A privacy awareness framework for NFT avatars in the metaverse

Dorottya Zelenyanszki
Griffith University
dora.zelenyanszki@griffithuni.edu.au

Zhé Hóu
Griffith University
z.hou@griffith.edu.au

Kamanashis Biswas
Australian Catholic University
kamanashis.Biswas@acu.edu.au

Vallipuram Muthukkumarasamy
Griffith University
v.muthu@griffith.edu.au

***Abstract*—Metaverse is a platform that offers unique user experiences. Users join the virtual worlds by using a virtual representation called an avatar. There is increasing use of Non-Fungible Token (NFT) avatars in metaverse applications that enable users to interact with each other in virtual worlds. These avatars contain not only personal information but also the behavioural footprints of users. Therefore, privacy preservation is critical for establishing a safe environment and enhancing the level of confidence in the metaverse. This paper proposes a novel privacy awareness framework that leverages machine learning (ML) algorithms to identify malicious patterns and alert users. This enables user awareness and the ability to adopt a required level of risk mitigation strategies. To test the framework, we have designed and implemented a test-bed where we can deploy various attack scenarios and collect data. We also present a case study on NFT cloning with the associated data collection.**

***Index Terms*—metaverse, NFT, avatar, privacy, blockchain, ML**

I. Introduction

Metaverse is often introduced as the new version of the Internet that presents an application platform for several sectors, and its need has accelerated after the pandemic with the wave of thirst for digital transformation. This has become possible due to the removal of the physical boundary and the ease of operation and control by sophisticated algorithms. In a metaverse, the participants, real and avatars, can meet and interact with each other regardless of their physical locations and current health conditions. These features make the metaverse the biggest game changer for almost every sector such as health, education, business, and governance [1].

In general, the metaverse can be viewed as a layered architecture as presented in Lim et al.'s work [2]. The first layer is the physical world which includes the real users who are connected to the virtual world through their metaverse avatars. In this layer, service providers use sensory devices to collect necessary information and provide them to the second layer (the virtual world) for content-creation. The third layer, the metaverse engine, includes all inputs from which the content is created and maintained. The final layer consists of communication, computation and storage-related elements.

An avatar holds both personal and behavioural information. The latter is collected from the user's behaviour in the virtual world, for example, the places the users visit and the metaverse elements they interact with. However, user interactions in virtual worlds also create several privacy-related issues [3]. First, behavioural information is highly useful for malicious participants. Second, digital footprints can be created from this information which can be used for user profiling.

In this research, we focus on the privacy aspects of NFT avatars. NFTs are unique tokens that can represent any type of objects such as art, music or a virtual user. They were first introduced on the Ethereum blockchain [4]. This paper mainly focuses on privacy threats pertaining to NFT avatars. Therefore, we first present an attack scenario where behavioural information is used to launch a privacy attack. Then, we propose a novel privacy awareness framework that integrates ML algorithms for pattern creation and privacy prediction in the metaverse. We also design and implement a test-bed that is used as a generic tool for data collection and evaluation of the framework.

II. Related work

Ghantous et al. [5] described the potential of ML in the metaverse space. They also introduced two strategies to enhance the metaverse using ML. The first strategy focused on improving the security of blockchain assets through the use of artificial neural networks and linear regression. On the other hand, the second strategy covered the usage of ML to predict and recommend NFTs for users. They proposed two recommendation systems (Collaborative Filtering Recommender System and Content-Based Filtering Recommender System) as valuable tools for enhancing the metaverse.

Amiri-Zarandi et al. [6] presented privacy-preserving ML techniques related to the Internet of Things (IoT) layers. In the perception layer, they mentioned the transmission of aggregated data instead of raw data to mitigate the risk of data leakage. To enhance privacy in the network and the application layers, they used federated learning (FL) and distributed deep learning modules. However, by using these techniques, noisy data is also produced during the process. The last layer is the application layer, where all the included applications have two functionalities: data storage and processing. Both of these require investment in the privacy aspect, for example, data storage needs authentication and access control. This layer also includes all the problems related to information privacy, and ML is used in several connected applications.

