New Techniques in Profiling Big Datasets for Machine Learning with A Concise Review of Android Mobile Malware Datasets

Gürol Canbek
Middle East Technical University
Informatics Institute
Ankara, Turkey
gurol@canbek.com

Seref Sagiroglu
Gazi University
Computer Engineering Department
Ankara, Turkey
ss@gazi.edu.tr

Tugba Taskaya Temizel
Middle East Technical University
Informatics Institute
Ankara, Turkey
ttemizel@metu.edu.tr

***Abstract*—As the volume, variety, velocity aspects of big data are increasing, the other aspects such as veracity, value, variability, and venue could not be interpreted easily by data owners or researchers. The aspects are also unclear if the data is to be used in machine learning studies such as classification or clustering. This study proposes four techniques with fourteen criteria to systematically profile the datasets collected from different resources to distinguish from one another and see their strong and weak aspects. The proposed approach is demonstrated in five Android mobile malware datasets in the literature and in security industry namely Android Malware Genome Project, Drebin, Android Malware Dataset, Android Botnet, and Virus Total 2018. The results have shown that the proposed profiling methods reveal remarkable insight about the datasets comparatively and directs researchers to achieve big but more visible, qualitative, and internalized datasets.**

Keywords—data profiling, data quality, big data, malware detection, mobile malware, machine learning, classification, Android, feature engineering

I. INTRODUCTION

A dataset is actually the main input solely characterizing the problem domain in artificial intelligence especially in machine learning (ML) studies such as classification or clustering. The ML models or algorithms and the approach to establishing those models are domain independent. For example, when we use support vector machines and/or decision trees to establish a mobile malware detection classifier (i.e. labeling a given instance of Android mobile application file as malign or benign software), obviously we need sufficient Android mobile malware datasets.

However, researchers who are concentrated on building a high-performing classifier by trying different feature sets with different ML models along with different hyper-tuned parameters may pay little attention in the sufficiency of the datasets used. Giving more importance to modeling may be due to the challenge in the application of a rather new machine learning algorithm (e.g. a deep-learning classifier) in a specific domain. Giving less importance to datasets may be due to the fact that researchers do not enjoy dataset activities while they take the most of their time [1]. After all, accessing to quality data and improving the quality datasets are the biggest bottleneck in successful artificial intelligence initiatives [1].

Both assessing the quality of the obtained datasets and improving the quality of the available datasets require methods evaluating the quality in datasets. This is more critical with respect to big data reality where the volume (e.g. the number of samples), variety (e.g. curse of dimensionality), and velocity (e.g. the number of new samples) are increasing [2]. The researchers overwhelmed by these 3-Vs, could not easily discern the other aspects such as veracity (e.g. the level of representativeness in problem domain), value (e.g. contribution to achieving high-performance), variability (e.g. freshness of the samples), and venue (e.g. the source of the samples).

Dataset profiling can be an effective method to realize these aspects at the forefront systematically. If a dataset is profiled from different characteristics, it is possible to maintain a qualitative dataset by merging other datasets or samples in those datasets according to these characteristics. Profiling also makes conducting initial activities much more convenient and motivates the researchers to proceed further. In this study, we proposed and demonstrated four techniques that can be conducted to gain more insight in a systematic manner:

- Basic profiling
- Timeline profiling
- Duplicate samples profiling
- Density/sparsity profiling

The literature addressed quality of information rather earlier. Lee *et al.* approaches to information quality from intrinsic (e.g. accuracy, consistency), contextual (e.g. relevance, completeness), representational (e.g. understandable, format), accessibility (e.g. availability, security) dimensions. Leaving aside the statistical methods such as individual descriptive statistics (e.g. record count, null value percentage, distinct value percentage, most frequent values), the literature covers the quality of datasets from specific domain's perspective such as the quality in spatial datasets [3], health-data [4], natural disasters [5]. Data quality is also studied from the requirements of data mining from architectural perspectives such as data lakes and polystores [6].

