Do NFTs' Owners Really Possess their Assets? A First Look at the NFT-to-Asset Connection Fragility

Ziwei Wang
12250053@mail.sustech.edu.cn
Southern University of Science and Technology & University of Birmingham
Shenzhen, Guangdong, China

Jiashi Gao
12131101@mail.sustech.edu.cn
Southern University of Science and Technology
Shenzhen, Guangdong, China

Xuetao Wei*
weixt@sustech.edu.cn
Southern University of Science and Technology
Shenzhen, Guangdong, China

ABSTRACT

Most NFTs (Non-Fungible Tokens) use multi-hop URLs to address the off-chain assets due to the costly on-chain storage, but the path from NFTs to the underlying assets is fraught with instability, which may degrade its value. Hence, this paper aims to answer the question: Is the NFT-to-Asset connection fragile? This paper makes a first step towards this end by characterizing NFT-to-Asset connections of 12,353 Ethereum NFT Contracts (6,234,141 NFTs in total) from three perspectives, storage, accessibility, and duplication. In order to overcome challenges of affecting the measurement accuracy, *e.g.*, IPFS instability and the changing availability of both IPFS and servers' data, we propose to leverage multiple gateways to enlarge the data coverage and extend a longer measurement period with non-trivial efforts. Results of our extensive study show that such connection is very fragile in practice. The loss, unavailability, or duplication of off-chain assets could render the value of NFTs worthless. For instance, we find that assets of 25.24% of Ethereum NFT contracts are not accessible, and 21.48% of Ethereum NFT contracts include duplicated assets. Our work sheds light on the fragility along the NFT-to-Asset connection, which could help the NFT community to better enhance the trust of off-chain assets.

CCS CONCEPTS

• **General and reference** → **Measurement**; • **Security and privacy** → **Web application security**.

KEYWORDS

Blockchain, NFTs, Fragility, Trust, Characterization

ACM Reference Format:
Wang, et al. 2023. Do NFTs' Owners Really Possess their Assets? A First Look at the NFT-to-Asset Connection Fragility. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583281

*Xuetao Wei is the corresponding author.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583281

1 INTRODUCTION

The year of 2021 is remarkable for NFTs since the market of NFTs has been expanded to $7B. Record-breaking prices have been witnessed, such as the "Everydays: The First 5,000 Days" ($69.3M) [48] and "The Merge" ($91.8M) [37]. The value of an NFT largely comes from the digital asset it represents, as well as the unique and identifiable ownership of such asset backed by blockchain.

Due to the costly storage on the blockchain (on-chain), most NFTs save their assets off the blockchain (off-chain). Centralized storage, *i.e.*, storing in NFT providers' own servers or cloud services, and decentralized storage, *e.g.*, *InterPlanetary File System* (IPFS) [13], Filecoin [10], Arweaves [3], is used to host these assets. Decentralized storage is generally considered with better sustainability and tamper-resistance as data is not controlled by a centralized entity. In practice, there are roughly two ways to bind the on-chain NFT with the off-chain asset: ❶ the metadata field in NFTs records multi-stage storage addresses to point to the off-chain asset, which refers to the NFT-to-Asset connection in this study; ❷ a hash value or the Merkle root of a group of data is recorded in the metadata of NFTs. Note that in this work, we focus on the asset path of NFTs and won't discuss the effect of hash proof on the asset safety since the NFTs' hashes are unable to be retrieved with standardized `ABI`. It will be a future work to combine cryptography with the NFT to enhance NFT assets' safety.

Previous work has investigated NFTs on price forecasting [62, 66], integrity [56, 74], plagiarism [59, 69], wash trading [73], and anonymity [52, 70]. However, the study of the NFT-to-Asset connection that decides the value of NFTs, is still unaware, which hinders us from understanding whether NFTs' assets are stored safely in practice and trusting the value of these NFTs. Therefore, we conduct a systematic characterization and analysis of the NFT-to-Asset connection fragility of NFTs on Ethereum. To be more specific, our study attempts to answer the following research questions:

- **RQ1** Where are the NFTs' off-chain assets? Are the NFTs' assets in peril of aggregation?
- **RQ2** What is the degree that NFTs are connected to the off-chain assets? At which stage do NFTs lost connection with the assets?
- **RQ3** To what extent are NFT assets duplicate? At which step does the repetition occur? Which sort of NFTs are highly repetitive?

We initiate our study with non-trivial efforts by collecting NFT contracts based on Etherscan's "*NFT*" contract tag, and subsequently retrieve the data from on-chain *TokenURI* to off-chain *Metadata* and *Asset*. We have to overcome several challenges to improve

the measurement accuracy, *e.g.*, IPFS instability, and the changing availability of both IPFS and servers' data. In our measurement framework, we propose to leverage multiple gateways to enlarge the data coverage and extend a longer measurement period to cover as much of the time as possible when servers went online (refer to Appendix A.2). Our measurement framework focuses on the angles of storage (refer to Section 4), accessibility (refer to Section 5) and the content (refer to Section 6), aiming to uncover the sustainability, the availability and redundancy of NFT. Through this multi-perspective profiling, we present a thorough and comprehensive understanding of the NFT-to-Asset connection, where the interesting and valuable findings reveal the present stage of NFTs in the wild is concerned.

The connection between NFTs and their assets is actually fragile, bringing at least but not limited to the following concerns: ❶ **Lost and inaccessible** (refer to Section 4.1.1, Section 4.1.2, Section 5.1). The value of NFTs heavily depends on the safety of the underlying asset. Under the existing off-chain settings, owning NFTs doesn't really possess the assets. We find that assets of 3,118 (25.24%) NFT contracts are not accessible. Whenever the actual asset is lost, ownership of such NFTs becomes worthless. Even with the more recognized decentralized storage, infrequently accessed data is vulnerable to loss or inaccessibility. ❷ **Counterfeit** (refer to Section 6.1, Section 6.2). As the transparent nature of the public blockchain, assets of NFTs can be accessed by anyone, which can be easily used to generate new NFTs by others. Such hypothesis is proved by our study. We find that 2,653 (21.48%) NFT contracts include duplicated assets. ❸ **Unclear custodian responsibility** (refer to Section 4.2). Once smart contracts are deployed on blockchain, NFTs' owners are not necessarily responsible for maintaining the assets by themselves. We observe that 4,801,699 (77.02%) NFTs of centralized storage are aggregated on merely 1,184 third-party storage sites. The loss of these data renders unclear responsibility among buyers, sellers, and third-party storage platforms.

In summary, our contributions are as follows:

- To the best of our knowledge, our work is the first to present a systematic characterization of the NFT-to-Asset connection fragility between on-chain NFTs and the associated off-chain assets from three perspectives: storage, accessibility, and duplication.
- In order to overcome challenges of affecting the measurement accuracy, *e.g.*, IPFS instability and the changing availability of both IPFS and servers' data, we propose to improve the measurement accuracy by leveraging multiple gateways to enlarge the data coverage and extend a longer measurement period with non-trivial efforts,.
- Our study provides empirical evidence to suggest that the current NFT-to-Asset connection is very fragile, which has non-trivial effects on NFT safety and trust. Key interesting and unexpected findings are listed as follows:
  - **Off-chain assets are aggregated** (response to **RQ1**). We uncover evidence that 10 centralized platforms account for hosting 79.04% of NFTs' off-chain assets, showing that the NFT off-chain assets are indeed centralized.
  - **Instability of decentralized storage** (response to **RQ1**). We show that decentralized storage has not yet achieved its goal of providing everlasting data storage in practice as we find that only 33.77% of NFTs' decentralized assets can be fully retrieved.
  - **The 2-stage storage state are inconsistent** (response to **RQ2**). We identify a deceptive practice that though 48 NFT collections appear to be decentralized (according to their on-chain URLs), their underlying assets are actually centralized.
  - **Duplicated NFTs** (response to **RQ3**). We find that NFT assets are severely duplicated. Of the NFT contracts that contain duplicated assets, 36.82% of such have duplication rate greater than 95% regarding to assets' URL.
- Our observations shed light on the fragility along the NFT-to-Asset connection, which brings important implications to the NFT community: ❶ NFT holders or prospective investors need to pay closer attention to the risks associated with the NFT-to-Asset connection and the uniqueness of their assets. ❷ NFT developers can enhance the strength of the NFT-to-Asset connection by integrating the NFT with other techniques, such as storage incentive, digital rights management, and binding mechanism, to increase the sustainability, traceability, and uniqueness of the NFT. ❸ For the NFT applications, a more precise division of storage responsibilities and usage rights with users is needed to improve the NFT's credibility.