Kang et al. [7] stated that the metaverses gain the data they use from industrial IoT, which they use to improve efficiency. However, issues like data leakage endanger privacy.

Fig. 1. An example attack scenario.

To solve this, they proposed a cross-chain-empowered privacy-preserving framework that uses FL where the cross-chain interaction provides secure model aggregation. The workflow for this framework is as follows: Learning tasks are published, and the learning requests are sent to the main chain. This chain sends the task to the relay chain, which is a cross-chain management platform that forwards and verifies the data and also handles the connection between several blockchains. After that, the task is performed in both virtual and physical worlds, and participants from both spaces also take part in it. Once it is finished, the updated models are verified and secured on the subchains. After that, it is transferred to the main chain. From these tasks, a new global model can be generated, which will be downloaded by all workers. This way, the whole system benefits from all tasks.

III. An attack scenario in metaverse

Falchuk et al. [8] presented a mechanism that allows the metaverse users to create multiple clones of their NFT avatars. This mechanism acts as a privacy-preserving tool that the users can use to confuse malicious participants. However, this mechanism also sets up a privacy risk. The malicious actors can also create clones of other NFT avatars. They may create a visually identical NFT avatar and impersonate the end-user behind the original avatar, which may mislead other users. If the attacker also collects behavioural information from the original avatar, that can further enhance the clone and the malicious user can create more problematic issues in the virtual world. With impersonation, these users may get financial gain or extract highly sensitive personal information from other victims. The attack scenario is depicted in Fig 1.

IV. The proposed privacy awareness framework

This section presents the conceptual design of the proposed privacy awareness framework. It is designed to create privacy patterns that use previously measured behavioural information with the help of ML. Based on the outcomes, the user can be notified regarding a potential privacy issue in the virtual world in real-time. The patterns also create a foundation for privacy prediction mechanisms implemented by ML. Since the prediction is sent back to the virtual world in the form of a notification, the proposed framework helps to create a self-learning virtual world where participants gradually build their privacy awareness through daily interactions. The conceptual design of the framework is illustrated in Fig 2.

A. Elements of the framework

Fig 3 presents a layered architecture of the metaverse proposed by Lim et al. [2] and our framework within it. Unlike this generic architecture, our framework utilises ML algorithms for privacy prediction. Here, we describe the elements of the proposed framework.

The first layer is the physical layer through which real-world users connect to the metaverse. This layer serves as a model for content creation in the virtual world. The second layer is the virtual world which includes two major elements:
NFT avatars: As defined earlier, it is the physical user's virtual representation that enables the user to participate in unique experiences offered by the metaverse.
Metaverse elements: This group includes the requirements, available actions, functionalities, limits of the virtual world and its contents with which the NFT avatars interact. For example, an NFT avatar can go to a museum in the metaverse and attend an exhibition of virtual art.

The third layer is the metaverse engine which contains the pattern analysis-related elements such as follows.
Database: Database is used to store the collected behavioural data. It also includes the existing privacy patterns so they can be used later for pattern creation/analysis and as input for the privacy-prediction ML mechanisms.
ML component: This includes the ML part of the framework. It communicates with the database and, based on the available data, handles three processes: pattern creation, privacy prediction, and notification.

B. Data collection aspect

Behavioural information can depend on several factors. For example, the creators of the metaverse are able to set system limits that may disable certain types of interactions. It also has a strong dependence on the type of the NFT avatar. Human-like representations are more open to digital footprint creation because they may contain facial or verbal elements which offer a lot of information regarding the end user. While NFT avatars with a limited appearance mainly cover factors that are usually not as direct as the previously mentioned ones.