Profiling is describing a data or dataset from a specific perspective. In relational database management system perspective, inclusion dependencies or functional dependencies are mostly derived from relational database profiling beyond primary and/or foreign key relationships [7]. Systematic data profiling is achieved by the third-party tools running SQL (structured query language) queries according to user's interest and returns the metadata, which can be used in database query optimization, data cleansing, data integration [7]. Ellefi *et al.* compile the semantic, qualitative, statistical, and temporal (i.e. dynamicity) profiling attributes

978-1-7281-0472-0/18/$31.00 ©2018 IEEE

of the data on the web, most of which based on information quality dimensions summarized above [8]. The quality of datasets in machine learning is not addressed sufficiently in the literature. Sessions and Valtorta, observe the obvious degrading effects of different combinations of accurate and inaccurate data on a Bayesian network [9]. Today, the dataset repositories such as OpenML and Kaggle uses various statistical attributes to describe each dataset they store. But, the attributes are numeric and hard to interpret and compare.

The rest of the paper is organized as follows. Section II introduces the example domain Android mobile malware detection and the malware datasets that we profiled in this study. Section III summarizes our implementation named MalWareHouse to profile the datasets. Section IV describes our new profiling techniques and demonstrates them in the surveyed datasets. Section V evaluates the overall profiling results of our profiling in malware datasets. The last section summarizes the study, discusses the benefits of profiling datasets, and highlights the contributions to big data and machine learning studies.

Extra Materials: The tables, extra materials, and some scripts are provided at https://github.com/gurol/dsprofiling.

II. Android Malware Detection Domain and The Surveyed Five Datasets

Android malware detection via machine learning is one of the most studied domain in malware detection in cyber security. Comparing the desktop malware, mobile malware is a complex domain having different characteristics from defensive and offensive perspectives [10]. As the number of new mobile application releases/updates increase dramatically and considering the hundreds of third-party mobile application markets where the security controls may not be forced and applied enough, machine learning based malware analysis and detection become the only effective solution [10].

Although the literature proposes several approaches to detect Android mobile malware, the datasets are not as diverse as them. For our another study, we surveyed 60 studies between 2009 and 2018 that use machine learning binary classification based on static malware analysis and saw that AMGP dataset is in 65% of the studies. 27% includes Contagio Mobile dataset, and 22% uses Drebin datasets (see the extra materials at https://github.com/gurol/dsprofiling). In 47% of the studies use also different datasets, for example, a dataset obtained from an anti-virus company. We included the datasets in Table I to demonstrate our profiling approach in an example mobile malware detection domain.

TABLE I. The five Android Mobile Malware Datasets

Dataset (Abbreviation)	Year	Reference
Android Malware Genome Project (AMGP)	2013	[11]
Drebin	2014	[12]
Android Botnet (ABot)	2015	[13]
Android Malware Dataset (AMD)	2017	[14]
VirusTotal Academic Malware Samples (VT2018)	2018	

Note that the datasets reviewed besides AMGP and Drebin, namely AMD, ABot, and VTotal2018 were not used in the surveyed studies. Therefore, profiling these datasets including AMGP and Drebin is remarkable because the findings by making the critical review can guide the researcher about the quality of the datasets.

III. Profiling Platform Implementation

We have implemented an Android Mobile Malware Analysis Platform named MalWareHouse in order to store the analysis of samples collected from different dataset repositories. The platform that was implemented in Python, MongoDB, and R, analyses the samples per dataset, stores the information in a NoSQL database and includes the queries to extract and visualize profiling information. We used the platform to extract information related to the profiled datasets.

IV. New Profiling Methods and Their Demonstration

This section describes four profiling techniques introduced. In the heading of Each technique is given by the "V" aspects of big data that the technique is addressed to connect the process to big data domain.

Feature space: In machine learning, the attributes in a dataset are called features. Feature space is the other dimension of the dataset apart from sample space. In this study, we chose standard Android application permissions requested by each sample. Android's permission mechanism limits the specific operations performed by the mobile applications and/or provides ad-hoc access to specific pieces of data at the end user's control. The permissions and/or the combination of permissions requested by a mobile application could signal the potential maliciousness [15]. Therefore, permissions are a well-representative feature space for Android mobile malware datasets. We eliminated non-standard or custom permissions overall. Therefore, the maximum feature space size is 151 standard permissions [16].

Each profiling technique is described and then conducted in five Android mobile malware datasets.

