AndroZoo: A Retrospective with a Glimpse into the Future

Marco Alecci
SnT, University of Luxembourg,
Luxembourg, Luxembourg
marco.alecci@uni.lu

Pedro Jesús Ruiz Jiménez
SnT, University of Luxembourg,
Luxembourg, Luxembourg
pedro.ruiz@uni.lu

Kevin Allix
Independent researcher,
Rennes, France
kallix@kallix.net

Tegawendé F. Bissyandé
SnT, University of Luxembourg,
Luxembourg, Luxembourg
tegawende.bissyande@uni.lu

Jacques Klein
SnT, University of Luxembourg,
Luxembourg, Luxembourg
jacques.klein@uni.lu

ABSTRACT

In 2016, we released AndroZoo, a continuously expanding dataset of Android applications that aggregates apps from various sources, including the official Google Play app market. As of today, AndroZoo contains approximately 24 million APK files, making it, to the best of our knowledge, the most extensive dataset of Android APKs accessible to the research community. Currently, an average of 500 000 APKs are downloaded every day, with our initial MSR paper counting more than 880 citations on Google Scholar.

Over time, we have consistently expanded AndroZoo, adapting to app markets' evolution and refining our collection process. Additionally, we have started collecting supplementary data that could be valuable for various Android-related research projects and that we are providing to users, such as app descriptions, upload dates, ratings, lists of permissions, and many other kinds of data.

This paper begins with a retrospective analysis of AndroZoo, offering statistics on APK files, apps, users, and downloads. Then, we introduce the new data we are releasing to users, offering insights and examples of their usage.

KEYWORDS

Android Applications, APK, Software Repository, Metadata

ACM Reference Format:
Marco Alecci, Pedro Jesús Ruiz Jiménez, Kevin Allix, Tegawendé F. Bissyandé, and Jacques Klein. 2024. AndroZoo: A Retrospective with a Glimpse into the Future. In *21st International Conference on Mining Software Repositories (MSR '24), April 15–16, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3643991.3644863

1 INTRODUCTION

AndroZoo is an Android app repository formally introduced to the research community through an MSR data showcase paper published in 2016 [3]. In December 2023, AndroZoo contains approximately 24 million APKs collected since 2013 from various markets. AndroZoo has been made available to the research community, aiming to offer the research community unrestricted, scalable access to an up-to-date collection of Android apps. One of our goals was to overcome common limitations faced by the research community, such as reliance on small and quickly aging fixed datasets whose statistical significance and access conditions are not always documented, which can lead to non-reproducible experiments and undetectable biased results.

Today, ❶ over 2000 researchers worldwide have requested access to AndroZoo, ❷ around 500 000 APKs are downloaded daily, and ❸ the MSR paper accounts for more than 880 Google Scholar citations, demonstrating both the impact of AndroZoo and its usefulness to the research community.

With this new paper, our contribution is twofold: First, we propose a retrospective of AndroZoo by providing statistics on APKs files, apps, users, and downloads (cf. Section 2). Second, we present the new pieces of information that are made available to users (cf. Section 3) and some use cases (cf. Section 4). Finally, we discuss possible limitations, followed by concluding remarks.

2 ANDROZOO STATISTICS

In this section, our goal is to present some interesting statistics regarding the evolution of AndroZoo since 2016, which marks the year of our initial paper release.

2.1 Number of APKs

As of December 2023, at the time of writing this paper, our dataset comprises 23 923 632 APKs, 448 TiB of memory. To the best of our knowledge, AndroZoo represents the most extensive dataset of Android APKs accessible to the research community.

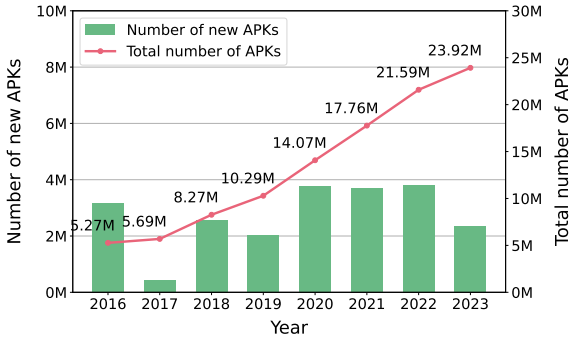

Figure 1: Number of APKs Overview

MSR '24, April 15–16, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0587-8/24/04
https://doi.org/10.1145/3643991.3644863

Figure 1 displays the number of APKs collected per year through green bars, while the total count of APKs within our dataset is represented by the red line. The number of APKs collected annually typically ranges between 2 million and 4 million, except for 2017, when we retrieved fewer than half a million apps due to organizational issues. In 2023, the decrease in collected apps could be attributed to fewer apps in the Google Play Store and the closure of one of the largest Chinese markets, namely *Anzhi*. As highlighted in our previous publication [3], the continuous expansion of our dataset makes it particularly suitable for evolutionary studies [6, 7, 9, 11, 13].