2 BACKGROUND

2.1 Non-Fungible Token

2.1.1 Token Standard. NFTs are driven by standard smart contracts that conform to specific interfaces, *e.g.*, *Non-Fungible Token Standard* `ERC-721` [7], *Multi Token Standard* `ERC-1155` [4], *Composable Non-Fungible Token Standard* `ERC-998` [8]. Under the same token standard, NFTs have identical functions to achieve compatibility. Descriptive information about NFTs is usually recorded in the *metadata* field of NFT, including function *name()*, *symbol()*, and *tokenURI()*. The return value of *TokenURI* is usually an external URL pointed to a `JSON` table that is saved in another repository. Another small proportion of *TokenURI* will return the metadata or the asset directly. Other NFT protocols improve NFTs' compatibility and interoperability, such as the *Semi-Fungible Token Standard* `ERC-3525` [6] and the *NFT Royalty Standard* `ERC-2981` [5].

2.1.2 NFT-to-Asset Connection. As stated in the introduction, NFTs use a multi-hop URL to link to the asset. A typical NFT storage structure is that the *TokenURI* is stored on the blockchain, which points to the *Metadata* corresponding to the NFT. *Metadata* is usually a `JSON` table with a certain schema, including title, type, image link, properties, etc. The *Metadata* contains the *AssetURI* to the final storage location of the underlying digital asset. Not all NFTs conform to this storage structure. For example, some NFTs save their assets directly on-chain; Others may omit the step of *TokenURI* and save the *Metadata* on the blockchain; Also, some NFTs may have no *Metadata*, but their *TokenURI*s are pointed to the off-chain assets. Saving data on-chain is undoubtedly an optimum yet costly solution, as data will be verified and permanently hosted by nodes included in the blockchain network.

2.2 Data Storage

2.2.1 Centralized Storage. Centralized storage refers to the storage on NFT marketplaces (*e.g.*, Opensea [40], Nifty Gateway [39], Rarible [43]), NFT applications, cloud services (*e.g.*, Google Cloud [30], Amazon Web Services [20]). Hosting on such servers, the file size is almost unlimited and the data response becomes faster, but the data is highly centralized and the NFT owner has no actual control over it. Once the servers stop responding or are attacked, the lovely images of NFTs will disappear. The data also faces the risk of being tampered with, and there is no corresponding measure to ensure the rights of NFT holders. For example, the NFT originator CryptoKitties [27] is centralized. Kitties on-chain is merely a series of numbers, and their corresponding identities and properties are on the company's servers, which are isolated from its on-chain ownership.

2.2.2 Decentralized Storage. In decentralized storage, data are replicated among multiple nodes and the data will never disappear as long as there is a node in the world that stores this file. Decentralized storage uses *Content Identifier* (`CID`) to address the file, which is the hash value of the file that cannot be tampered with. Filecoin and Arweave are both the incentive layer based on IPFS, aiding with blockchain technology, so as to motivate miners to actively keep data correctly and long-lastingly.

Although decentralized storage has made great progress in tamper resistance and data persistence, there are still some risks in this method. ❶ *The infrequently accessed data may also be lost as the earlier file will be deleted when the node reaches its upper limit of nodes' storage capacity*. For this reason, there are now IPFS storage custodial services like Pinata [42], which help IPFS data to be hosted forever. ❷ Even if the data exists in IPFS, *it can also be massively aggregated in the IPFS nodes of some NFT applications*. Since these NFTs are often obtained directly through their gateways, these data will be rarely backed up in other IPFS nodes. ❸ *The response time of IPFS can be extremely laggy and the possibility of successful retrieval is not guaranteed* (a brief demonstration is given in Appendix A.2.1) since the node does not necessarily store the requested file, nor does it necessarily be online.

2.3 Asset Consistency

2.3.1 Hash value. Saving the cryptographic hash of the data on the blockchain is the most widely adopted approach to preserving data integrity. The asset consistency can be verified easily by comparing the asset hash to the on-chain record hash. `CID` of decentralized storage itself is the hash of the data. Once data is published, it cannot be modified. Data can be updated only by publishing another modified data, and the `CID` becomes different. Another method is to store the data hash in groups to save space, such as using the Merkle root hash of a batch of data. For example, the digital art collection Hashmasks [11] generates hash provenance by hashing the concatenating images, and the image's URL is not saved on the blockchain.

2.3.2 On-chain Storage. To be more straightforward, data on the blockchain live permanently and will not be tampered with, which is an optimal but costly solution for NFT data storage. For example, Akomba Commemorative Token (`AkCT`) [19] are quotes from crypto-celebrities; ColorverseFounder (`CVF`) [26] and Mandala Tokens (`MANDALA`) [36] are the `SVG`s of the mosaic of color blocks. A more applicable and gas-saving solution is to deploy a `SVG` generator on-chain, which can generate the image from the NFT properties directly, such as the top decentralized exchange Uniswap V3 (`UNI-V3-POS`) [17], and DeFi-NFT Aavegotchi (`GOTCHI`) [2].

3 DATASET AND EXPERIMENTAL SETUP

3.1 Data Collection

According to the NFT-to-Asset structure, data collection is divided into the following sub-processes, and **several factors affecting measurement accuracy are discussed in Appendix A.2**. Overall, we overcome these challenges to obtain comprehensive data with non-trivial efforts.

NFT Contract. In the beginning, we obtain 12,353 NFT contract addresses from Etherscan [9], where NFT contracts are tagged [15]. NFTs in a same contract are also called an NFT *collection*.

The *TokenURI*. The *TokenURI* data is collected from the Ethereum Mainnet and Opensea. We make our endeavor to ensure that the data comes from the Ethereum Mainnet for the most part, as the *TokenURI* provided by other platforms isn't always consistent with the one in Ethereum. For example, the Waifus NFTs have no publicly available *TokenURI* according to its contract, but it is given by OpenSea that the *TokenURI* is *https://api.waifusion.sexy/v1/opensea/i*, which conflicts the record on the blockchain. We build a full archive OpenEthereum node of Ethereum mainnet for synchronization, which enables full contract state information[16]. The workflow starts with requesting the total supply of NFTs in an NFT contract (function in `ABI`: *totalSupply()*). If the contract contains NFTs, then we sequentially request the *TokenURI* and its ownership with token id (functions in `ABI`: *tokenByIndex()*, *tokenURI()*, *ownerOf()*).

However, a portion of *TokenURI*s are difficult to be obtain from Ethereum for the following reasons: ❶ the NFT contract does not always conform to the standard form, *i.e.*, the `ERC-721`, `ERC-1155`, and some variants; ❷ contracts are not open-source, which is not available on Etherscan; ❸ the *TokenURI* related function can not be called from outside the contract, which prevents us from invoking the contract function to collect the information. Therefore, for contracts that are not accessible from Ethereum, we obtain it with OpenSea "*asset*" API [41], and extract the *TokenURI* from *token_metadata* field. But unfortunately, OpenSea's API has a maximum access limit, which is controlled by parameters "*offset*" and "*limit*". That is, only the first 10,050 NFTs can be obtained. 23 contracts exceed the upper limit of OpenSea during this process, thus we try to collect from these contracts manually. Among them, *TokenURI* of 10 contracts are successfully retrieved, and 6 of them can be obtained but the value is empty, hence the full NFT dataset of these contracts is obtained. Besides, 7 of them are unable to get due to the non-public contracts' codes.

The *Metadata*. Metadata is obtained by requesting the *TokenURI*. For location-based URL (in HTTP style), we request the data ordinarily. For content-based URL (IPFS and Arweave's `CID`), we request the data from hybrid sources: a locally build IPFS node, and the public gateways (`ipfs.io` [35], `gateway.pinata.cloud`

[33], `infura-ipfs.io` [34] and `dweb.link` [14]), considering the response instability of *IPFS* (refer to Appendix A.2.1).

