1. Terms and Conditions for the use of MAL2 datasets

Introduction
The MAchine Learning detection of MALicious content (MAL2) project applies Deep Neural Networks and Unsupervised Learning in two specific scenarios: a) advance discovery of fraudulent eCommerce and b) evaluate the capabilities of detecting Potentially Harmful Apps in Android. Currently there is a lack of technology supporting an integrated solution of large-scale feature extraction and Neural Network training. The goal of the MAL2 project is (i) to release Open Source framework which provides integrated functionality along the required pipeline – from data extraction, feature composition up to Neural Network Training and analysis of results (ii) to execute its components at large-scale within Hadoop and GPU cluster support (iii) to publish the harvested Ground-Truth dataset, the extracted features as well as the trained Neural Network in both application domains on open data platforms. To visualize the projects results and to raise awareness for cybercrime prevention in the general public, two demonstrators are deployed at Watchlist Internet that allow live-inspection on the trustworthiness of eCommerce sites and Android Apps.

The MAL2 project is supported by funding from the federal ministry BMK within the 6th call for proposals in ICT of the Future 2017 by the Austrian national funding agency FFG.
https://projekte.ffg.at/projekt/3044975

Project coordination:
AIT Austrian Institute of Technology GmbH
Project partners:
Kuratorium Sicheres Österreich (KSÖ)
X-Net Services GmbH
Österreichisches Institut für angewandte Telekommunikation (ÖIAT)
IKARUS Security Software GmbH
https://malzwei.at

Release of MAL2 Ground-Truth datasets (deliverable D2.1). To support scientific and industrial research activities and to advance cybercrime prevention two well curated and large-scale datasets are released by the MAL2 project in both application domains, free to use for scientific and academic noncommercial purposes after approval by the MAL2 consortium. The dataset includes

a) 861 GB of Android Malware Samples (APK files) and 1.9 GB of extracted features
b) 20 GB of archived Fake-Shops (html, css, javascript) and 345 MB of features and metadata

A dedicated Ground-Truth data composition task within the project made sure to assure the quality, and coverage of the collected samples. If you find your application or website listed and disagree with the expert evaluation, please feel free to contact us at https://malzwei.at/kontakt

a) MAL2 Android Malware Samples Dataset
A total of 150 Thousand Apps was expected to be needed for achieving a meaningful and comprehensive evaluation within the Android Malware detection scenarios of MAL2. The MAL2 Android Malware Ground-Truth data set was developed in two iterations. The first iteration was kept small with 56,392 APKs and was used to test the proof-of-concept prototype in the project. 45,676 APKs (of which a total of "Benign" 27,965) were used as training, 5,076 as test and 5,640 as validation data sets. This contains samples from 430 different PUA families and 25 Trojan families. In the final iteration, 790 thousand APK data records consisting of malware, adware, probably clean and Google Play samples were selected, and their correct assignment was verified using the IKARUS scanner. As Malware is not protected by IPR, this subset of data is publicly released as Ground-Truth dataset by the MAL2 project for scientific research purposes and includes 861 GB of Malware Samples, collected by IKARUS between 2017 and 2018 as well as 1.9 GB of extracted features using the DREBIN framework for constructing the labelled Ground-Truth. The consortium has undergone a best-effort approach of removing all IPR related material, but is not liable for application samples or other IPR protected sources which possibly were included by incident within the dataset..

The dataset is structured as follows: Each included file contains a metadata representation which is associated by the malwares APK name and each row in the metadata contains a textual representation of the string pattern found in the after disassembling the file which was extracted by using the developed MAL2 framework.

b) MAL2 Fraudulent eCommerce 'Fake-Shop' Dataset

Consumers report suspected suspicious cases of fraudulent online ecommerce sites to Watchlist Internet and are entered into a fake-shop database developed within the MAL2 project. The experts of ÖIAT subsequently manually review these samples in a timely manner by using a standardized multi-stage validation process. Suspicion raised by consumers were confirmed in the course of the MAL2 project in 96% of the cases (2,814 samples in 2019). The evaluation of the tool-based evaluation process shows that 85% of the fake shops and 83% of the counterfeit shops were clearly identified as such after stage one of the process ("online research"). A further 13% and 16%, respectively, followed in stage two "payment options". In stage three "review of the imprint information" additional 1.8% and 0.3% were confirmed. No corpus on fraudulent eCommerce shops existed prior the project. Therefore, an ongoing initiative by the MAL2 project was established to scrape and archive a corpus dataset of fraudulent ecommerce website-archives during the duration of the first year of the project and to apply manual annotations and a standardized checklist to establish a Ground-Truth for evaluating the performance of the trained machine learning detection models. The scraping of the website data is based on the open source web crawling framework Scrapy and is written in Python. The results of the expert-based fake shop evaluation process are part of the fake shop data set of 3480 samples (20 GB of data) which were collected during 2019.

