However, regarding the lower part of the panel, the Cramer's V sense check reveals no great difference between the statistics for these three pairs, compared with the rest. We conclude that the finding of weak dependence for these three pairs is most likely due to sampling error, with only 10,000 tokens generated from well over one million possible distinct tokens, and that the BAYC generative algorithm was indeed set to generate traits independently. The lower panel of Table 7 depicts the results for Doodles, and here a completely opposite picture

emerges. Virtually every pair of traits shows highly significant dependence, with the notable exception of the piercing trait which appears to have been allocated independently of the other Doodles traits.

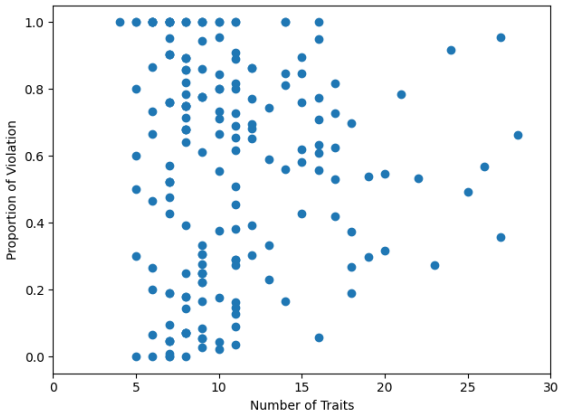
From the data provided by Nadini et al. (2021) we filter out those with less than 8,888 tokens, yielding 196 collections. Then Figure 11 displays the frequency of traits for the 191 collections (BAYC and Doodles included) having 30 traits or less, and the proportion of violations calculated as the number of rejections of the chi-squared test for independence divided by the total number of distinct trait pairs.[41] The right-hand scatter plot indicates that the majority of collections have more than 10% non-independent traits. There are only five collections for which all traits are completely independent, namely Dope Shibas, OnChainMonkey, Fly Frogs, Rabbit College Club and PyMons. Another 19 collections (including BAYC) have a proportion of violation less than 10%, and which could be regarded as having independent traits from examination of the Cramer's V statistics.

Figure 11: **Frequency of Traits and Scatter Plot of Trait Number vs Proportion of Violations.** The proportion of violations counts the number of rejections of the chi-squared test for independence in the contingency table and divides this by the total number of distinct trait pairs.

(a) Frequencies of Trait Numbers



(b) Trait Number vs Proportion of Violations



Bar and QR Codes for 196 Collections

Finally we show the bar and arithmetic-harmonic QR codes that were introduced in Section 4.2 for the 196 collections with 8,888 or more PFPs. The plots are presented in alphabetical order of the collection's name. Each collection is identified by its name and below this we report the result of a Kolmagorov-Smirmoff (KS) test for equality of score distributions, followed by its p-value in parentheses. Below this we depict the bar code and below that the arithmetic-harmonic QR code, with a red line at 45°. Other QR plots are possible but we select the arithmetic-harmonic one because the divergence between these two scores is greater than the divergence between any other pair of Pythagorean

[41] We exclude five collections with more than 30 traits: VOX Series 1 (68 traits), Afro Droids (32 traits), Crypto-Pills by Micha Klein (42 traits), Animetas (48 traits) and MonsterBlocks (166 traits).

means, so the arithmetic-harmonic QR code should be more dispersed. Given the number of collections we have displayed here we confine ourselves to some general comments. This way, we aim to guide readers how to interpret the characteristic of any of these collections, simply by regarding these codes.

Some collections have a single horizontal line in their bar code, albeit of varying width. This includes the first four: Al Cabones, Avastar, Adam Bomb Squad and Afro Droids. Notice that the QR codes of these collections are also highly dispersed, *i.e.* the ranks derived from the two rarity metrics are very different for most of the tokens. Only a few lie on the 45° line. What does this imply? In the current environment the ranking of rarity is highly non-standardized. There are numerous NFT marketplaces but no consistent, universally-accepted ranking methodology. One marketplace could list a high offer price based on its assessment that the PFP is quite rare, whereas another market place could rank the PFP as very common and therefore list a much lower offer price. This implies that it will be hard to find a clear relationship between price and rarity for such collections. Roughly one-third of the collections shown here fall into a similar category.

However, a single line in the bar code does not imply that the QR code is highly dispersed. For example, several collections have QR plots where all points lie close to the 45° line, meaning that there will be less divergence between the ranks assigned by the various metric used, so that a closer agreement between prices listed on different marketplaces should be possible. This in turn should bring more clarity to the price-rarity relations for such collections. Notable examples in this category (with varying degrees of closeness) include Cool Cats, CryptoGhost, Pluto, Panda Dynasty, Starchain Official and The Wonder Quest. CryptoGhost is very interesting because its bar code is a single line at zero, meaning that the rarity score for every PFP is independent of the metric. So the QR code has clusters with equal rarity ranks (it does not look this way, because we have not increased the size of the point when multiple tokens have the same coordinates). Note that Cool Cats and Pluto both have bar codes with a horizontal line near 1, indicating maximum potential for divergence between different rarity scores. Nevertheless, the clustering of the QR code around the 45° line indicates that the harmonic and arithmetic mean rarity rank densities must be very close.