If the metaverse project already has determined privacy patterns, it is highly recommended to initiate the database with them so they may be further improved with additional data. These existing patterns can be based on discovered vulnerabilities/threats or already implemented and used privacy mechanisms. For example, Falchuk et al. [8] mentioned some privacy mechanisms in their work that can be transformed into patterns. For instance, the avatars can create multiple clones of themselves, allocate a space in the virtual world just for themselves or transport to a different part of the metaverse when they think it is necessary.

C. Machine Learning

The role of ML in the proposed framework is to extract and classify the behavioural information that is required for the patterns to notify users and make predictions. To achieve this, numerous attributes are set for the patterns and data collection regarding them. For example, for the presented cloning attack scenario, we can set several instructions for the avatars and since they are going to perform the instructions in a different order with a varying time factor, a significant

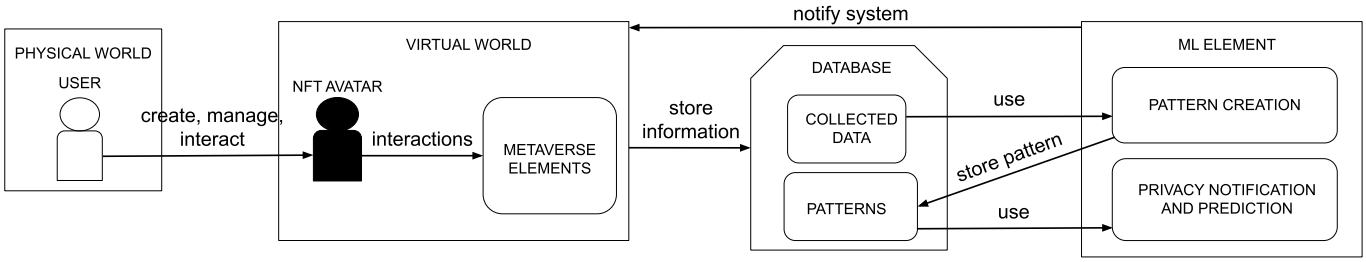

Fig. 2. An overview of the proposed privacy awareness framework.

amount of information can be extracted regarding potential malicious clones.

Our framework is designed to support different ML algorithms for identifying privacy violations and malicious behaviours. Since the processed datasets are often structured, we expect that algorithms such as random forest and gradient boosting will be particularly suitable for classification tasks, while regression algorithms are needed for tasks like computing similarity scores between NFT clones. We also envisage that in a mature and large virtual world network, more considerations should be taken to protect the user's privacy when ML is deployed. In this case, FL can be used in a distributed fashion without transferring user data across the Internet. More specifically, we will adopt Markovic et al.'s method that integrates random forest into FL for detecting intrusions [9].

D. ML related privacy risks

Although in this research, we are proposing a privacy awareness framework by using ML, existing literature suggests that ML algorithms themselves are open to several types of privacy attacks, such as reconstruction or model inversion attacks from malicious users coming from both inside and outside of the system [10]. To address this issue, based on the work of Truong et al. [11], we propose to employ privacy preservation in the FL module. This includes several existing methods which give us a good foundation to design the ML component with enhanced privacy.

More specifically, we consider security and privacy threats in FL from three perspectives: at the server end, the communication phase, and at the client end. On the server, we assume that the attacker could infer illegitimate information from parameter updates. To defend this, we adopt Bonawitz et al.'s Secure Aggregation Protocol [12] on both the server side and the client side of FL, which prevents the server from introspecting individual model updates. The communication will be protected by traditional protocols such as SSL/TLS and HTTPS. For the client side (local nodes), we adopt Geyer et al.'s approach [13] to apply perturbations (random noise) to local model parameters so that adversaries cannot train the original model even if they obtain the parameters meanwhile maintaining the predictive performance of the global model.

