IEEE INFOCOM 2023 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) | 978-1-6654-9427-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOMWKSHPS57453.2023.10225937

# Vetting Privacy Policies in VR: A Data Minimization Principle Perspective

Yuxia Zhan*, Yan Meng*, Lu Zhou†, and Haojin Zhu*‡
*Shanghai Jiao Tong University, Shanghai, China, †Xidian University, Xi'an, China
Email: {dabeidouretriever, yan_meng, zhu-hj}@sjtu.edu.cn, zhoulu@xidian.edu.cn

*Abstract*—**Virtual Reality is thought to be the prototype of the next-generation Internet, consisting of more I/O devices and interactive methods than traditional mobile systems. Hence VR developers need to inform users what data is collected and shared, which is generally conveyed by privacy policies. Existing research has examined the consistency between the VR app's privacy policy and its corresponding actual behaviors. However, few studies paid attention to the data minimization principle, *i.e.*, whether a privacy policy claims to collect no more data than it practically needs to implement the app's functionalities. In this poster, we targeted a mainstream VR platform and analyzed the data minimization principle compliance of privacy policies for all 1,726 VR apps in this platform. Experiment results show that 48.1% VR apps potentially violate the data minimization principle. Moreover, the comparative experiments reveal significant differences in the distribution of data collection between VR and non-VR apps.**

*Index Terms*—**Virtual reality, Privacy policy analysis, Data minimization principle**

## I. Introduction

To provide an immersive experience, VR introduces more I/O devices as well as novel interactive methods, which brought more potential privacy risks than traditional mobile systems. De Guzman *et al.* [1] discovered that much private information about users (*e.g.* body shape, room scale, mentality *etc.*) can be inferred from the data collected by VR devices (*e.g.* motion data, eye tracking data *etc.*). Adams *et al.* [2] surveyed that both developers and users express equal extent concerns about personal data collection and management of VR apps. Hence, as the main way to convey such information, privacy policy plays an important role between first-party (developers or companies) and second-party (users). A well-written privacy policy is expected to satisfy two important principles about data collection: **1) Transparency**, it should state clearly what data the app actually collects; **2) Minimization**, it should *only* state what it actually needs.

However, far too little attention has been paid to studying the privacy policies of VR apps. To the best of our knowledge, only OVRSEEN [3] studies this problem by examining the consistency between what a privacy policy claims and what the corresponding VR app actually does. But results of OVRSEEN are based on analyzing 102 privacy policies of top-selling apps in Oculus, which potentially has bias referring to the actual privacy situation in VR ecosystem. Moreover, OVRSEEN does not consider the data minimization principle of privacy policies. To fill up this gap, we conducted the first large-scale privacy policy analysis of data minimization principles of all 1,726 VR apps in Oculus platform and its authorized third-party App Lab.

Our research reveals the severe compliance issue in VR privacy policies, that 48.1% of VR apps potentially violate the minimization principle prescribed by most privacy laws, the majority of which is due to the collection of email and IP overbroadly. Besides, through the comparative experiments between VR and non-VR apps, we find notable differences in their data collection distributions. VR apps tend to collect more biometric and gender information. The overbroad behaviors of VR apps are biased mainly on biometric data, device ID, email, and IP. We hope our findings will arouse the community's attention to privacy issues in VR apps and lay the ground for relative research.

## II. Methodology

### A. VR apps dataset collection

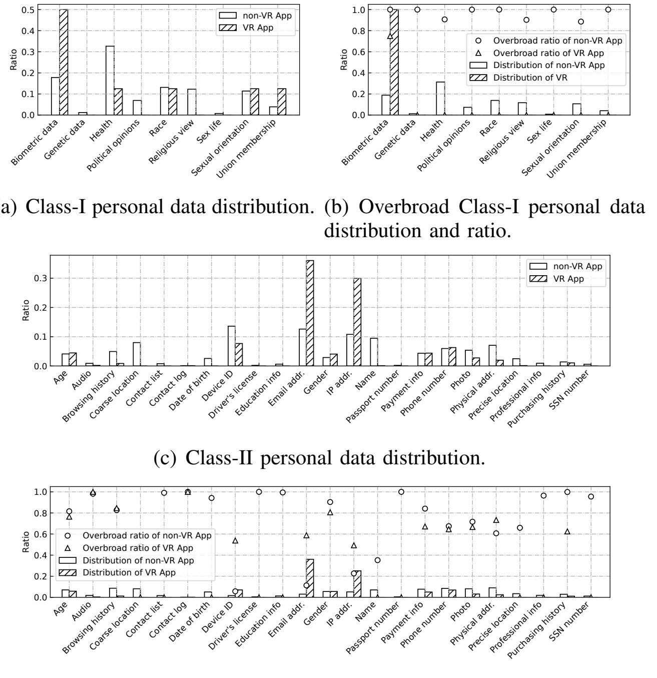
We chose Oculus Quest and its authorized third-party community, App Lab, as our target VR platform for the following reasons: 1) Quest, owned by Meta, occupies around 80% of the stand-alone consumer VR market share, and the VR app published there is required to append a privacy policy; 2) Quest has a close relationship with a third-party developer community, App Lab (in SideQuest), which enlarges the size of VR app samples from 390 to 1,726. We used WebScraper and Selenium to crawl privacy policy links as well as other helpful information (description, genres, publisher *etc.*) of every VR app automatically. After data cleaning, we obtained 1,703 samples in total.