The *Asset*. A standard `ERC-721` metadata format contains 2 URLs, one is in the *image* field and the other is in the *external_url* field [7]. Both the URLs are usually in different locations, such as an *image* of the IPFS-style URL and an *external_url* of the location-based URL for backup. We thus extract the *AssetURI* automatically from the following fields: *image*, *external_url*, *url*, and *animation_url*. Then we manually collect URLs from the irregular metadata. The following process is the same as that in the *Metadata* part, which is to collect content-based URLs and location-based URLs separately to obtain the asset.

3.2 Overall Statistics

A detailed data source is given in Appendix A.1. In summary, we have 6,234,141 NFTs according to 12,353 contracts. 10,916 (88.37%) contracts information are collected via Ethereum Mainnet directly and 1,437 (11.63%) contracts information are collected from Opensea. We find that 3,123 (25.28%) contracts are empty, *i.e.*, a zero *totalSupply*, and hence they will not be measured in the following. Table 1 provides the data collection from *TokenURI* to the *Asset* in general. Ideally, each step of the NFT-to-Asset path should be injective, that is the number of *TokenURI*s, *Metadata*, *AssetURI*s and *Asset*s should be equal, ignoring the very few NFTs which are exceptional to this normal path. However, we find that *the decrease of the un-collected data in the next stages is greater than the number of those exceptional NFTs. The number of assets collected at the end of the path is only 45.75% of that of NFT, indicating that the issue of ineffective URL is more severe than expected.* A detailed analysis of NFT data loss is presented in Section 5.

	Raw Count	Data Size	Hit Rate
Contract	12,353	\	\
NFT	6,234,141	\	\
TokenURL	5,846,287	1.53 GB	93.78%
Metadata	4,566,766	119.88 GB	78.11%
AssetURL	3,992,859	413.54 MB	87.43%
Asset	2,851,894	3.29 TB	71.42%

Table 1: Summary of collected NFTs and the associated data.

4 THE STORAGE

A typical storage scheme of NFT associated data can be divided into 2 stages: the *Metadata* stage and the *Asset* stage. We investigate the proportions of NFTs storage on centralized servers and decentralized storage networks, then conduct a progressively deeper investigation towards the exact server clusters where NFTs' associated data is hosted. Next, we take the 2-stage storage as a whole to investigate the stability and consistency of the NFT-to-Asset connection.

4.1 The 2-stage storage

To ascertain whether the NFT storage is sustainable, we first look at the location distribution in the two storage phases and provide the correlation between the two storage layers. Then, we conduct a detailed study of the scenario in which NFTs for centralized storage are constantly clustered on a few large platforms.

4.1.1 Do NFTs prefer decentralized storage? In the 1st-stage of storage, 5,862 contracts (4,802,766 NFTs) adopt the centralized storage, and 2,385 contracts (510,154 NFTs) use the decentralized storage. Therefore, the average number of NFTs in each contract of decentralized storage is lower. In the second phase of storage, assets are usually stored in different locations for backup to reduce the risk of asset loss. The number of contracts (NFTs) for centralized storage decreases obviously by 28.14% (7.21%) due to the direct storage of assets and the unavailability of URLs. Though the number of contracts for decentralized storage decreases by 4.65%, the number of NFTs increases by 12.33%.

4.1.2 Connection of 2-stage storage. We examine the relationship between the two stages of NFT storage, which is the direction of the 2nd-stage storage following specific 1st-stage storage, in order to determine the stability and consistency of the two-stage storage. We narrow the measurement scope to NFTs that successfully connect to the underlying assets, meaning that both *TokenURI* and *AssetURI* exist and the assets are able to be retrieved. NFTs with on-chain assets and NFTs that fail to connect to their assets are ignored since these assets have either been permanently stored in the blockchain or have been lost. Besides, some NFTs may exist one or more different URLs in *TokenURI* and *AssetURI* to guarantee asset safety. The NFT-to-Asset connection is divided into three categories according to the storage type and the consistency, which are *Decentralized*, *Semi-decentralized*, *Centralized*.

The result is presented in Table 2, with taking multi-URLs of *AssetURI* into account. Among the 6,619 NFT contracts that are included in the measurement scope, there are 2,034 (30.73%), 444 (6.71%), and 4,141 (62.56%) contracts in the *Decentralized*, *Semi-decentralized*, and *Centralized* categories, respectively. 68.59% of the NFT collections use more than one platform for asset back-up, and the major portion of NFTs without assets back-up are stored on the server. *However, we discover that there exists a deceptive practice, where 48 NFT collections that have decentralized metadata but centralized assets. In this way, the safety of the assets is not assured even though the metadata of NFTs is permanently preserved and tamper-resistant.*

4.2 NFTs are clustering on centralized servers

A large proportion of NFTs choose to keep their data on third-party storage platforms instead of maintaining a dedicated server themselves: ❶ NFT marketplaces, which allow NFT developers and independent collectors to publish NFT works on the blockchain through the exchange market's own channels, while providing relevant data storage services; ❷ The cloud service providers, where users first publish their NFT works on to it and then upload the given URLs to the blockchain; ❷ The server of the application provider, where the data is maintained and managed by the application itself. *In this way, NFT off-chain content is further clustered on a few servers, which goes against the primal decentralized intention of blockchain.*

Therefore, we intend to assess the extent to which NFTs are concentrated on a few servers. We examine the NFTs' hosting location by the domain from HTTP-based *TokenURI*, and the domains

Decentralized						Semi-decentralized						Centralized		
Metadata	***Asset***	**#**	***Metadata***	***Asset***	**#**	***Metadata***	***Asset***	**#**	***Metadata***	***Asset***	**#**	***Metadata***	***Asset***	**#**
AR	AR	48	IPFS	AR	1	IPFS	SE	**20**	SE	AS+SE	1	SE	SE	1,851
AR	AS	1	IPFS	AS	40	IPFS	SE+SE	**24**	SE	IPFS+SE	95	SE	SE+SE	2,284
AR	AR+AR	1	IPFS	IPFS	469	OC	SE	**4**	SE	IPFS+SE+SE	2	SE	SE+SE+SE	6
AR	AR+SE	271	IPFS	IPFS+AS	3	SE	AR	3						
OC	IPFS	4	IPFS	IPFS+IPFS	14	SE	AS	108						
OC	AR+AR	3	IPFS	AR+SE	3	SE	IPFS	134						
OC	AR+IPFS	13	IPFS	IPFS+SE	1,159	SE	IPFS+IPFS	20						
OC	IPFS+SE	1	IPFS	IPFS+SE+SE	3	SE	AR+SE	33						
Sum					2,034	**Sum**					444	**Sum**		4,141

Table 2: Distribution of NFT-to-Asset paths. The storage types are in line with Section 4.1.1: Server(SE), Arweave(AR), IPFS(IPFS), Asset(AS), On-chain(OC).

with the highest concentration of contracts are presented in Table 3. Of 5,858 contracts (4,801,699 NFTs) that adopt centralized storage, there are 1,184 different storage sites. *The top-10 domains hold 65.67% (79.04%) of the total number of contracts (NFTs).* The NFT marketplaces are oligarchical with Opensea [40], NiftyGateway [39], and Mintable [38] considering the number of collections (in proportions of 17.94%, 16.38%, 15.47%). While further examining the NFT storage on these marketplaces, we observe that the Opensea NFTs' metadata is on its own server, and the assets are hosted on GoogleUserContent (97.87%) [30] and Opensea (2.13%). Nevertheless, the NFTs' metadata of 158 (10.28%) contracts on Opensea fail to retrieve. The NiftyGateway target assets are saved on Cloudinary (99.47%) [25] and AmazonAWS (0.53%) [20], and we observe that the asset acquisition result of 25 (2.60%) NiftyGateway contracts (1795 NFTs) is *not found*. Compared to the above marketplaces, Mintable's asset layer has no uniform schemes, and the NFTs' assets are directly hosted on the metadata-level URL's location in a one-hop way, or are on various storage such as IPFS, Arweave, CloudFront [24], Imgur [31], AmazonAWS, etc. The NFTs in 116 (12.8%) contracts are not available, which is the highest loss rate among the three marketplaces.

When measured by the number of NFTs, the aggregation of NFTs is slightly more scattered than the measured by contracts, which is reflected in the smaller aggregation of the NFTs of the top-3 domains (in proportions of 12.77%, 10.34%, 10.27%) and the larger aggregation of the other domains. However, the blockchain game Gods Unchained [29] ranked 1st with 613,328 NFTs, though operating normally, *TokenURI* recorded on Ethereum cannot access the data in the next level without exception. Another highly ranked but risky NFT project is Cybertopia [28]. We obtain 33,413 (13.3%) metadata from a total of 251,001 NFTs. The collected metadata contain no valid information that can represent the identity of NFT, but only duplicate data.