A. Basic profiling (Volume, Variety, Veracity, Venue)

Basic profiling gives initial insight about datasets in very high-level. Sample space size (number of samples in a dataset), feature space size (number of attributes or columns in a dataset), size (physical size of the dataset), and the statistics related to other domain-specific prominent criteria comprise basic profiling. The criteria in basic profiling are most likely to be addressed first when we are talking about a dataset. Table II lists the five criteria we recommended for malware datasets.

TABLE II. Basic Profiling Criteria

Criteria	V's	AMGP	Drebin	AMD	ABot	VT2018
Sample space size (m)		1,260	5,555	23,743	1,929	4,725
Feature space size (n)	Volume	65	94	105	78	111
Size (in GB)		1.5	6.8	58.1	2.6	18.4
Malware family	Variety	49	> 20	71	14	N/A
Malware variant	Veracity Venue	N/A	N/A	135	N/A	N/A

Even the basic profiling tells us important aspects of the datasets with respect to different big data aspects. The datasets can be compared according to these profiles. For example, AMD has the largest number of samples and file sizes and it holds extra information about malware namely malware

family/variants. Note that basic profiling can be conducted in a more granular manner. For example, instead of giving total size, average size per sample can be stated without an information lost. Values can be represented by percentages.

B. *Timeline profiling (Veracity, Variability, Velocity)*

Though basic profiling provides initial insight, we cannot be sure about the content of the datasets. We propose the following criteria giving insights related to the timeline of the whole dataset:

- *Age*: the years between the youngest sample and the oldest sample.
- *Freshness*: the years (or months, days) outdated since the youngest sample.
- *API (Application Programming Interface) level range*: the minimum and maximum Android API-level numbers of all the samples.

To assure the homogenous date and date related criteria distribution, the age of the datasets should be maximum (ten years as of 2018), freshness should be minimum (close to zero), and API range reflects the whole latest spectrum. Table III visualize the results of timeline profiling. Both the freshest and the oldest dataset is VT2018. The API range of VT2018 is also the most up-to-date comparing the others. The latest API-level is 28 at the time of writing. Table III proves that the most used datasets in Android malware detection domain, namely Android Malware Genome Project (AMGP) and Drebin are quite out-of-date with respect to date and API coverage. The classifiers trained on these datasets cannot discriminate new threats and evolving malware. Note that the oldest and youngest dates of the samples per dataset calculated by retrieving the time stamp of files in Android application package in MalWareHouse platform.

C. *Duplicate samples profiling (Value, Venue)*

As observed in Android malware datasets, datasets can include the same samples. The duplicated samples in a dataset decrease the value gained from the dataset and waste the time of researchers and consume the processing power. The duplication can be detected by comparing the unique identifier(s) of the samples (i.e. primary key). Hash values are commonly used for identifying malware samples. We used SHA-1 (Secure Hash Algorithm 1) to avoid collocations

TABLE III. TIMELINE PROFILING CRITERIA

Dataset	Oldest	Youngest	API	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Freshness
AMGP	2008	2011	12	2.9 years old											-7 years
Drebin	2008	2012	16	4.1											-6
AMD	2008	2016	25	7.6											-2
ABot	2009	2015	21		6.6										-3
VT2018	2009	2018	27		9.0										0

potentially caused by MD5 (Message-Digest algorithm 5) in MalWareHouse platform. Fig. 1 visualizes the distribution of duplicates samples among per combination of five datasets. Fig. 1 (a) shows the number of duplicated in a Venn diagram. It is straightforward to interpret which datasets have more duplicates and with which datasets. Interestingly, the diagram shows that AMGP is almost included in Drebin dataset (only 11+14 = 25 samples are different). ABot has common samples across AMGP, Drebin, and AMD. VT2018 has no duplicates with any datasets. Summing any non-zero values inside the closed curves yield the unduplicated sample size 35.531.

Fig. 1 (b) is a different representation of common elements in dataset profiling where the number of datasets can be more than six. The horizontal bars at the left show the distribution of samples per dataset. Whereas vertical bars show the samples per individual datasets (single points at the bottom banded area) and per combination of datasets (more than two points). The vertical bars can be sorted according to datasets or the frequency of duplicate samples. Note that we used UpSetR package for the visualization in Fig. 1 (b) [17].