2.2 Number of Apps

An Android app can have multiple versions, each resulting in different APK files that share the same package name (e.g., com.android.chrome). AndroZoo stores multiple APK files for each app sharing the same package name, while keeping track of multiple versions stored [3]. This feature allows users to monitor the lineage of apps, representing the evolutionary history of a specific application.

Table 1: Top 10 apps by number of APKs

Package Name	#APKs
com.chrome.canary	1986
org.mozilla.fenix	1811
wp.wpbeta	910
dating.app.chat.flirt.wgbcv	826
com.blackforestapppaid.blackforest	822
com.brave.browser_nightly	787
com.topwar.gp	728
com.opodo.reisen	688
com.edreams.travel	679
com.styleseat.promobile	675

If we consider unique apps AndroZoo contains 8 708 304 apps with an average of 2.74 versions (and so APKs) for each app. Table 1 reports the top 10 apps by the number of collected APKs. The highest count belongs to `com.chrome.canary` (i.e., the alpha version of Google Chrome), for which we gathered a total of 1986 distinct APK files, representing various versions of the app.

In Table 2, we have assessed app lifespans within our dataset by calculating the duration between their oldest and newest APK creation dates. The table shows how many unique apps cover various spans of years with their APKs. Notably, around 10 000 apps have APKs spanning over 10 years, encompassing AndroZoo's entire lifespan. For instance, `com.whatsapp` is among the longest-lasting apps, spanning 10 years and 7 months from its initial APK in June 2013 to its latest in November 2023.

Table 2: Lifespan of apps in AndroZoo

#Years	#Apps	#Years	#Apps	#Years	#Apps
10	9347	6	37 099	2	315 206
9	20 072	5	84 931	1	432 536
8	20 171	4	108 962	0	2 732 016
7	37 378	3	186 800		

2.3 Number and Location of Users

At the time of writing this paper, we have distributed a total of 2158 API keys to researchers. Figure 2 provides an overview of the number of new users per year (displayed in green bars) alongside the total number of users with access to our dataset (illustrated by the red line). Seven years after its public release, more and more researchers request access to AndroZoo, with almost 400 new users in 2023.

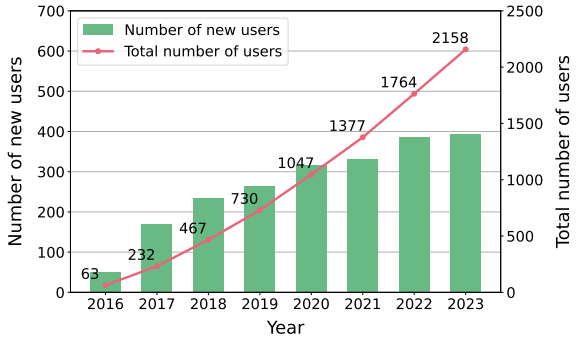

Figure 2: Number of Users Overview

Additionally, Figure 3 shows the geographical distribution of users who have access to AndroZoo, while Table 3 lists the top ten countries by the number of users. We linked user-requested institutional email suffixes to countries (e.g., ".it" for Italy). Common suffixes like ".org" or ".edu" were manually verified to identify the associated country. We have issued API keys to researchers from 86 countries, highlighting AndroZoo's global adoption as a standard tool among researchers.

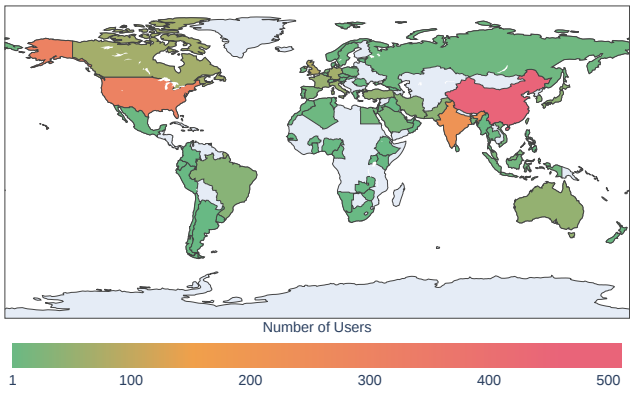

Figure 3: User Distribution across the world

2.4 Downloads

In this section, we present AndroZoo download statistics spanning from November 2021 to November 2023.