E. Workflow

The workflow of the proposed framework is as follows: The user enters the virtual world with an NFT avatar which is based on the data the user transfers from the physical world. During the creation process, personal information is put into the avatar; however, that type of information is not our focus in this research. Within the metaverse, the NFT avatars interact with other metaverse elements like with another avatar or other types of NFT assets. For example, the avatars are moving around in the virtual world, purchasing an NFT asset or creating new user content for the system. Those interactions create behavioural information that can be used for profiling purposes. We collect this information and put it into a database.

We can also assume that the privacy issues of Web 2.0 are going to be transferred to the Web 3.0 in new and emerging forms. For example, avatar cloning has a parallel with the cloning of social media profiles. In this case, malicious actors often target the victim's user friend list [14]. This is probably a scenario that is going to be followed in the metaverse as well. The information that the malicious clone can collect from other avatars about the victim user can be used to further enhance the cloned avatar and thus, more advanced behavioural data and highly sensitive personal information can be extracted. The stored information is transferred to the ML component where it is used for looking for determinable privacy patterns. Once a pattern is identified, it is going to be stored in the database so the upcoming data can be compared to it by using suitable ML methods.

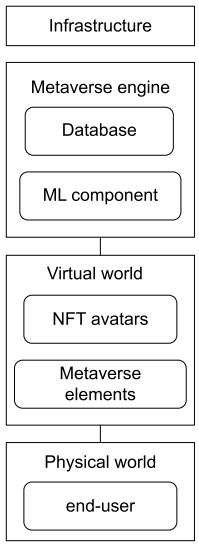

Fig. 3. A layered architecture of metaverse.

If the pattern discovers a privacy risk, the virtual world will get a notification, so the users behind the corresponding NFT avatars can adopt appropriate precautions to enhance their privacy and mitigate the discovered vulnerability. The identified pattern group and the database form the base for the privacy prediction process. If a prediction is made, the users are notified. This way, they have the ability to prevent an already identified malicious activity before it takes effect. The base also creates a foundation to identify and potentially predict new malicious activities in addition to those that are well-known in the virtual world. Because of this, the system

is capable of continuous self-learning.

V. Test-bed: Virtual world environment

Data collection and database creation are essential in this work. However, due to the fact that the current metaverse applications are relatively new, the publicly available data is limited and requires collaboration with the industries. In order to lay a foundation for the creation of a database for this research, a test-bed environment was designed and implemented. It is a virtual world imitation in a 2D format that includes a map where NFT avatars can be added, and the user can move them around. Once the concept model is established, this setup can be extended for more complex applications.

A. Currently available functionalities

At this initial stage, the test-bed includes the following functionalities:

- Mint and add NFT avatars to the user wallets: The users can connect to the test environment with a wallet address, and they can upload an image, set up a name and description, and with the metadata that is created from these, mint and add an NFT avatar. They can also create multiple avatars in the same way.
- Movement of NFT avatars: The NFT avatars are added to the map, and the users are able to move their own avatars with keyboard combinations on the map.
- Change the visibility of NFT avatars: The users can change the visibility of their avatars. If they choose to make an avatar invisible, it will be removed from the map.
- Avatar cloning: The users are able to clone NFT avatars using three options: cloning based on metadata; InterPlanetary File System (IPFS) image URL (this is the case if someone uses the same file as well because the image hash is going to be the same); or an image which is highly similar to the original file. The last one points out the case when malicious users slightly change the picture to avoid detection. In the test environment, the first and the third option is reproducible through the added inputs, while the second option is included in a back-end script described in Section VI.

The test-bed virtual world can be further enhanced in order to be adaptable for deploying other privacy-related attack scenarios. Also, other elements for further enhancement and evaluation of the framework can be incorporated as needed.