5 THE ACCESSIBILITY

Given that the NFT-to-Asset path may include several layers of connectivity and storage platforms, the failure of the NFT value will be caused directly by the inaccessibility of data at any point. In this section, we evaluate the availability of NFT assets by analyzing the completeness of data collection.

Contract Aggregation				
***Metadata* Location**	**# of Contract**	**(%)**	**# of NFTs**	**(%)**
api.opensea.io	1,051	17.94	20,115	0.42
api.niftygateway.com	960	16.38	160,800	3.34
metadata.mintable.app	906	15.47	25,310	0.53
coludfunctions.net	529	9.03	41,142	0.86
locksmith.unlock-protocol.com	143	2.44	2,740	0.05
herokuapp.com	105	1.79	72,336	1.51
api2.cargo.build	99	1.69	110,947	2.31
azurewebsites.net	63	1.08	48,279	1.01
amazonaws.com	61	1.04	496,512	10.34
factory.chocomint.app	60	1.02	747	0.01
Total # of Contract				5,858

Table 3: Aggregation of contracts and NFTs on the centralized servers.

5.1 Accessibility at each stage

The NFT-to-Asset connection is made up of four interlocking parts: *TokenURI*, *Metadata*, *AssetURI*, and *Asset*. The NFT data lost of the above four parts in the NFT-to-Asset path can thereby be represented by the validity rate of three transition stages, which we denote as: ❶ *TokenURI* → *Metadata*; ❷ *Metadata* → *AssetURI*; ❸ *AssetURI* → *Asset*. The access rate is calculated by the proportion of data that is successfully retrieved from its predecessor. For example, the NFT access rate of stage ❶ refers to the number of collected metadata divided by the number of *TokenURI*s.

In general, of the 9,235 NFT collections with *TokenURI* retrieved from Ethereum, the number of NFT collections which is not fully retrieved is almost doubling between stages, which is 1,421 (15.39%) of ❶, 2,594 (28.09%) of ❷, and 4,004 (43.36%) of ❸, respectively. As presented in Figure 1, the distribution of access rate in each stage exhibits different characteristics. First, the access rate of each stage is mainly centered around 1.0, indicating that *most NFTs remain valid*. The number of fully valid NFTs is 7,814 (84.61%) of ❶, 6,641 (71.91%) of ❷, and 5,231 (56.64%) of ❸, respectively, while it shows *an accelerated decline trend towards the end of the connection*. Secondly, per Figure 1 the distribution of ❶ shows a tendency of more at both ends and less in the middle, which indicates that the access rate of NFTs is centered near 1.0 and 0.0, *i.e.*, *most NFTs are either completely available or completely unavailable*. The unavailability

of *TokenURI* may be due to the fact that the data obtained from *TokenURI* is not itself in the form of a URL (see Section 4.1.1). This may also indicate that the publisher has shut down the service, since metadata is typically held by the publisher for NFTs stored on the Server. The access rate from *AssetURI* to *Asset* is more diffuse, with a higher percentage of NFTs in the middle interval of the figure. Such changes indicate that asset storage is much more unstable. This instability is not due to the human factor such as a service close, but rather because a large portion of assets is hosted on fragile storage platforms with poor data persistence, which is of vital importance for NFT safety. The next section explains further on causes of data storage and retrieval failures.

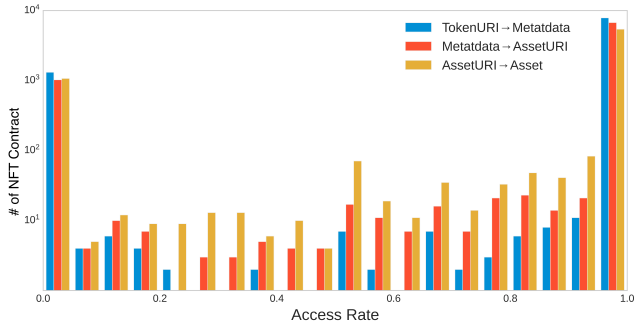

Figure 1: The distribution of access rate. *TokenURI → Metadata*, *Metadata → AssetURI*, *AssetURI → Asset* refer to ❶-❸ in Section 5.1.

5.2 What causes data loss?

Two perspectives are used to analyze the NFT lost data in further detail: the causes of the NFT loss and the features of these missing data. To verify the factor that leads to the NFT loss, we divide the access rate of *TokenURI* towards *Asset* (denoted by α) into three intervals and compared the influence of the factors on the NFT data loss. The results are shown in Table 4.

(1) **Storage Platform.** Per Table 4, the percentage of completely lost NFTs (α=0), among centralized storage and decentralized storage is 53.16% *vs.* 46.84%, which shows no significant difference between storage types. For NFTs with incomplete requests (0<α<1) and full acquisition (α=1), the ratio of centralized storage to decentralized storage increases strikingly, about three times when 0<α<1, and twice when α=1. Meanwhile, as described in Section 4.1.1, considering that the NFTs' ratio of choosing centralized storage (5,862, 71.08%) to decentralized storage (2,385, 28.92%) is about 2.45, this is roughly consistent with the ratio of 0<α<1 by twice and three times. Therefore, the proportion of the centralized NFTs with α=0 is actually below this ratio, which indicates that *the assets of the centralized storage currently have superior persistence and accessibility to some extent.*

(2) **Asset Size.** There is no asset size for NFTs with a 0 access rate. NFTs with 0<α<1 are characterized by a large number of NFTs within contracts, nearly 1.77 M NFTs in 843 contracts, which poses challenges to maintaining asset integrity. However, the average size of these assets is only 1.06 MB, which is slightly smaller than the NFT assets with α=1. We thereby conduct a two-sample t-test on the average asset size of the set of 0<α<1 and α=1 to test whether the means of the two independent samples are significantly different. The test is two-sided and the null hypothesis H_0 assumes that the means of the two samples are the same. Results show that the statistic t=-2.846 and the *p-value*=0.004. As the *p-value* is smaller than 0.025, the null hypothesis can be rejected with 95% confidence, and there exist significant differences between the average asset size. Therefore, *the average file size of NFTs with α=1 is indeed larger than that of 0<α<1.*

(3) **URL Format.** The *AssetURI* formats of completely lost NFTs are presented. For NFTs with α=0, 2,653 (78.44%) of asset access failures are caused by illegal URLs, and the primary cause for this failure is that the return value of these URLs is null. The influence of invalid URLs is tiny for NFTs of $\alpha \neq 1$, and only a few of these NFTs are unreachable due to empty URLs.

6 THE ASSETS

In this section, we study the NFT plagiarism from the perspective of NFT duplication, where the assessment target composes of the repeated URLs and the repeated underlying asset, aiming to locate the exact phase where duplication massively occurs.

6.1 Duplicated URL

We observe that a large number of NFT collections have encountered the problem of high duplication of corresponding URLs. To understand the extent to which NFTs suffer from non-uniqueness, we make an assessment of the above situations respectively. Considering that a large proportion of NFTs adopt two-stage storage, such repetitiveness can thus be generalized into 4 fundamental situations, which are presented in Figure 2. The duplicated rate of the NFT corresponding URLs is calculated as the number of redundant URLs divided by the total number of URLs. For instance, suppose a contract has 10 NFTs and 4 unique *TokenURI*s, indicating that 6 of the *TokenURI*s are redundant. The duplicated rate is thus 60%.