Throughout these two years, we received 365 604 948 download requests from 692 different users, resulting in a total of over 4 PiB

Table 3: Top 10 countries by number of users

Country	#Users	Country	#Users
China	511	Germany	68
USA	308	Italy	58
India	222	France	56
UK	86	Australia	51
Canada	69	Turkey	49

of data sent. Table 4 provides an overview of the statistics computed from our daily logs collection, categorized by both daily and monthly data. The first two rows display average data, specifically the number of downloads and download volume, while the last three rows show the highest recorded values for the number of downloads, download volume, and the count of active users.

On average, we receive approximately 500 000 download requests each day, totaling 5.8 TiB of data. These numbers underline the significant importance of AndroZoo to researchers, demonstrating its substantial daily usage within the research community. In terms of user activity, we peaked at 130 active users in a month, further highlighting the usage of AndroZoo as a standard tool in research.

Table 4: Download Statistics from 11-2021 to 11-2023

	Day	Month
Average Number of HTTP requests	502 083	15 393 045
Average Download Volume	5.8 TB	170 TB
Highest Number of HTTP requests	7 815 246	40 345 028
Highest Download Volume	31 TB	587 TB
Highest Number of Active Users	43	130

3 NEW METADATA WITHIN ANDROZOO

To enable new research investigations, we recently decided to make available new metadata within AndroZoo, in addition to the raw APK files. Starting in June 2020, our Google Play crawlers started to collect and save the metadata presented by Google. This metadata corresponds to what allows the official Google Play app to display information about apps. Some of that information is provided by developers (app name and description, developer name, and address), and some is provided by Google, like the upload date, the number of downloads, and the ratings.

Metadata has been collected from Google Play since June 2020, and amounts today to over 16 million records, providing details on over 4 million apps. Furthermore, our metadata covers over 13 million different versions, enabling a look back in time. For example, the 211 metadata entries about 132 versions of the app `com.snake.io.slither.worms` allow to follow the app from its creation in December 2020, with only a handful of downloads, to its latest version in December 2023, with over 100 million downloads.

Like the rest of AndroZoo, metadata are collected continuously, and several thousands of new records are obtained every day. Table 5 presents the number of metadata entries we collected per year.

Accessing the Data. The HTTP API, previously only available for downloading APKs from AndroZoo, can now also be used to download metadata entries, as detailed on the AndroZoo website. We also distribute the metadata using two files (that are regularly updated – currently every week):

- `gp-metadata-full.jsonl.gz` is a JSONLines file containing all the metadata crawled from Google Play, with an additional field `az_metadata_date` that denotes the date when this metadata was acquired. The file is 7.1 GB compressed as of December 2023.
- The file `gp-metadata-aggregate.jsonl.gz` is a JSONLines file described by the schema detailed on AndroZoo's website, containing metadata aggregated by package name. We have added new fields that consolidate multiple fields into a unified view for each package name. For instance, rather than specifying download count values for individual app versions within this file, we compute and present the minimum and maximum download counts for the entire app history. The file is 1.2 GB compressed as of December 2023.

Files and API documentation are both available on the following page:

https://androzoo.uni.lu/gp-metadata

While the complete documentation for the metadata is available on AndroZoo's website, we highlight some of the attributes found in the metadata below.

Rating. Metadata contain multiple fields related to the rating of an app, as evaluated by its users. These fields are: ❶ Rating; ❷ Number of N star ratings (N goes from 1 to 5); ❸ Number of ratings; and ❹ Number of comments. These fields could potentially be used to determine the quality of an APK (e.g., an APK with 100K ratings and a rating of 4.7 indicates a high-quality app).

Based on the maximum rating value computed in the aggregate file[1], we realized that 2 800 287 (66%) apps have never been rated, since their maximum rating is 0 (note that the minimum input for rating an app in Google Play is 1). Additionally, based on this maximum rating, about 25% of the apps have never reached a rating over 4.