B. Implementation

For testing purposes, the test virtual world uses a local hardhat node but later it is going to be transferred to the Goerli test network so other users can join and participate in the testing and evaluation. The test virtual world is implemented as a Next.js web application, and JavaScript is used to create the necessary elements. The map and its elements are created and loaded with the use of the Phaser HTML5 framework. Moralis is used to create a server where the data regarding the NFT avatars can be stored and it is connected to our local hardhat node. It is also synced with the events that are emitted in the smart contract functions, which makes it possible to store data that are included in these events. This is how we store every information regarding a newly added NFT avatar. For connecting to the smart contracts and calling their functions from the front-end side, the ethers.js library is used. Moralis APIs can also be used to interact with smart contracts and NFTs after the transfer to the test network. At this stage, hardhat nodes are not supported in the APIs. The smart contracts are created by the Solidity programming language, and they adopt existing smart contracts from the OpenZeppelin Contracts library. The back-end functionalities, such as running the scripts, are handled by hardhat.

C. Automated data collection

To create an initial database, we have to add the first set of patterns so future ones can be created based on these. Since cloning attack is a very well known privacy risk, we have decided to include it as the first pattern.

Although the web application allows users to create and clone NFT avatars in order to gain enough data for sophisticated pattern creation, the data collection process has to be automated. For this purpose, a script was developed that randomises the NFT avatar minting and adding process based on 20 free pictures from unsplash[1]. It stores the images and the metadata in IPFS through the Pinata SDK and places every image and hash of metadata in local variables so they can be used later in the cloning. Some of the selected pictures were also slightly modified, which serves as the base for the third presented option for cloning an NFT avatar. The script first adds a randomised number of normal NFT avatars and sends them to addresses that belong to the local hardhat node. After that, it clones a randomised number of NFT avatars. For that, it uses one of the mentioned cloning options. In the tokenURI's case, it uses one of the existing tokenURIs and fully clones the avatar NFT. In the case of the image URL, it gets an existing image URL from one of the tokenURIs and adds that as the image URL for the new avatar NFT's metadata. The other parts of the metadata are randomly filled. The final option is cloning based on the slightly changed image where for the new image a completely different image URL is created. The script always randomly chooses which option to use for cloning. By running the script, we get a randomised NFT avatar collection that can be used for adding the first mechanisms to detect the cloning privacy issue.

VI. Case study: NFT cloning identification

In this section, we demonstrate the cloning avatar issue as a pattern in our privacy awareness framework. At this stage, only the information included in the metadata is checked, but in future improvements, the behavioural data is going to be added as described in Section IV-C.

To check NFT cloning, another script was created. This script goes through every added avatar and checks whether

[1]https://unsplash.com/

TABLE I
AN ANALYSIS OF NFT AVATARS AND THEIR CLONES.

TokenId	Clones	Similarity
2	Clone #31	100.00%
	Clone #48	63.33%
	Clone #54	63.33%
4	Clone #11	100.00%
	Clone #31	100.00%
5	Clone #18	100.00%
	Clone #46	63.33%
7	Clone #50	63.33%
	Clone #56	63.33%
8	Clone #37	100.00%
	Clone #42	63.33%
9	Clone #40	100.00%
	Clone #52	100.00%
0	Clone #29	100.00%
3	Clone #14	100.00%
6	Clone #35	100.00%
17	Clone #21	63.33%
20	Clone #44	96.67%

there are any previously added NFTs which has either the same tokenURI, image URL or a highly similar image. If there is an existing NFT in the database, it also calculates a matching score based on how much similarity exists in the metadata. In the tokenURI's case, this score is going to be 100%. To check the cloning based on the tokenURI or the image URL, only a string comparison is performed. For image-based cloning, the Jimp image processing library is used which provides an image similarity check based on the perceived distance and the pixel difference. If the cloning is done either by the image URL or an almost equal image, that means that one-third of the metadata is equal; therefore, the similarity percentage is at least 30%, and the rest of the actual percentage is calculated based on the similarity of the other parts of the metadata object. The script creates a JSON object with the cloning information. Table I presents the outcomes of cloning attacks in the test-bed environment. For better readability, the results are displayed in a table format and sorted by the number of clones, starting with the maximum value. Please note that avatars without any identified clones are not included in the table. There are several clones with a 100% score which means that they were most probably added by a tokenURI. This score is the result of the fact that when the cloning is based on a tokenURI the whole metadata object is copied, therefore, there will be no differences between the two NFT avatars at all. The clones with other scores can be the result of either the image URL-based or slightly changed image-based cloning. The entries with 63.33% similarity are a result of the image being similar as well as a subset of the metadata properties being identical, whereas Clone #44 has 96.67% similarity score indicating that the metadata objects are almost identical.