*6.1.1 Type **(1)**.* In this case, we observe that 1,756 (23.50%) NFT contracts share *TokenURI* to varying degrees. *The repetition rate of NFTs with repeated URLs is mainly centered toward 100%.* As can be seen from Figure 3, the number of NFTs is increasingly dense when the repetition rate is higher than 50%. Of these, *527 (30.01%) contracts have a repetition rate higher than 95%.*

*6.1.2 Type **(2)**.* The calculation of duplicated *AssetURI* may be impacted by the following reasons: ❶ The *AssetURI* usually includes multiple fields for recording URLs, rendering different duplicated rates of each field of URLs (refer to Section 3.1).; ❷ String of metadata URL field is the URL base shared between NFTs within an NFT collection, while the *AssetURI* is already included in other fields, resulting in an overestimated duplicated rate. To address the above problems, the duplicated rates of each field of URLs are calculated separately, and the duplicated rate is determined by the lowest value. Results show that the *AssetURI*s are *severely* duplicated. *There exists 2,417 (37.95%) NFT collections that have duplicated AssetURI*

Interval	# of Contract	# of NFT	Storage Platform		File Size		URL Format		
			IPFS + AR	Server	Overall	Average	Empty	Non-URL	Asset
$\alpha = 0$	3,382	\	226 + 18 (53.16%)	215 (46.84%)	\	\	2,746	28	149
$0 < \alpha < 1$	843	1,773,287	190 + 27 (25.93%)	620 (74.07%)	1.80 TB	1.06 MB	5	0	0
$\alpha = 1$	5,010	1,130,648	1,360 + 331 (33.77%)	3,317 (66.23%)	1.50 TB	1.39 MB	2	0	0

Table 4: The influence of storage platform, asset volume, and URL format on the NFT preservation. The arrival rate of *TokenURI* towards *Assets* is represented by α.

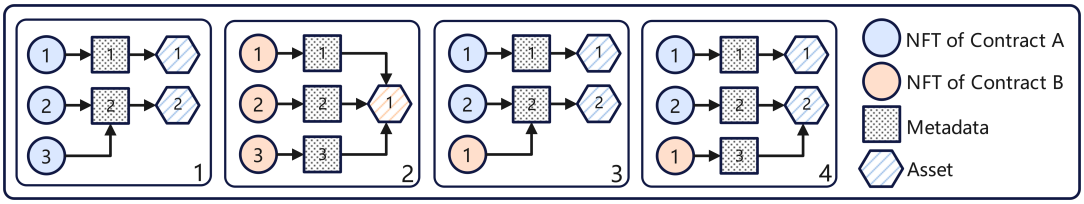

Figure 2: Types of URL confusion. ❶ NFTs within a same contract have identical *TokenURI*; ❷ NFTs within a contract have identical *AssetURI*; ❸ NFTs of different contracts have identical *TokenURI*; ❹ NFTs of different contracts have identical *AssetURI*.

NFTs, and 890 (36.82%) redundant NFT collections have duplicated rate larger than 95%. Per Figure 3 the number of NFT collections with repeated *AssetURI* increases significantly compared to *TokenURI*. Furthermore, the 661 newly included duplicated NFT collections almost have duplicated rate larger than 75%. *This phenomenon indicates a decoupling between NFT's on-chain metadata and off-chain assets, as a large number of different TokenURIs refer to identical AssetURI.*

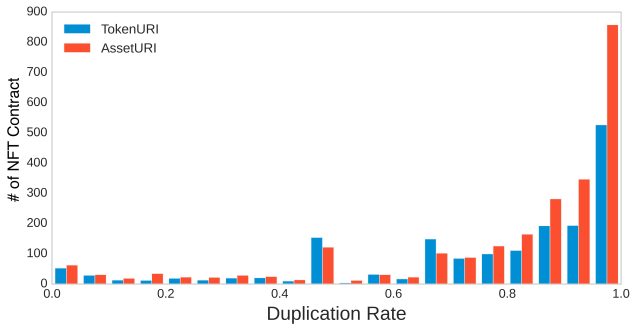

Figure 3: The distribution of duplicated rate of NFTs' *TokenURI* (Section 6.1.1) and *AssetURI*(Section 6.1.2), excluding cases where URIs are not duplicated.

*6.1.3 Type **(3)**.* We detect 26,983 *TokenURI*s which are shared among 528 contracts, and the most widely shared URL is used by 15 contracts. We then count the addresses that occurred most frequently in all shared contract groups, which is the number of other contracts that has *TokenURI* overlapped with it, aiming to observe the association of shared addresses between contracts. We find that the most involved NFT has *TokenURI* overlapped with 17 other NFTs. And common characteristics of the shared-URL contracts are that most of them have the same contract name, and it probably includes counterfeit contracts, or the contracts are repeatedly deployed for project testing or release updates. It suggests that *the TokenURI duplication is more of a duplication of invalid information and multiple contracts belonging to the same project.*

*6.1.4 Type **(4)**.* We observe 23,340 (0.64%) URLs are shared by NFTs from different contracts out of 3,652,097 *AssetURI*s, for a total number of 491 such contracts. Meanwhile, we generate the set of NFT contracts for each *AssetURI*, and observe 788 different contract sets. The number of shared URLs of each contract set ranges from 1 to 10,000, however, 98.10% of contract sets have less than 100 shared URLs. As with Type **(3)**, the addresses with the highest number of shared URLs tend to have similar token names. For example, the contract set with 10,000 shared URLs is made up of 4 addresses named Su Squares (SU) [47], and the contract set that shares 4,606 URLs is made up of 3 addresses named 89 seconds Atomized (SNP001) [18]. The content similarity between the contracts is not high. The most severe situation is that one contract shared the URLs with 15 other contracts, while 89.21% of the contracts are shared with only one other contract. That is, *the vast majority of AssetURI cross-contract sharing is taking place on a tiny scale.*

6.2 Identical Underlying Content

Assets obtained from different URLs may still probably identical. Analysis of duplicate URLs alone is insufficient to illustrate the degree of the duplication of assets. The identical assets are detected using the hash function at first (*e.g.*, `MD5`, `SHA256`). Whereas these hash functions yield different results against such minor change, which is unable to detect the nearly identical content. Therefore, a similarity hash algorithm is then applied to measure the extent to which assets are duplicated (*e.g.*, `dHash`, `aHash`, and `pHash`), which provides a fault-tolerant measurement on the asset similarity.

6.2.1 Asset Hash. The quantitative results are in Table 5. The duplication rate under the measure of asset hash is consistent with the method in Section 6.1.1. The number of duplicated NFT collections increases from 2,417 (at *AssetURI* level in Section 6.1.2) to 2,653. In other words, *the NFTs of 236 collections exist duplicated assets when there is no duplication of AssetURI (i.e., identical content*

from different AssetURIs). The average duplication rate within these NFT collections is 34.68%. As a whole, there are a total number of 1,412,662 unique hashes of the assets, of which 95.96% has a frequency equal to 1 that these NFT assets have no replica, indicating that *the majority of the NFT contents are still unique.* Of the rest of the 57,079 replicated content hashes, the duplication of most of the hashes is less than 10^2, and 109 hashes are repeated more than 10^3.

Furthermore, we find that 12,688 duplicated assets are identically owed by multiple contracts and 44,391 of them are duplicated within a single contract. It suggests that *77.77% of the duplication occurs within an NFT collection.* The average repetitions of duplicated assets among multiple contracts and within a contract are 28.24 and 25.63, respectively, which shows no significant difference. We then verify the most associated contracts empirically, and we observe that a large proportion of the contract groups that share identical content are correlated in the contract names and symbols. They may contain counterfeit contracts, or they may be different contracts published by application providers, which is in line with the results of duplicated URLs (refer to Section 6.1.3, Section 6.1.4).

6.2.2 Image Fuzzy Hash. The widely-used hash algorithm *Perceptual Hash* (`pHash`) and *Block Mean Value based Hash* (`BlockHash`) are selected to measure the NFT image similarity [60]. We use the open-source library `ImageHash` [12] and `blockhash` [1] for the experiment. The images are resized to $256 * 256$ at first and are then applied to the chosen hash algorithm. The precision of fuzzy hash is controlled by the hash size range in $[8, 16, 32, 64, 128, 256]$ bits, as it determines the granularity of image segmentation. In general, a finer granularity of the image segmentation results in lower repeatability. The result is explained as the increase in duplication, which is the duplication under perceptual hash minus the duplication under ordinary hash (`MD5` results in Section 6.2.1), where the calculation of duplication is again consistent with the method stated in Section 6.1.1. Notice that the measurement scope is narrow to the pure image NFT collection as the perceptual hash is limited to image content, which causes a negative increment to the mixed-type NFTs.