### B. Counterpart-based model

Since there is no legal specification for *minimization*, we utilized the counterpart-based model from [4] to analyze the minimization principle of VR apps. This model first extracts data Collecting/Sharing Practices (CSPs) from the privacy policy of each app based on NLP tools like PoliCheck [5], then compares CSPs between a target app and its *counterparts*, *i.e.*, apps that have similar functionalities with the target app. The insight behind this method is that, if group of apps provide similar functionalities, then they are expected to collect similar scope of data to support these functionalities. Under this assumption, if a CSP exists in a target app but is not collected by more than half of its counterparts, then this

‡ Haojin Zhu is the corresponding author.

978-1-6654-9427-4/23/$31.00 ©2023 IEEE



(a) Class-I personal data distribution.

(b) Overbroad Class-I personal data distribution and ratio.

(c) Class-II personal data distribution.

(d) Overbroad Class-II personal data distribution and ratio.

Fig. 1. CSPs' distribution and overbroad ratio of VR / non-VR Apps.

CSP is regarded as *overbroad*. The key step of this method is to find proper counterparts for each target app. Hence we proposed a multi-sources similarity computing algorithm that integrates 1) recommendations from professional websites (like SteamPeek), 2) genres of apps, and 3) descriptions of apps, into consideration. We set a similarity threshold as a standard for *being similar* and choose top-11 similar apps as counterparts of the target app.

### C. Minimization principle analysis

For each CSP of the target app, we compared it with corresponding CSPs from its counterparts and then checked whether this CSP is overbroad to examine whether it potentially violates the data minimization principle. Like [4], data types are divided into two classes, Class I (highly sensitive in general, like biometric data), and Class II (relatively sensitive and can easily identify users in the physical world, like phone number). We conducted this analysis on our VR dataset and compare them with that of non-VR apps in [4].

## III. Result

We found counterparts for 1,351 VR apps and extracted 1,534 CSPs from 36.8% (497/1,351) VR privacy policies in total. Note that the performance of underlying NLP tools for extracting CSPs is degraded due to domain shift from non-VR apps to VR apps, and we will discuss this limitation in Section IV. Distribution of CSPs from VR / non-VR apps of Class-I and Class-II data types are shown in Fig1(a), 1(c) respectively. For Class-I CSPs, VR apps tend to collect more biometric data and trade union membership information, while non-VR apps tend to collect more health information, political opinions, and religious views from users. For Class-II CSPs, data collection of VR apps is centralized and distributed on the email address (36.0%) and IP address (29.9%).

Fig1(b) and 1(d) demonstrate the distribution and ratio of overbroad CSPs from VR / non-VR apps of Class-I and Class-II data types respectively. The overbroad ratio is defined as $\frac{\#\ overbroad\ CSPs}{\#\ total\ CSPs}$ for a certain data type. In total, 48.1% (239/497) VR apps are reported to have at least one overbroad CSP, hence potentially violating the data minimization principle. Surprisingly, biometric data is the only Class-I data type that has overbroad behaviors in VR apps. For Class-II overbroad CSPs, distribution and overbroad ratio of device ID (7.0% / 53.8%), email address (36.1% / 58.8%), and IP address (25.1% / 49.3%) in VR apps are significantly higher than that in non-VR apps. Though the distribution of overbroad physical address CSPs in VR apps is lower than that in non-VR apps, it has a relatively higher overbroad ratio (73.3%), which means if a VR app collects physical address, then it is more likely to be an overbroad CSP than in non-VR scenario.

## IV. Conclusion and future works

In this study, we conducted a large-scale privacy policies analysis towards data minimization principles of VR apps for the first time. Our initial results uncover the severe privacy policy compliance issues in the VR ecosystem as well as the bias of CSPs' distributions between VR and non-VR apps. Two directions could be considered for further research. First, how to maintain the high performance of underlying NLP tools when utilizing them in a new domain (*e.g.* from non-VR apps to VR apps) is still an open problem [6]. Second, more privacy laws and their principles for privacy policies should be involved to give a comprehensive compliance analysis of VR apps. We leave these issues as future works in this direction.

## Acknowledgment

This research was supported by the National Natural Science Foundation of China (No. 61972453 and 62132013).

## References

[1] J. A. De Guzman, K. Thilakarathna, and A. Seneviratne, "Security and privacy approaches in mixed reality: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019.

[2] D. Adams, A. Bah, C. Barwulor, N. Musabay, K. Pitkin, and E. M. Redmiles, "Ethics emerging: The story of privacy and security perceptions in virtual reality," in *14th USENIX SOUPS*, 2018, pp. 443–458.

[3] R. Trimananda, H. Le, H. Cui, J. T. Ho, A. Shuba, and A. Markopoulou, "OVRSEEN: Auditing network traffic and privacy policies in oculus VR," in *31st USENIX Security*, 2022, pp. 3789–3806.

[4] L. Zhou, C. Wei, T. Zhu, G. Chen, X. Zhang, S. Du, H. Cao, and H. Zhu, "POLICYCOMP: Counterpart comparison of privacy policies uncovers overbroad personal data collection practices," in *32th USENIX Security*, 2023.

[5] B. Andow, S. Y. Mahmud, J. Whitaker, W. Enck, B. Reaves, K. Singh, and S. Egelman, "Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with PoliCheck," in *29th USENIX Security*, 2020, pp. 985–1002.

[6] S. Manandhar, K. Kafle, B. Andow, K. Singh, and A. Nadkarni, "Smart home privacy policies demystified: A study of availability, content, and coverage," in *31st USENIX Security*, 2022, pp. 3521–3538.