The dataset is structured as follows: A fake-shop dataset contains the scraped html, css and javascript of the sites main page. Within the features.csv file each column in the table represents a feature and each row represents a URL of the associated fake-shop dataset. The URL name is the last column and the extracted feature names are contained in the first row of the file. The values contained in the file are numerical representations of the string patterns found in the html, css and js files pertaining to each URL.The numerical representation is calculated using tf-idf (term frequency inverse document frequency), which is a numerical statistic that reflects how important a string is to a datapoint in a dataset.

All information on requesting access to the dataset are presented online at www.malzwei.at

Please be sure to state your institutions affiliation and research topic in a manner which allows the consortium of the MAL2 project to understand and evaluate the purpose you intend to use the data for. After a positive evaluation you will receive access through a tokenized download link and you agree that we may list your institutions name and logo on our site. The following terms and conditions apply.

1. Definition

1.1. Agreement shall mean these Terms and Conditions for the MAL2 datasets (see description above)

1.2. Data means up to ~861 GB of Android Malware Samples (APK files) and ~20 GB of scraped Fake-Shop websites (html, css, js files) including their corresponding metadata of extracted features and labelled Ground-Truth as CSV records.

1.3. User is the natural person who is granted personal, token-protected access to that Data

1.4. Research Institute means the legal party the User represents who is a research organization according to EC 198/2014/01

1.5. We or us are the parties of the project as listed in the grant agreement of the FFG project MAL2[1] as a group as well as any party of the MAL2 project individually, depending on the meaning in the Agreement.

2. Accepting these Terms and Conditions

2.1. The User shall read this Agreement carefully before using the Data provided by us. The User accepts this Agreement by clicking "I accept the terms of the agreement" box where this option is made available to the User at malzwei.at. The User provides his data, the scope and intent output the Data is planned to be used for as well as information on the affiliation with the Research Institute he represents through the foreseen web-form.

[1] https://projekte.ffg.at/projekt/3044975

2.2. Through clicking on the “I accept the terms of this agreement” box the User agrees to be bound by this Agreement and binds the legal party he represents. Furthermore the User warrants that he/she has the full legal authority to bind the legal party he represents to this Agreement. If he/she does not have the requisite authority, he/she may not accept the Agreement on behalf of the legal party mentioned on the web-form.

3. Scope

The following term of use shall apply to using the Data. The use of the Data is only permitted if the User in the name of his Research Institute accept these Terms and Conditions as described in section 1 and 2.

4. Registration/access

4.1. To receive access to the Data of Mal2 a prior submission on malzwei.at via a web-form is required. The first and last name and email address of the User as well as the name and some details of the research institution and intended research scope and data usage are to be given. Based on the information provided the MAL2 consortium will evaluate each case individually, and cast a vote based on the consortium agreement. The requestee is notified and will receive a token-based download link to the data after a positive evaluation. The User hereby agrees that the MAL2 project may list the institutions logo and name on the website.

4.2. Access as a User can be only granted to natural person.

4.3. The access details for the requested Data are only granted and directly linked to the person that has applied for. The access link and token may not be disclosed or shared with any other natural person, co-workers or a third party. The sharing of such an access token will be considered a material breach of the Agreement. The downloaded Data however may be distributed to co-workers of the same institution that are considered jointly working on the specified research topic as long as the affiliation of the User and scope of research remains.

4.4. If details of the Users or his affiliation, relationship to the Research Institute he registered for change, we have to be informed immediately of the change through the online contact form provided on the malzwei.at website.

5. Data provided by us

5.1. We will grant the Research Institution access to the requested Data in form of compressed tar archives, with the usual limits of accessibility of our servers and available bandwidth, free of charge. **We retain the right to change the Data as well as modalities for the use thereof in the future upon prior and timely adequate notice of all affected Users**. We will try to maintain access to the Data during the cause of the MAL2 project, but the Research Institution and the User have no claim to a permanent and continuous availability of the Data.

5.2. **The Research Institute acknowledges and agrees the Data provided under this Agreement is expressly provided on an “as is” basis with no warranties, including, but not limited to, any implied warranties arising out of any course of dealing, custom or usage of trade, merchantability, fitness for particular purpose or non-infringement of the right of third parties.**

5.3. The User is aware that part of the Data contains Malware and therefore the use and storage of the Data is in the sole responsibility of the User and his/her Research Institute, especially also in regards to applicable export regulations.