Table 5 shows the results. As the size of the perceived hash increases, the number of duplicate assets tends to approximate the conventional hash algorithm. When the hash size is 8, the number of duplicate assets of `pHash` and `BlockHash` is 3.37% and 6.55% higher than that of the original hash algorithm, respectively. However, when the hash size is greater than 32, the difference in similarity tends to be negligible. `pHash` in general is more sensitive to asset similarity, and the number of duplicate assets increases almost linearly, whereas `BlockHash` increments with small hash sizes are much larger than its increments with large hash sizes. *The proportion of the duplicate assets within a single contract versus across multiple contracts are all around 99% and 1% with variance no greater than 0.9%.* So in this sense, neither `pHash` nor `BlockHash` shows a significant difference from `MD5`. Nevertheless, the average frequency of the multi-contract hash of `BlockHash` is considerably larger than that of `MD5` and `pHash` around 5 to 7 times, indicating that *the images shared by multiple contracts are largely identical under the metric of* `BlockHash`.

	Unique Hash (#)	Duplicated Hash (%)	Multi-Contract		Single Contract	
			prop.	*freq.*	*prop.*	*freq.*
Hash						
`MD5`	1,412,662	4.04%	0.89%	28.25	99.11%	25.63
pHash						
size = 8	1,046,644	7.41%	0.86%	41.42	99.14%	17.93
size = 16	1,154,287	4.98%	1.07%	28.79	98.93%	24.75
size = 32	1,188,263	5.40%	1.03%	28.62	98.97%	21.11
size = 64	1,226,512	4.40%	1.00%	28.58	99.00%	25.10
size = 128	1,226,742	4.40%	1.00%	28.40	99.00%	25.15
size = 256	1,226,787	4.40%	1.00%	28.40	99.00%	25.16
BlockHash						
size = 8	690,782	10.59%	0.87%	77.79	99.13%	15.95
size = 16	931,937	6.80%	0.16%	205.96	99.84%	15.73
size = 32	998,743	5.19%	0.16%	194.54	99.84%	17.96
size = 64	1,025,215	4.60%	0.17%	179.18	99.83%	19.23
size = 128	1,039,439	4.28%	0.17%	170.02	99.83%	20.09
size = 256	1,040,032	4.11%	0.18%	159.68	99.82%	20.93

Table 5: The similarity of NFT assets measured with conventional Hash, pHash, and BlockHash. *The number of Unique Hash* is the number of unique NFT assets. The *Dup Hash*, and *Non-dup Hash* refer to whether the hash corresponding asset has duplication. The *Multi-Contract* and *Single Contract* refer to whether the duplicated hash corresponding asset exists in multiple contracts or within a single contract.

7 RELATED WORK

➤ **Market Trend**. Studies on factors that affect the NFT market are diverse, such as the rarity and scarcity of NFTs, [65, 68], the price of cryptocurrency [50, 58], the influence from social network [53, 63], the creator of NFT [72], the sales auction [64], *etc.* Attempts have also been made to forecast the value of NFTs using statistical techniques [51], machine learning [63, 66], and deep learning [62, 63, 66].

➤ **Security and Privacy.** NFTs are facing a number of security and privacy issues [56, 74]. Research concerning the NFT security includes wash trading [73] and plagiarism [59, 69], and the NFT privacy-preserved solutions have been proposed in [52, 70].

➤ **Infrastructure.** Decentralized storage network, *e.g.*, Storj [46], BitTorrent [23], and SAFE [44], are compared technically in [54], of which the BitSwap [22] is improved to reduce communication overhead and increase the content discovery rate [57]. From another perspective, the incentives of NFT such as rewards and punishment mechanisms [61] are proposed to regulate the NFT trading behavior.

8 CONCLUSION

In this work, we have analyzed the phases along the path of NFT-to-Asset connection in detail, including the storage distribution, accessibility, and degree of duplication. Our results have shown that a non-trivial portion of NFTs have been disconnected from their assets and have highly repetitive data within their collections, which significantly undermines the NFTs' value. We have further discovered that decentralized storage does not show advantages in accessibility, response time, and anti-loss for NFTs, which still needs to be strengthened.

ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China under Grant 2021YFF0900300 and in part by the National Natural Science Foundation of China (Project No. 72031003). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