Meanwhile, during the last two weeks, we collected 123 339 rating values. In 76 258 of the cases, the value is 0[2]. Figure 4 shows the distribution of the 47 081 remaining values. More than 50% of the rated apps have a rating over 4.

Number of downloads. Google only presents an approximation of the number of downloads of an app. This approximation uses the format N+ – meaning N times or more – to represent the number of downloads. This number is also quite limited, using only 0, 10^x, and $5 * 10^x$ (e.g., 0, 1, 5, 10, 50, 100, 500, etc.). Therefore, one application could have been downloaded 9M times, and Google Play

Table 5: Number of metadata entries collected by year

Year	# Metadata entries
2020	4 053 419
2021	4 469 732
2022	4 639 095
2023	2 887 427
Total	16 049 673

[1]This is the maximum rating we collected for a given package name.
[2]The rating of an app is 0 when the app has not been rated yet

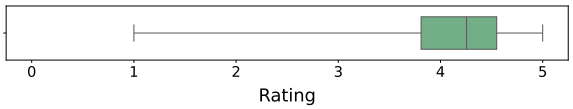

Figure 4: Rating Distribution of Rated Apps from 23-11-2023 to 06-12-2023

would report it has been downloaded 5M+ times, leaving out 4M downloads.

As we can observe in Figure 5, in our metadata, there are 209 616 apps (almost 5% of the apps) that have never been downloaded. Meanwhile, there are less than 100 apps with 1B+ downloads. Some examples of these extremely popular apps (with 10B+ downloads) are Google Chrome, Google Maps, Gmail, and YouTube.

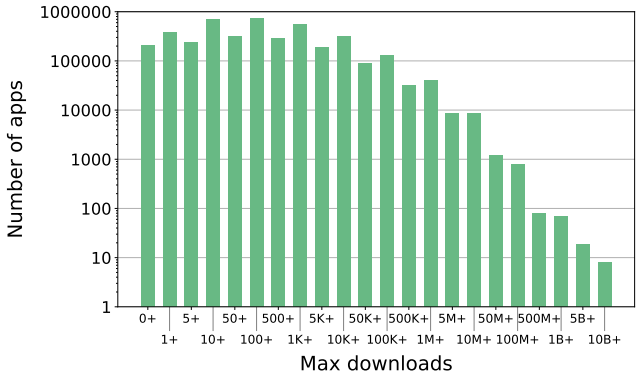

Figure 5: Number of Apps per Max Downloads.

Description. Two description fields are collected from Google Play: ❶ Description; and ❷ Short description. The description is a short text written by the developer(s) of an app to explain what an app does. According to Google's guide to create compliant app descriptions[3], the description of an app can not exceed 4000 characters, while the short description should summarize the description using 80 or fewer characters. While we collect descriptions written in English when available, many apps only have descriptions in another language.

Upload date. One of the problems found in AndroZoo was the lack of a reliable date on which the APKs were developed. We could not rely on the `dex_date` field since in November 2023, there are 8 697 841 APKs whose `dex_date` is set to the year 1980, 7 103 688 to 1981, and so on. The metadata made available solves this issue by introducing the field `uploadDate`, which is the reliable, Google-provided date this version of the app was uploaded to Google Play. In contrast to the `dex_date`, the first `uploadDate` found in AndroZoo is from January 2010, holding our statement that researchers can rely on the `uploadDate`.

Permissions. The list of permissions required by apps is used in many papers that study Android apps security.

Privacy policy link. Recently, there have been multiple works [10, 21] which explore the privacy policy of Android apps to determine whether their code complies with their policy. With this field, it becomes trivial to select apps whose developers have provided a link to an official privacy policy.