5.4. For the avoidance of doubt, we are not obligated to provide these Data. We retain the right, to immediately prevent or restrict access to the Data or parts thereof or take any other action as necessary in case of technical problems, infringing or objectionable material, inaccurate listings, or any other action or prohibition infringing applicable law or the MAL2 projects aim or for any other reason in the sole and absolute discretion of us. Therefore, any Data can be deleted by us immediately.

5.5. Should the User have any problems accessing, the Data, he can reach us at the contact form on malzwei.at. We will contact him as soon as possible.

6. Use of the Data by Research Institute

6.1. The Research Institute agrees not to do any of the following: (i) use the data for any other purpose than the intended research scope and purpose than communicated in the web-form on which the request was granted for and approved by the MAL2 consortium, (ii) sell, rent or otherwise sub-data (iii) duplicate, copy or otherwise exploit the data for a commercial or (v) edit or otherwise modify any Data without our consent (iv) publish results based on the data without naming the MAL2 project by either a link to the website or citing one of the MAL2 projects publications.

7. Indemnity

7.1. Research Institute agrees to indemnify, defend and hold us harmless from and against any claims, costs, liabilities and expenses – including reasonable attorneys 'fees- paid or payable to an third party arising from (i) Research Institute or Users breach of this Agreement, (ii) any claim that Research Institute or Users has infringed another's intellectual property rights through the use of the Data (iii) any violation of applicable law by Research Institute or User (iv) any violation of applicable law through the downloading or publishing of the Data.

7.2. As the Data is provided free of charge, we will not be liable for any loss or damages of any nature. We will not be liable for any consequential, claims, indirect or special loss or damage though the use of the Data. We will make every effort to ensure that the Data is free from viruses or defects except those listed expressly in the Data; however, we cannot guarantee that the use of the Data will not cause damage to the end device that is used by User to access the Data or that the Data does not infringe the rights of third parties.

8. Data protection

We will treat all personal data of the User in responsible manner. We will use, store and process the data resulting from the application form only for the purpose or this Agreement and treat is as confidential in line with the provisions of the applicable data protection laws. For the management of the access, it is necessary to store the names and email addresses of contact persons provided by User. Further the Research Institutes Logo and Name will be listed on malwzei.at. By agreeing to this Agreement, the User agrees to their name and email address being visible to us. The User may revoke his consent through a message through the project's contact form. In such a case we will delete all access to the Data and User and his/her Research Institute will be precluded from using the Data or results based on that Data.

9. Termination

This Agreement shall terminate immediately in case of breach of this Agreement by User or Research Institute. Should a Research Institute choose to discontinue using and delete all Data provided by us.

We can terminate this Agreement at any time trough information to the email address of the User.

10. Miscellaneous

10.1. This document comprises any and all agreements entered into by the Research Institute and us. There are no written or oral ancillary agreements. We reserve the right, at our sole discretion, to modify or replace this Agreement, or change, suspend, or discontinue all or parts of Data at any time by posting a notice on the projects website malzwei.at or by sending the User an email. It is the Users responsibility to check this Agreement periodically for changes. Users and Research Institute continued use of the Data following the postings of any changes to this Agreement constitutes acceptance of those changes.

10.2. All disputes or claims arising out of or in connection with this Agreement including disputes relating to its validity, breach, termination or nullity shall be finally settled under the Rules of Arbitration of the International Arbitral Centre of the Austrian Federal Economic Chamber in Vienna (Vienna Rules) by one or three arbitrators appointed in accordance with the said Rules. The provisions on expedited proceedings are applicable. The number of arbitrators shall be one. The substantive law of Austria shall be applicable under exclusion of the United Nations

Convention on Contracts for the International Sale of goods, 1980. The language to be used in the arbitral proceedings shall be English.

10.3. The Research Institute irrevocably waive any objection which he or she might at any time have towards the International Arbitral Centre of the Austrian Federal Economic Chamber in Vienna, being nominated as the forum to hear and determine any proceedings and to settle any disputes and agree not to claim that the courts of Vienna are not convenient or appropriate forum.

10.4. Should any provisions of this Agreement be or become wholly or partly invalid or unenforceable, this shall not affect the validity or enforceability of the remaining provisions. In this event, the invalid or unenforceable provision shall be substituted by such valid/enforceable provision, which comes as close as possible to the legal and economic purposes pursued by MAL2 with such invalid/unenforceable provision.

10.5. This Agreement shall be governed in its entirety by the laws of the Republic of Austria excluding any legal norms referring to other legal systems. This includes disputes on its conclusion, binding effect, amendment and legal consequences of this agreement.