D. *Density/sparsity profiling (Volume, Variety, Value)*

The feature space of the datasets especially the binary or categorized ones is usually sparse. The general density or sparsity levels of the datasets is another criterion for dataset profiling as formulized in (1) and (2).

$$density = \frac{number\ of\ observed\ features}{m \times n} \tag{1}$$

$$sparsity = \frac{number\ of\ unobserved\ features}{m \times n} = 1 - density \tag{2}$$

$$density_n = \frac{n}{n_{max}} \tag{3}$$

$$sparsity_n = 1 - density_n \tag{4}$$

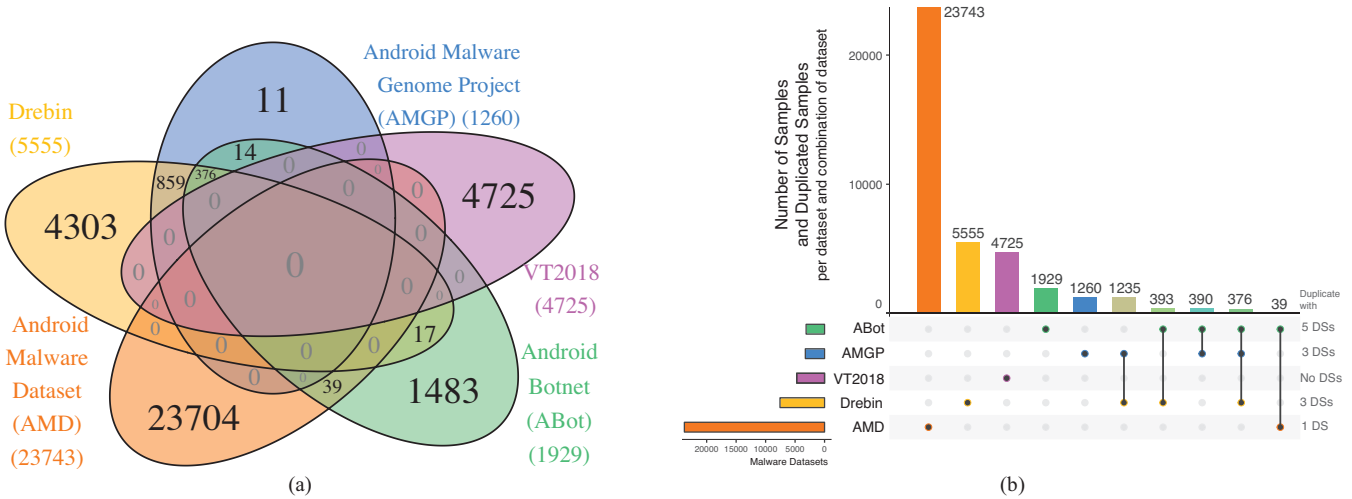

Fig. 1. Visualization of profiling the sample-space duplications (a) in a Venn diagram (b) in an UpSet diagram (sorted by nonduplicate, duplicate frequency)

We also defined density and sparsity for feature space that measures the ratio of the observed features in a dataset to the universal or maximum feature space. As described above, our maximum feature space is 151 Android application permissions.

Table IV shows the density and sparsity profiling values for the datasets. Sample and feature space size are also included to interpret the values. Table IV shows that AMGP is the densest dataset, but it has the least sample and feature space sizes and Drebin is the most sparse dataset. Feature space sparsity is a critical criterion because sparse feature space could not represent the problem domain and a classifier, for example, trained in such a dataset could not distinguish the instance with missing features.

TABLE IV. DENSITY/SPARSITY PROFILING CRITERIA

Criteria	AMGP	Drebin	AMD	ABot	VT2018
Density	18%	10%	11%	16%	15%
Sparsity	82%	90%	89%	84%	85%
Feature space density	43%	62%	70%	52%	74%
Feature space sparsity	57%	38%	30%	48%	26%
Sample space size (*m*)	1,260	5,555	23,743	1,929	4,725
Feature space size (*n*)	65	94	105	78	111

The next section evaluates the overall results of all the profiling techniques conducted.

V. RESULTS

Table V summarizes the overall findings per criteria per profiling together. It also lists the seven big data "V" aspects covered per technique.