REFERENCES

[1] 2016. *BlockHash by Commonsmachinery.* https://github.com/commonsmachinery/blockhash
[2] 2021. *Aavegotchi Onchain SVGs.* Retrieved April 6, 2021 from https://docs.aavegotchi.com/overview/onchain-svgs
[3] 2021. *Arweaves.* Retrieved April 6, 2021 from https://www.arweave.org/
[4] 2021. *ERC-1155: Multi Token Standard.* Retrieved April 6, 2021 from https://eips.ethereum.org/EIPS/eip-1155
[5] 2021. *ERC-2981: NFT Royalty Standard.* Retrieved April 6, 2021 from https://eips.ethereum.org/EIPS/eip-2981
[6] 2021. *ERC-3525: Semi-Fungible Token.* Retrieved April 6, 2021 from https://eips.ethereum.org/EIPS/eip-3525
[7] 2021. *ERC-721: Non-fungible token standard.* Retrieved April 6, 2021 from https://eips.ethereum.org/EIPS/eip-721
[8] 2021. *ERC-998: Composable Non-Fungible Token Standard.* Retrieved April 6, 2021 from https://eips.ethereum.org/EIPS/eip-998
[9] 2021. *Etherscan.* Retrieved April 6, 2021 from https://etherscan.io/
[10] 2021. *Filecoin.* Retrieved April 6, 2021 from https://filecoin.io/
[11] 2021. *Hashmasks Provenance Record.* Retrieved April 6, 2021 from https://www.thehashmasks.com/provenance.html
[12] 2021. *ImageHash by Johannes Buchner.* https://github.com/JohannesBuchner/imagehash
[13] 2021. *InterPlanetary File System (IPFS).* Retrieved April 6, 2021 from https://ipfs.io/
[14] 2021. *IPFS Gateway: "dweb.link".* https://github.com/olizilla/dweb.link
[15] 2021. *NFT List provided by Etherscan.* https://cn.etherscan.com/tokens/label/nft
[16] 2021. *Openethereum.* https://openethereum.github.io/
[17] 2021. *Uniswap SVG NFT.* Retrieved April 6, 2021 from https://docs.uniswap.org/protocol/reference/periphery/libraries/NFTSVG
[18] 2022. *89 Seconds Atomized.* https://snark.art/89seconds/assets/artworks/89seconds/89-seconds-Atomized-White-Paper.pdf
[19] 2022. *Akomba Commemorative Token.* https://opensea.io/collection/akomba-commemorative-token
[20] 2022. *Amazon Web Services.* https://aws.amazon.com/
[21] 2022. *Arcana NFT.* https://arcana.network/blog/launching-private-nfts-on-arcana/
[22] 2022. *Bitswap Protocol.* https://docs.ipfs.tech/concepts/bitswap/
[23] 2022. *Bittorrent.* https://www.bittorrent.com/
[24] 2022. *CloudFront.* https://aws.amazon.com/cloudfront/
[25] 2022. *Cloudinary.* https://cloudinary.com/
[26] 2022. *Colorverse Founder.* https://colorverse.medium.com/
[27] 2022. *CryptoKitties.* https://www.cryptokitties.co/
[28] 2022. *Cybertopia.* https://cybertopia.world/
[29] 2022. *Gods Unchained.* https://godsunchained.com/
[30] 2022. *Google Cloud.* https://cloud.google.com/
[31] 2022. *Imgur.* https://imgur.com/
[32] 2022. *IPFS Gateway Checker.* https://ipfs.github.io/public-gateway-checker/
[33] 2022. *IPFS Gateway: "gateway.pinata.cloud".* https://www.pinata.cloud/dedicated-gateways
[34] 2022. *IPFS Gateway: "ipfs.infura.io".* https://infura.io/product/ipfs
[35] 2022. *IPFS Gateway: "ipfs.io".* https://docs.ipfs.tech/concepts/ipfs-gateway/#gateway-providers
[36] 2022. *Mandala Tokens.* https://www.enterthemandala.io/
[37] 2022. *Merge by Pak.* https://www.niftygateway.com/marketplace/collectible/0xc3f8a0f5841abff777d3eefa5047e8d413a1c9ab
[38] 2022. *Mintable.* https://mintable.app/
[39] 2022. *Nifty Gateway.* https://www.niftygateway.com/
[40] 2022. *Opensea.* https://opensea.io/
[41] 2022. *OpenSea "Asset" API.* https://docs.opensea.io/reference/api-overview
[42] 2022. *Pinata.* https://www.pinata.cloud/
[43] 2022. *Rarible.* https://rarible.com/
[44] 2022. *SAFE Network.* https://github.com/safenetwork
[45] 2022. *Secret NFT.* https://scrt.network/about/secret-nfts
[46] 2022. *Storj.* https://www.storj.io/
[47] 2022. *Su Sqares.* https://tenthousandsu.com/
[48] 2022. *"The First 5,000 Days" by Beeple.* https://opensea.io/collection/beeple-everydays
[49] 2022. *Unlock Protocol.* https://unlock-protocol.com/
[50] Lennart Ante. 2022. The non-fungible token (NFT) market and its relationship with Bitcoin and Ethereum. *FinTech* 1, 3 (2022), 216–224.
[51] Hong Bao and David Roubaud. 2022. Non-Fungible Token: A Systematic Review and Research Agenda. *Journal of Risk and Financial Management* 15, 5 (2022), 215.
[52] Ammar Ayman Battah, Mohammad Moussa Madine, Hamad Alzaabi, Ibrar Yaqoob, Khaled Salah, and Raja Jayaraman. 2020. Blockchain-based multi-party authorization for accessing IPFS encrypted data. *IEEE Access* 8 (2020), 196813–196825.
[53] Simone Casale-Brunet, Mirko Zichichi, Lee Hutchinson, Marco Mattavelli, and Stefano Ferretti. 2022. The impact of NFT profile pictures within social network communities. *arXiv:2206.06443* (2022).
[54] Erik Daniel and Florian Tschorsch. 2022. IPFS and friends: A qualitative comparison of next generation peer-to-peer data networks. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 31–52.
[55] Erik Daniel and Florian Tschorsch. 2022. Passively Measuring IPFS Churn and Network Size. *arXiv:2205.14927* (2022).
[56] Dipanjan Das, Priyanka Bose, Nicola Ruaro, Christopher Kruegel, and Giovanni Vigna. 2021. Understanding security Issues in the NFT Ecosystem. *arXiv:2111.08893* (2021).
[57] Alfonso De la Rocha, David Dias, and Yiannis Psaras. 2021. Accelerating Content Routing with Bitswap: A multi-path file transfer protocol in IPFS and Filecoin.
[58] Michael Dowling. 2022. Is non-fungible token pricing driven by cryptocurrencies? *Finance Research Letters* 44 (2022), 102097.
[59] Darius Galiș, Ciprian Pungilă, and Viorel Negru. 2022. A Fast NDFA-Based Approach to Approximate Pattern-Matching for Plagiarism Detection in Blockchain-Driven NFTs. In *Springer DaWaK*.
[60] Qingying Hao, Licheng Luo, Steve TK Jan, and Gang Wang. 2021. It's Not What It Looks Like: Manipulating Perceptual Hashing based Applications. In *ACM CCS*.
[61] Haya R Hasan, Khaled Salah, Ammar Battah, Mohammad Madine, Ibrar Yaqoob, Raja Jayaraman, and Mohammed Omar. 2022. Incorporating Registration, Reputation, and Incentivization Into the NFT Ecosystem. *IEEE Access* 10 (2022), 76416–76433.
[62] Shrey Jain, Camille Bruckmann, and Chase McDougall. 2022. NFT Appraisal Prediction: Utilizing Search Trends, Public Market Data, Linear Regression and Recurrent Neural Networks. *arXiv:2204.12932* (2022).
[63] Arnav Kapoor, Dipanwita Guhathakurta, Mehul Mathur, Rupanshu Yadav, Manish Gupta, and Ponnurungam Kumaraguru. 2022. Tweetboost: Influence of social media on nft valuation. *arXiv:2201.08373* (2022).
[64] Pavel Kireyev. 2022. NFT Marketplace Design and Market Intelligence. (2022).
[65] Amin Mekacher, Alberto Bracci, Matthieu Nadini, Mauro Martino, Laura Alessandretti, Luca Maria Aiello, and Andrea Baronchelli. 2022. How rarity shapes the NFT market. *arXiv:2204.10243* (2022).
[66] Matthieu Nadini, Laura Alessandretti, Flavio Di Giacinto, Mauro Martino, Luca Maria Aiello, and Andrea Baronchelli. 2021. Mapping the NFT revolution: market trends, trade networks, and visual features. *Scientific reports* 11, 1 (2021), 1–11.
[67] Constantinos Patsakis and Fran Casino. 2019. Hydras and IPFS: a decentralised playground for malware. *International Journal of Information Security* 18, 6 (2019), 787–799.
[68] Dinuka Piyadigama and Guhanathan Poravi. 2022. An Analysis of the Features Considerable for NFT Recommendations. *arXiv:2205.00456* (2022).
[69] Ciprian Pungila, Darius Galis, and Viorel Negru. 2022. A New High-Performance Approach to Approximate Pattern-Matching for Plagiarism Detection in Blockchain-Based Non-Fungible Tokens (NFTs). *arXiv:2205.14492* (2022).
[70] Rui Song, Shang Gao, Yubo Song, and Bin Xiao. 2022. ZKDET: A Traceable and Privacy-Preserving Data Exchange Scheme based on Non-Fungible Token and Zero-Knowledge. (2022).
[71] Dennis Trautwein, Aravindh Raman, Gareth Tyson, Ignacio Castro, Will Scott, Moritz Schubotz, Bela Gipp, and Yiannis Psaras. 2022. Design and evaluation of IPFS: a storage layer for the decentralized web. In *ACM SIGCOMM*.
[72] Kishore Vasan, Milán Janosov, and Albert-László Barabási. 2022. Quantifying NFT-driven networks in crypto art. *Scientific reports* 12, 1 (2022), 1–11.
[73] Victor von Wachter, Johannes Rude Jensen, Ferdinand Regner, and Omri Ross. 2022. NFT Wash Trading: Quantifying suspicious behaviour in NFT markets. *arXiv:2202.03866* (2022).
[74] Qin Wang, Rujia Li, Qi Wang, and Shiping Chen. 2021. Non-fungible token (NFT): Overview, evaluation, opportunities and challenges. *arXiv:2105.07447* (2021).

A APPENDIX

A.1 Data Source

We use the results of calling the *totalSupply* function and the *TokenURI* function of the NFT contract to describe the basic situation of the NFT contract and the proportion of data collected from different sources. The contracts are divided into 5 circumstances. ❶ *Complete*: the *totalSupply* is identical to the available *TokenURI*; ❷ *Not Callable totalSupply*: The *totalSupply* is not callable or the return is null; ❸ *Empty Supply*: the *totalSupply* is zero, indicating that the contract is empty; ❹ *Great Supply*: The *totalSupply* is greater than 1 million; ❺ *Not Callable TokenURI*: the *totalSupply* is a non-zero value, but *TokenURI* is not callable or the returns are null.

Data sources are Ethereum Mainnet and Opensea. According to Table 6, from the results of the initial call to the Ethereum, there are 8,056 contracts in *Complete* status of which we successfully collect the entire *TokenURI* information. There are 752 contracts in *Not Callable totalSupply* status, and there are 685 *Not Callable tokenURI* contracts. There are two reasons why the functions are not callable, one is that the contract is not in the standard format, and another is that it might be the ERC-998 NFT contract as it does not provide an interface of *totalSupply()* or *tokenURI()*. Besides, the number of *Empty Supply* and *Great Supply* contracts are 2,851 and 9, respectively. We thereby request the *Not Callable totalSupply* and *Not Callable TokenURI* contracts to Opensea. Among the 752 *Not Callable totalSupply* contracts, 549 of which then become *Complete* and 203 of which are *Empty Supply*; Among the 685 *Not Callable tokenURI* contracts, 615 of which then become *Complete* and 70 of which are *Empty Supply*.

	❶	❷	❸	❹	❺	Sum
Ethereum Mainnet	8,056	**752**	2,851	9	**685**	12,353
Opensea + ❷	549	\	203	\	\	**752**
Opensea + ❺	615	\	70	\	\	**685**
Aggregate Result	9,195	\	3,123	9	\	12,353

Table 6: Contract information and data source. ❶ *Complete*, ❷ *Not Callable totalSupply*, ❸ *Empty Supply*, ❹ *Great Supply*, ❺ *Not Callable TokenURI*.