[3]https://support.google.com/googleplay/android-developer/answer/13393723?hl=en

4 USE CASES

The paper by Martin et al. [15] studies many of the available attributes of Android apps (from app stores), highlighting their potential usage for research. Adding the new metadata to AndroZoo will allow our dataset to back these studies by providing a large number of attributes that researchers could deem useful. Indeed, there are already some studies [19, 20] which used a combination of these newly available properties to determine whether an Android app is malicious or papers like the one by Ochiai et al. [17] which uses these attributes to classify Android apps. By including information about the description of Android apps in Google Play, our dataset could support papers like CHABADA [12] and AutoCog [19], where by comparing the app description against its API usage or its permissions, researchers were able to identify evident anomalies. Other works also heavily rely on app description [1, 2, 4], and thus could benefit from the new metadata AndroZoo provides. The availability of the list of declared permissions will ease the work of many papers [5, 14, 16, 19] which use this attribute. Additionally, the inclusion of the upload date (provided by Google) makes the results from evolutionary and longitudinal studies [6–9, 11, 13] more reliable, while potentially removing the bias highlighted in the paper by Pendlebury et al. [18], of models used to identify Android malware. Also, due to the recent applicability of GDPR[4] (2018), there have been some studies [10, 21] that use the privacy policy to determine if apps comply with these regulations. Providing the link to the privacy policy will facilitate the work of researchers.

5 LIMITATIONS

Over time, collecting metadata is faced with the same limitations outlined in our first paper [3] such as the impossibility to enumerate all apps available in Google Play, and therefore to collect all data, or even compute the fraction of all available data that we did collect.

The metadata we collect is obtained through an independent implementation of the protobuf [5] protocol used between the official Google Play app and Google's servers. Due to the lack of documentation of this protocol, the implementation we rely on is missing several fields, thus having several raw pieces of data we cannot derive the meaning of. Nonetheless, the vast majority of the fields are available and self-explanatory, and they already provide a trove of useful data.

Finally, the metadata discussed in Section 3 has inherent limitations, such as the inability to provide precise download numbers for each app, offering only an estimation. However, these constraints are intrinsic to Google Play and are beyond our control.

6 CONCLUSIONS

In this paper, we have offered a retrospective overview of AndroZoo, presenting statistics on APKs, apps, users, and downloads. This look back underscores both the importance of AndroZoo

[4]https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en

[5]https://protobuf.dev

and its peculiarities: A long-lived, still evolving dataset that, seven years after its public release, is now more popular in the research community than ever.

We then presented the new data that we provide to AndroZoo users, such as descriptions, upload dates, ratings, permission lists, and many others, providing some examples of use cases. This metadata addition enables AndroZoo to not only provide raw APKs, but also to offer the data necessary to select relevant objects of study in an otherwise overwhelmingly large dataset. It also significantly enhances the capacity to study the evolution of apps and their adoption over significant periods of time.

We hope that AndroZoo, augmented with this new data, can further contribute to ongoing research and enable the exploration of new potential research topics on Android apps, and further improve the reproducibility of Research on Android Apps. All the details regarding AndroZoo and its usage, are available at:

https://androzoo.uni.lu

7 ACKNOWLEDGEMENT

This research was funded in whole, or in part, by the Luxembourg National Research Fund (FNR), grant references REPROCESS Project (16344458) and UNLOCK Project (18154263).

REFERENCES

[1] A. A. Al-Subaihin, F. Sarro, S. Black, L. Capra, M. Harman, Y. Jia, and Y. Zhang. 2016. Clustering Mobile Apps Based on Mined Textual Features. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Ciudad Real, Spain) *(ESEM '16)*. Association for Computing Machinery, New York, NY, USA, Article 38, 10 pages. https://doi.org/10.1145/2961111.2962600

[2] Marco Alecci, Jordan Samhi, Tegawendé F Bissyandé, and Jacques Klein. 2024. Revisiting Android App Categorization. *46rd International Conference on Software Engineering (ICSE), IEEE/ACM.*

[3] Kevin Allix, Tegawendé F. Bissyandé, Jacques Klein, and Yves Le Traon. 2016. AndroZoo: Collecting Millions of Android Apps for the Research Community. In *Proceedings of the 13th International Conference on Mining Software Repositories* (Austin, Texas) *(MSR '16)*. ACM, New York, NY, USA, 468–471. https://doi.org/10.1145/2901739.2903508

[4] Afnan Alsubaihin, Federica Sarro, Sue Black, and Licia Capra. 2019. Empirical comparison of text-based mobile apps similarity measurement techniques. *Empirical Software Engineering* 24 (12 2019). https://doi.org/10.1007/s10664-019-09726-5

[5] Daniel Arp, Michael Spreitzenbarth, Malte Hübner, Hugo Gascon, and Konrad Rieck. 2014. DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket. *Symposium on Network and Distributed System Security (NDSS)*. https://doi.org/10.14722/ndss.2014.23247

[6] Haipeng Cai. 2020. Assessing and Improving Malware Detection Sustainability through App Evolution Studies. *ACM Trans. Softw. Eng. Methodol.* 29, 2, Article 8 (mar 2020), 28 pages. https://doi.org/10.1145/3371924

[7] Haipeng Cai, Xiaoqin Fu, and Abdelwahab Hamou-Lhadj. 2020. A study of runtime behavioral evolution of benign versus malicious apps in android. *Information and Software Technology* 122 (2020), 106291. https://doi.org/10.1016/j.infsof.2020.106291

[8] Haipeng Cai and Barbara Ryder. 2020. A longitudinal study of application structure and behaviors in android. *IEEE Transactions on Software Engineering* 47, 12 (2020), 2934–2955.