TABLE V. OVERALL DATASET PROFILING RESULTS

Profiling	V's[1]	Criteria	AMGP	Drebin	AMD	ABot	VT2018
Basic	Volume	Sample space size (*m*)	1260	5555	**23743**	1929	4725
		Feature space size (*n*)[2]	65	94	105	78	111
		Size (in GB)	1.5	6.8	58.1	2.6	18.4
	Variety Veracity Venue	Malware family[3]	**49**	**> 20**	**71**	**14**	N/A
		Malware variant[3]	N/A	N/A	**135**	N/A	N/A
Timeline	Veracity Variability Velocity	Age (years)	2.9	4.1	7.6	6.6	**9**
		Freshness (years)	-7	-6	-2	-3	**0**
		API-level range[3]	12	16	25	21	**27**
Duplicate samples	Value Venue	Unduplicated samples	11	4303	**23704**	1483	4725
		Duplicated samples	1249	1252	39	446	0
Density / sparsity	Volume Variety Value	Density[4]	18%	10%	11%	16%	15%
		Sparsity[4]	82%	90%	89%	84%	85%
		Feature space density	43%	62%	70%	52%	**74%**
		Feature space sparsity	57%	38%	30%	48%	26%
General heuristic dataset profile			Low (-5)	Normal (1)	High (4)	Normal (1)	High (4)

(1) The 7 V's covered: value, variability, variety, velocity, venue, veracity, volume
(2) Evaluated in feature space density/sparsity
(3) Domain-specific but could be adapted into other domains
(4) Informative. not imply any superiority

In order to assess the datasets globally, we marked the high and low profile criteria with bold values and underlined values, respectively. For example, 23,743 samples of VT2018 dataset corresponds to the highest number whereas 43% feature space density is the lowest ratio among the datasets. The criteria in gray text are redundant. Malware family, malware variant, and API-level range are domain specific. Nevertheless, they can be adapted in other domains.

According to our heuristic evaluation of the criteria, Android Malware Dataset (AMD) and VirusTotal Academic Malware Samples (VT2018) are the high-profile datasets whereas Android Malware Genome Project (AMGP) that belongs to a leading comprehensive study 2013–2014 in the literature is the low-profile dataset. Total 14 (12 non-redundant) criteria together define the profiles comparatively. Refer to online extra materials at https://github.com/gurol/dsprofiling for detailed information.

VI. DISCUSSION AND CONCLUSION

In this study, we have proposed a concise step-by-step profiling approach to gain insights about a group of datasets in different dimensions. The profiling consists of four categories: (1) basic, (2) timeline, (3) duplicate samples, (4) density/sparsity. We employed the profiling method in five Android mobile malware datasets highly used in the literature and separated the low, normal, and high-profile ones. Some visualization methods are also employed and suggested to convey the profiling criteria in more comprehensible manner (i.e. Venn and UpSet diagrams for sample-space duplications and timeline profiling). The comparative results clearly give a quantitative/qualitative degree about datasets.

We highlight the following aspects related to dataset profiling in ideal:

- Labeling datasets as low/high profile is very convenient and helps to eliminate uncertainty over datasets.
- Stating "we achieved 0.97 classification accuracy in Android Malware Genome Project dataset" is quite different when we included the dataset's profile.
- Profiling datasets have a positive effect on increasing their quality especially in a quantitative manner. It motivates the researchers to achieve and use high-profile datasets so that the overall research level increases more competitively.
- Any new dataset shared in a domain will necessitate re-profiling all the datasets.
- Knowing the profile of a dataset leads to address and improve the shortcomings.
- Altogether, we suggest that profiling related datasets in a domain should be an established behavior among the research community.

As big data has become the reality faced in artificial intelligence, our profiling techniques help researchers to maintain the life cycle of datasets confidently and enrich the dataset by importing new datasets from different repositories in more conscious manner.

Especially in machine learning perspective, our profiling techniques can help to compare the real performance of different classifiers trained on different datasets. For example, a high-performing classifier trained on only Android Malware

Genome Project (AMGP) is less credible than a less-performing classifier trained on only Android Malware Dataset (AMD). Because the profiling of the latter's dataset exhibits low profile overall as shown in Table V.

Although the specific criteria embodied into our profiling techniques are not novel, the conceptualization of each technique and overall heuristic assessment method are new. Moreover, the introduced profiling techniques are described in terms of seven "V" aspects of big data, namely volume, variety, velocity, veracity, value, variability, and venue so that they are mutually connected with big data.