A match does not necessarily mean that the clone is created by a malicious user; it may have been added by a regular user for different purposes. To extend the cloning comparison with the identification of malicious intent, the analysis of behavioural information is highly important.

VII. CONCLUSION AND FUTURE WORK

NFT avatars are an essential part of the metaverse. While they enable the users to join virtual worlds and interact, the sensitive data which could be captured can lead to personalised attacks. In this paper, we proposed a privacy awareness framework that utilises ML for pattern creation, analysis, privacy threats prediction and user notification. We also included a test-bed virtual world and presented how automated data collection can be added to the creation of an initial database for pattern generation. For simplicity, this test-bed was designed and implemented in a 2D format. However, our future work aims to implement the proposed framework in a 3D environment.

We also plan to extend this work using behavioural data as well which is going to be based on the existing malicious approach of Web 2.0 and available functionalities of existing metaverse projects. In future work, other privacy attack scenarios will be considered and the privacy issues of ML will be explored more comprehensively as well.

REFERENCES

[1] J. Thomason, "Metahealth - how will the metaverse change health care?" *Journal of Metaverse*, vol. 1, no. 1, pp. 13 – 16, 2021.

[2] W. Y. B. Lim, Z. Xiong, D. Niyato, X. Cao, C. Miao, S. Sun, and Q. Yang, "Realizing the metaverse with edge intelligence: A match made in heaven," *IEEE Wireless Communications*, pp. 1–9, 2022.

[3] Y. Wang, Z. Su, N. Zhang, D. Liu, r. xing, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," 03 2022.

[4] L. Ante, "The non-fungible token (nft) market and its relationship with bitcoin and ethereum," *SSRN Electronic Journal*, 2021.

[5] N. Ghantous and C. Fakhri, "Empowering metaverse through machine learning and blockchain technology: A study on machine learning, blockchain, and their combination to enhance metaverse," 07 2022.

[6] M. Amiri-Zarandi, R. A. Dara, and E. Fraser, "A survey of machine learning-based solutions to protect privacy in the internet of things," *Computers & Security*, vol. 96, p. 101921, 2020.

[7] J. Kang, D. Ye, J. Nie, J. Xiao, X. Deng, S. Wang, Z. Xiong, R. Yu, and D. Niyato, "Blockchain-based federated learning for industrial metaverses: Incentive scheme with optimal aoi," 06 2022.

[8] B. Falchuk, S. Loeb, and R. Neff, "The social metaverse: Battle for privacy," *IEEE Technology and Society Magazine*, vol. 37, no. 2, pp. 52–61, 2018.

[9] T. Markovic, M. Leon, D. Buffoni, and S. Punnekkat, "Random forest based on federated learning for intrusion detection," in *Artificial Intelligence Applications and Innovations*, I. Maglogiannis, L. Iliadis, J. Macintyre, and P. Cortez, Eds. Cham: Springer International Publishing, 2022, pp. 132–144.

[10] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 49–58, 2019.

[11] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the gdpr perspective," *Computers & Security*, vol. 110, p. 102402, 2021.

[12] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," ser. CCS '17, 2017, p. 1175–1191.

[13] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," 2017. [Online]. Available: https://arxiv.org/abs/1712.07557

[14] G. Jethava and U. P. Rao, "A novel defense mechanism to protect users from profile cloning attack on online social networks (osns)," *Peer-to-Peer Networking and Applications*, vol. 15, 09 2022.