[9] Paolo Calciati and Alessandra Gorla. 2017. How Do Apps Evolve in Their Permission Requests? A Preliminary Study. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*. 37–41. https://doi.org/10.1109/MSR.2017.64

[10] Cheng Chang, Huaxin Li, Yichi Zhang, Suguo Du, Hui Cao, and Haojin Zhu. 2019. Automated and personalized privacy policy extraction under GDPR consideration. In *Wireless Algorithms, Systems, and Applications: 14th International Conference, WASA 2019, Honolulu, HI, USA, June 24–26, 2019, Proceedings 14*. Springer, 43–54.

[11] Jun Gao, Li Li, Pingfan Kong, Tegawendé F. Bissyandé, and Jacques Klein. 2021. Understanding the Evolution of Android App Vulnerabilities. *IEEE Transactions on Reliability* 70, 1 (2021), 212–230. https://doi.org/10.1109/TR.2019.2956690

[12] Alessandra Gorla, Ilaria Tavecchia, Florian Gross, and Andreas Zeller. 2014. Checking app behavior against app descriptions. In *Proceedings of the 36th international conference on software engineering*. 1025–1035.

[13] Dongjie He, Lian Li, Lei Wang, Hengjie Zheng, Guangwei Li, and Jingling Xue. 2018. Understanding and Detecting Evolution-Induced Compatibility Issues in Android Apps. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering* (Montpellier, France) *(ASE '18)*. Association for Computing Machinery, New York, NY, USA, 167–177. https://doi.org/10.1145/3238147.3238185

[14] Naveen Karunanayake, Jathushan Rajasegaran, Ashanie Gunathillake, Suranga Seneviratne, and Guillaume Jourjon. 2020. A multi-modal neural embeddings approach for detecting mobile counterfeit apps: A case study on Google Play store. *IEEE Transactions on Mobile Computing* 21, 1 (2020), 16–30.

[15] William Martin, Federica Sarro, Yue Jia, Yuanyuan Zhang, and Mark Harman. 2017. A Survey of App Store Analysis for Software Engineering. *IEEE Transactions on Software Engineering* 43, 9 (2017), 817–847. https://doi.org/10.1109/TSE.2016.2630689

[16] Annamalai Narayanan, Charlie Soh, Lihui Chen, Yang Liu, and Lipo Wang. 2018. Apk2vec: Semi-Supervised Multi-view Representation Learning for Profiling Android Applications. In *2018 IEEE International Conference on Data Mining (ICDM)*. 357–366. https://doi.org/10.1109/ICDM.2018.00051

[17] Keiichi Ochiai, Fatina Putri, and Yusuke Fukazawa. 2019. Local App Classification using Deep Neural Network based on Mobile App Market Data. 186–191. https://doi.org/10.1109/PERCOM.2019.8767416

[18] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. 2019. TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time. arXiv:1807.07838 [cs.CR]

[19] Zhengyang Qu, Vaibhav Rastogi, Xinyi Zhang, Yan Chen, Tiantian Zhu, and Zhong Chen. 2014. Autocog: Measuring the description-to-permission fidelity in android applications. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 1354–1365.

[20] Andrea Saracino, Daniele Sgandurra, Gianluca Dini, and Fabio Martinelli. 2016. Madam: Effective and efficient behavior-based android malware detection and prevention. *IEEE Transactions on Dependable and Secure Computing* 15, 1 (2016), 83–97.

[21] Anhao Xiang, Weiping Pei, and Chuan Yue. 2023. PolicyChecker: Analyzing the GDPR Completeness of Mobile Apps' Privacy Policies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (Copenhagen, Denmark) *(CCS '23)*. Association for Computing Machinery, New York, NY, USA, 3373–3387. https://doi.org/10.1145/3576915.3623067