Datasheet Template

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

With the rapid emergence of graph representation learning, the construction of new large-scale datasets are necessary to distinguish model capabilities and accurately assess the strengths and weaknesses of each technique. By carefully analyzing existing graph databases, we identify 3 critical components important for advancing the field of graph representation learning: (1) large graphs, (2) many graphs, and (3) class diversity. To date, no single graph database offers all of these desired properties. We introduce MalNet, the largest public graph database ever constructed, representing a large-scale ontology of software function call graphs. MalNet contains over 1.2 million graphs, averaging over 17k nodes and 39k edges per graph, across a hierarchy of 47 types and 696 families. Compared to the popular REDDIT-12K database, MalNet offers **105× more graphs**, **44× larger graphs** on average, and **63× more classes**. We provide a detailed analysis of MalNet, discussing its properties and provenance, along with the evaluation of state-of-the-art machine learning and graph neural network techniques. The unprecedented scale and diversity of MalNet offers exciting opportunities to advance the frontiers of graph representation learning—enabling new discoveries and research into imbalanced classification, explainability and the impact of class hardness. The database is publicly available at www.mal-net.org.

B. What (other) tasks could the dataset be used for?

The dataset could be user for (1) graph representation learning, (2) graph classification, (3) explainable graph classification, and (4) imbalanced graph classification.

C. Who funded the creation dataset?

This work was in part supported by NSF grant IIS-1563816, CNS-1704701, GRFP (DGE-1650044) and a Raytheon research fellowship.

D. Any other comment?

We want to thank Kevin Allix and AndroZoo colleagues for generously allowing us to use their data in this research.

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Each instance is a graph containing nodes representing functions and directed edges representing calling relationships. Each graph contains two labels, a high level "type" label and a lower-level "family" label. In total, there are 47 type and 696 family labels.

B. How many instances are there in total (of each type, if appropriate)?

There are 1,262,024 function call graphs (FCGs) across 47 types and 696 families of malware. The distribution of FCGs can be found in the appendix.

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance is a graph written in edge list form. Each edge list is stored hierarchically, according to its type and family label.

D. Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance contains 2 labels—a family and type label.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The only relationship between individual instances is the shared family and type labels.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

As the number of software applications and malware instances is incredibly large, the database is a sample of the larger population. Leveraging the AndroZoo repository of benign and malicious Android software, we process and analyze APK files containing both type and family labels.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes, we provide recommended training/validation/test splits. We divide MalNet-Graph into three stratified sets of data: training, validation and test, with a split of 70/10/20, respectively; repeated for graph *type*, *family* and MALNET-TINY labels.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Function call graphs are assigned a general *type* (e.g., Worm) and specialized *family* label (e.g., Spybot) using the Euphony [1] classification structure. To generate these labels, Euphony takes a VirusTotal [2] report containing up to 70 labels across a variety of antivirus vendors and unifies the labeling process by learning the patterns, structure and lexicon of vendors over time. While Euphony provides state-of-the-art performance, this task is considered an open-challenge due to both naming disagreements [3], [4] and a lack of adopted naming standards [1] across vendors. To help address this issue, we collect and release the raw VirusTotal reports containing up to 70 antivirus labels for each graph.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and does not rely on external resources other than the hosted website run through Georgia Tech.

Any other comments? N/A

III. COLLECTION PROCESS

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

With the generous permission of the AndroZoo repository [5], [6], we collected 1,262,024 Android APK files, specifically selecting APKs containing both a *family* and *type* label obtained from the Euphony classification structure [1]. This process took about a week to download and 10TB in storage space when using the maximum allowed 40 concurrent downloads. In addition, we spent about 1 month collecting raw VirusTotal (VT) reports to release with MALNET, through VT's academic access, which allows 20k queries per day. Each VT report contains up to 70 antivirus labels per graph.

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The function call graphs were directly derived from the Android APK files obtained from the AndroZoo repository. The labels for each function call graph were obtained using the Euphony classification structure.

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset was obtained by processing all APK files with known type and family labels in the AndroZoo repository.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The authors of the paper were the only ones involved in the data collection process.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

It took approximately 1 month to collect all of the data.

IV. DATA PREPROCESSING

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

We leave each graph in its original state—retaining its edge directionality, disconnected components and node isolates (i.e., single nodes with no incident edges). Since we are dealing with highly malicious software, our goal is to mitigate the risk of releasing information that could potentially be used to reverse engineer malware. Thus, we numerically relabel the nodes of each graph, removing any associated attribute information. We use the type and family labels obtained from the AndroZoo repository.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, we spent about 1 month collecting raw VirusTotal (VT) reports to release with every graph instance, through VT's academic access, which allows 20k queries per day. Each VT report contains up to 70 antivirus labels per graph.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Yes, it does achieve the goal originally set out.

E. Any other comments

N/A

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

As the dataset requires over 443 GB in storage space, traditional data hosting was not an option. To accommodate the large dataset size, we setup a website and file storage server leveraging Georgia Tech's internal infrastructure.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

We release the dataset with a CC-BY license

C. Are there any copyrights on the data?

No.

D. Are there any fees or access/export restrictions?

No.

E. Any other comments?

No.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The lab group from Georgia Institute of Technology—Polo club of Data Science.

B. Will the dataset be updated? If so, how often and by whom?

As needed.

C. How will updates be communicated? (e.g., mailing list, GitHub)

On the website news section.

D. If the dataset becomes obsolete how will this be communicated?

On the website's news section.

E. Is there a repository to link to any/all papers/systems that use this dataset?

Not at the moment.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

There is no mechanism to currently do this. As the dataset is over 400 GB, this is a significant challenge.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

N/A

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

N/A

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

M. Any other comments?

No.

REFERENCES

[1] Médéric Hurier, Guillermo Suarez-Tangil, Santanu Kumar Dash, Tegawendé F Bissyandé, Yves Le Traon, Jacques Klein, and Lorenzo Cavallaro. Euphony: Harmonious unification of cacophonous anti-virus vendor labels for android malware. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 425–435. IEEE, 2017.

[2] Virus Total. Virustotal-free online virus, malware and url scanner. *Online: https://www. virustotal. com/en*, 2012.

[3] Médéric Hurier, Kevin Allix, Tegawendé F Bissyandé, Jacques Klein, and Yves Le Traon. On the lack of consensus in anti-virus decisions: Metrics and insights on building ground truths of android malware. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 142–162. Springer, 2016.

[4] Alex Kantchelian, Michael Carl Tschantz, Sadia Afroz, Brad Miller, Vaishaal Shankar, Rekha Bachwani, Anthony D Joseph, and J Doug Tygar. Better malware ground truth: Techniques for weighting anti-virus vendor labels. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, pages 45–56, 2015.

[5] Kevin Allix, Tegawendé F Bissyandé, Jacques Klein, and Yves Le Traon. Androzoo: Collecting millions of android apps for the research community. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pages 468–471. IEEE, 2016.

[6] Li Li, Jun Gao, Médéric Hurier, Pingfan Kong, Tegawendé F Bissyandé, Alexandre Bartel, Jacques Klein, and Yves Le Traon. Androzoo++: Collecting millions of android apps and their metadata for the research community. *arXiv preprint arXiv:1709.05281*, 2017.