The techniques are demonstrated with clear outputs enhanced by some visualization methods where possible. We positioned our profiling approach before feature engineering in a machine learning workflow. It greatly simplifies and relives the burden in proceeding steps. Though we demonstrated our profiling approach in five Android malware datasets comprising 85 GB of malware samples in total and defined some extra specific criteria for malware domain (malware family and variant), each profiling technique can be adapted into other domains. One of the key facts of this study with respect to Android malware detection using machine learning is that the researchers should improve the datasets and should not rely on solely Android Malware Genome Project dataset.

Although we could not give more information about how to conduct specific profiling techniques at least in a procedural specification because of lack of space, there is a considerable amount of original software development that enabled us to experiment the profiling in our reviewed datasets. One limitation we observed in our approach is that profiling big sized datasets consumes more time. Enhancing processing power resources and/or optimizing the algorithm can diminish such problems. The future work, we plan to extend the profiling categories and criteria. We appreciate any feedbacks, recommendations and collaboration offers.

Finally, it is expected that our contributions in this study help researchers to adapt their artificial intelligence and machine learning researches that require quantitative and qualitative datasets to the specific conditions, constraints, and requirements of big data. This is a critical factor in many emerging domains such as malware detection in cyber security.

ACKNOWLEDGMENT

The authors wish to thank the authors of the surveyed datasets.

REFERENCES

[1] "2017 Data Science Report," 2017.

[2] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 42–47.

[3] M. F. Goodchild and K. C. Clarke, "Data Quality in Massive Data Sets," in *Handbook of Massive Data Sets*, Boston, MA: Springer, 2002, pp. 643–659.

[4] H. Chen, D. Hailey, N. Wang, and P. Yu, "A review of data quality assessment methods for public health information systems," *International Journal of Environmental Research and Public Health*, vol. 11, no. 5, pp. 5170–5207, 2014.

[5] A. P. Chapman, A. Rosenthal, and L. Seligman, "The Challenge of 'Quick and Dirty' Information Quality," *Journal of Data and Information Quality*, vol. 7, no. 1–2, pp. 1–4, 2016.

[6] J. Stefanowski, K. Krawiec, and R. Wrembel, "Exploring complex and big data," *International Journal of Applied Mathematics and Computer Science*, vol. 27, no. 4, pp. 669–679, 2017.

[7] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2014.

[8] M. Ben Ellefi *et al*., "Dataset Profiling -a Guide to Features, Methods, Applications and Vocabularies," *Undefined*, vol. 1, pp. 1–5, 2016.

[9] V. Sessions and M. Valtorta, "The effects of data quality on machine learning algorithms," in *Proceedings of the 11th International Conference on Information Quality*, 2006, pp. 485–498.

[10] G. Canbek, S. Sagiroglu, and N. Baykal, "New Comprehensive Taxonomies on Mobile Security and Malware Analysis," *International Journal of Information Security Science (IJISS)*, vol. 5, no. 4, pp. 106–138, 2016.

[11] X. Jiang and Y. Zhou, *Android Malware*. Raleigh, NC, USA: Springer, 2013.

[12] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "DREBIN: Effective and explainable detection of Android malware in your pocket," in *Network and Distributed System Security (NDSS) Symposium*, 2014.

[13] A. Fitriah, A. Kadir, N. Stakhanova, and A. A. Ghorbani, "Android botnets: What URLs are telling us," in *International Conference on Network and System Security (NSS)*, 2015, pp. 78–91.

[14] F. Wei, Y. Li, S. Roy, X. Ou, and W. Zhou, "Deep Ground Truth Analysis of Current Android Malware," in *Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2017*, vol. 10327, Springer, 2017, pp. 252–276.

[15] G. Canbek, N. Baykal, and S. Sagiroglu, "Clustering and visualization of mobile application permissions for end users and malware analysts," in *The 5th International Symposium on Digital Forensic and Security (ISDFS)*, 2017, pp. 1–10.

[16] Android, "Manifest.permission," *Android Developers*, 2018. [Online]. Available: https://developer.android.com/reference/android/Manifest.permission.html.

[17] J. R. Conway, A. Lex, and N. Gehlenborg, "UpSetR: An R package for the visualization of intersecting sets and their properties," *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, 2017.