A.2 Factors Affecting Measurement Accuracy

A.2.1 ***IPFS Instability.*** As a large amount of NFTs' data is hosted on IPFS, the accuracy of the measurement results highly depends on the availability of IPFS data. During data collection, we observe that the data availability of IPFS exists inconstant characteristics (refer to Section 2.2.2), which is reflected in the following aspects: ❶ There are plenty of cases where data can't be found. ❷ The accessibility of the same data at the same time with different public gateways as well as the IPFS nodes built in our experiment shows different results. ❸ The data request via IPFS is extremely slow, rendering additional response failure. One reason is that no more nodes have saved this data and the data is indeed lost forever. Another reason is that the data is not lost, but possibly nobody online has it, or the node with the data is hidden behind the internal network, or the node with data does not publish the data in the way that an ordinary node can find it.

Solution: The IPFS data is collected with multiple gateways to enlarge the data coverage. The response duration and the hit rate of different IPFS gateways are presented in Figure 4 and Table 7 to describe the above instability. It can be seen from Figure 4 that the response time of different gateways varies greatly. The fastest response is Infura, IPFS.io, and Dweb.link almost cut off after a minute, while Pinata has the heaviest tail. The numerical results are summarized in Table 7. Although the time difference is large, the hit rates of the gateways differ only 3.43% at most, which is not higher enough but not significantly different. Overall, 88.18% of the URLs are fully obtained by all gateways, and the union of URLs that are successfully obtained by each gateway covers 97.98% of the sample set. Therefore, the above data shows that our method is effective to overcome the above instability.

A.2.2 ***The Changing Availability.*** We note that both IPFS and servers' data availability fluctuate at different moments. For example, NFTs from Unlock Protocol [49] fail to be accessed until the last round of data collection. The low availability occurs randomly and is hard to define as data loss as it will cause data loss to be overestimated.

Solution: We extend the data collection process into a six-month period to cover as much of the time as possible when servers went online. In each collection cycle, we first remove invalid files by keywords identification, such as "*Gateway Time-out*", "*Server Error*", *etc.*, and then keep requesting the uncovered data. New NFTs are acquired after each cycle of data collection, reducing false-positive judgement of data loss.

A.3 Discussion on Mitigation Strategies

We discuss mitigation strategies to resolve and alleviate the issues of NFT lost, inaccessible and duplicated. Potential solutions are suggested for NFT holders, developers, and applications, respectively.

A.3.1 NFT (potential) holders. For **NFT holders**, we recommend they keep a copy of the data themselves and keep assets in separate locations to reduce the risk of data loss. If the NFT is on a decentralized storage network, we recommend they recall the data regularly to keep NFT from being forgotten in the storage network and use a pinning service to enhance data durability. For **NFT potential holders**, they need to understand the potential risks of NFTs before purchase and examine carefully on: how the NFT is binding to the asset, on which platform it is hosting, whether there is hash value on-chain, whether the asset is unique, whether the NFT metadata in the smart contract can be modified at will, and whether the right to change is with the user or the application provider.

A.3.2 NFT developers. NFTs are in peril under the existing scheme, which poses new challenges to optimize the NFT mechanism to meet the security demands of current public NFTs and new emerging private NFTs [21, 45]. Notice that we will only discuss the solution for NFTs with off-chain storage, as on-chain storage is impractical for most scenarios due to the expensive on-chain storage cost. We encourage **NFT developers** to mitigate the weaknesses of NFTs from the angles of:

❶ **Enhance the NFT-to-Asset connection.** For public NFTs, the existing IPFS scheme meets the ownership traceability and tamper-resistance properties for public NFT data, as the data hash (that is the CID on IPFS) is recorded on-chain and forms an effective

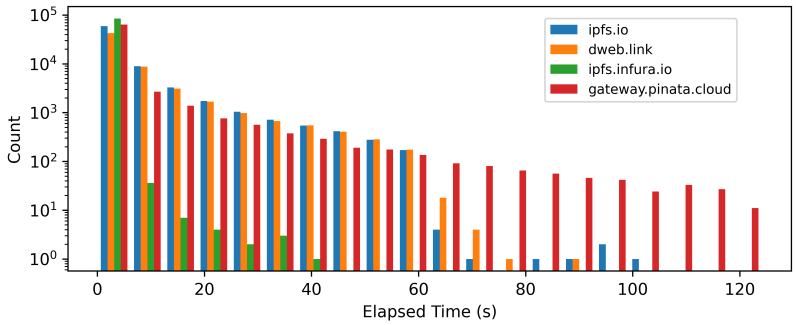

Figure 4: IPFS response elapsed with different gateways.

Gateways *https://<gateway>/ipfs/<cid>*	Average Response	Hit Rate	Failure Type			Hits	# of URL
			Server	Client	Other	4	88.18%
ipfs.io	5.19s	93.41%	4,159 (63.07%)	2,387 (36.20%)	48 (0.07%)	3	5.89%
dweb.link	6.32s	93.15%	4,848 (70.80%)	1,943 (28.38%)	56 (0.08%)	2	1.93%
ipfs.infura.io	1.45s	92.60%	4,661 (63.02%)	2,691 (36.38%)	44 (0.06%)	1	1.79%
gateway.pinata.cloud	3.32s	96.03%	1,417 (35.65%)	2,427 (61.05%)	131 (3.30%)	0	2.27%
Aggregate Hit Rate		97.98%					

Table 7: Comparison of IPFS gateways. We sample 100,000 IPFS-style URLs evenly from a set including both *TokenURI* and *AssetURI*, and 4 active gateways from the IPFS gateway selector [32]. The *CID* of IPFS is extracted from the sampled URLs, and then combine with the selected gateways to form a new IPFS URL for the serial request. The process is performed simultaneously among gateways to eliminate the network effect caused by the client.

constraint to off-chain asset. Since the NFT is public, the asset can naturally be seen and owned by anyone. Data published on IPFS is publicly retrieved, unless it is encrypted in advance. Therefore, the IPFS solution towards traceable ownership is not applicable to private NFT data.

For private NFTs, only the owners (past and current) of the NFT owns the digital asset. The record of asset hash enables an owner to self-prove the ownership of the off-chain data. In the current scheme, NFT sellers and buyers have the same data before and after the transaction. As in the process, the blockchain will only change the NFT's ownership pointer from the seller to the buyer with the asset unchanged (so they share identical data). When an illegal use of data occurred and ownership determination is required, no available scheme is sufficient to trace back the breach as far as we know, as it cannot determine which owner illegally leaks the data. We suggest distinguishing data owned by different owners by digital watermarking, and verifying such invisible but different data by on-chain fuzzy hash. The robust watermark will be the unique identification on the blockchain, and can thus be traced back to the ownership on the blockchain even when the multi-hop URL is broken.

❷ Improve the sustainability and availability of decentralized storage. We use IPFS to represent decentralized storage based on its dominant position in decentralized storage. The IPFS data collection in this work shows a great difference in accessibility and response duration when querying different gateways (nodes). Not only did our data access run into such problems, but also there are complaints from the community, that files that have been uploaded and pinned to IPFS are unable to access. We notice that some works have explored the IPFS network, such as the characteristics of the underlying P2P network [55, 71], the malicious resource occupancy attack [67], *etc.* However, research has not been conducted on the causes of the aforementioned problem, such as the distribution of content replica (Replication Protocol), the efficiency of node discovery (DHT), the efficiency of content exchange (Bitswap Protocol), *etc.*, which leaves a barrier to better understanding the underlying reasons for the IPFS's insufficient performance. We suggest the first step is to conduct a measurement of the effectiveness of the underlying protocol to identify the root cause. On this basis, the second step is to study and improve the bottleneck of each IPFS component to increase the performance.

A.3.3 NFT applications. Users(holders) are at significant risk as most of them have no actual control over their assets. We call on **NFT applications** to make a clear division of storage responsibility and the right of use with NFT holders. It prevents the problem of ambiguity of user and provider's responsibility when NFTs are lost in unavoidable situations. We also recommend NFT applications open-source their smart contracts and scrutinize the security via a trusted third party. Such security review will convince users of the authentication of the NFT structure, *e.g.*, how the metadata is generated and fed on-chain, what features of the NFT are included in the metadata, and who can modify the metadata. At last, they can actively upgrade the technical framework and make efforts to integrate new technologies to improve the reliability of NFTs.