

Contextual Weisfeiler-Lehman Graph Kernel For Malware Detection

Annamalai Narayanan, Guozhu Meng, Liu Yang, Jinliang Liu and Lihui Chen
Nanyang Technological University, Singapore.

annamala002@e.ntu.edu.sg, {gzmeng, yangliu}@ntu.edu.sg, liuj0081@e.ntu.edu.sg, elhchen@ntu.edu.sg

Abstract—In this paper, we propose a novel graph kernel specifically to address a challenging problem in the field of cyber-security, namely, malware detection. Previous research has revealed the following: (1) Graph representations of programs are ideally suited for malware detection as they are robust against several attacks, (2) Besides capturing topological neighbourhoods (i.e., structural information) from these graphs it is important to capture the context under which the neighbourhoods are reachable to accurately detect malicious neighbourhoods.

We observe that state-of-the-art graph kernels, such as Weisfeiler-Lehman kernel (WLK) capture the structural information well but fail to capture contextual information. To address this, we develop the Contextual Weisfeiler-Lehman kernel (CWLK) which is capable of capturing both these types of information. We show that for the malware detection problem, CWLK is more expressive and hence more accurate than WLK while maintaining comparable efficiency. Through our large-scale experiments with more than 50,000 real-world Android apps, we demonstrate that CWLK outperforms two state-of-the-art graph kernels (including WLK) and three malware detection techniques by more than 5.27% and 4.87% F-measure, respectively, while maintaining high efficiency. This high accuracy and efficiency make CWLK suitable for large-scale real-world malware detection.

keywords — graph kernels, malware detection, program analysis

I. INTRODUCTION

Malware detection has evolved as one of the challenging problems in the field of cyber-security as the attackers continuously enhance the sophistication of malware to evade novel detection techniques. Malware for various platforms such as desktop and mobile devices is growing at an alarming rate. For instance, Kaspersky reports [1] detecting 4 million malware infections in 2015 which is a 216% increase over 2014. This volume and growth rate clearly highlights an imperative need for automated malware detection solutions.

To perform automated malware detection, security analysts resort to program analysis and machine learning (ML) techniques. Typically, this process involves extracting semantic features from suitable representations of programs (e.g., assembly code, call graphs) and detecting malicious code or behavior patterns using ML classifiers [2]–[7].

A major reason for such tremendous growth rate in malware is the production of *malware variants*. Typically, the attackers produce large number of *variants* of the same malware by resorting to techniques such as variable renaming and junk code insertion [3, 4, 7, 8]. These variants perform same malicious functionality, with apparently different syntax, thus evading syntax-based detectors. However, higher level semantic representations such as call graphs, control- and data-flow graphs,

control-, data- and program-dependency graphs mostly stay similar even when the code is considerably altered [3, 5, 7, 8]. In this work, we use a common term, *Program Representation Graph* (PRG) to refer to any of these aforementioned graphs. As PRGs are resilient against variants, many works in the past have used them to perform malware detection. In essence, such works cast malware detection as a *graph classification problem* and apply existing graph mining and classification techniques [8]. Some methods such as [2, 3, 5] note that ML classifiers are readily applicable on data represented as vectors and attempt to encode PRGs as feature vectors. Typically, these techniques face two challenges:

- **(C1) Expressiveness.** PRGs are complex and expressive data structures that characterize topological relationships among program entities. Representing them as vectors is a non-trivial task. In many cases vectorial representations of PRGs fail to capture all the vital information. For instance, AppContext [2], a well-known Android malware detection approach represents apps as PRGs and ends up capturing features from individual nodes without their topological neighbourhood information. With such loss of expressiveness, attacks that span across multiple PRG nodes could not be effectively detected.
- **(C2) Efficiency.** The scale of malware detection problem is such that we have millions of samples already and thousands streaming in every day. Many classic graph mining based approaches (e.g., [8]) are NP hard and have severe scalability issues, making them impractical for real-world malware detection [3, 9].

Graph Kernels. One of the increasingly popular approaches in ML for graph-structured data is the use of graph kernels. Recently, efficient and expressive graph kernels such as [11]–[14] have been proposed and widely adopted in many application areas (e.g. bio- and chemo-informatics [15]–[18]). Some of them support explicit feature vector representations of graphs (e.g. [11]–[13]). Thus both the aforementioned challenges C1 and C2 are effectively addressed by these graph kernels. Therefore, it just suffices to use a graph kernel together with a kernelized ML classifier (e.g., SVM) and we have a scalable, effective and ready-to-use malware detector. Recently, three approaches [3], [6] and [9], have successfully demonstrated using these general purpose graph kernels for malware detection.

Research Gap. However, a major problem in using these general purpose graph kernels on PRGs is that, they are not designed to take domain-specific observations into account.

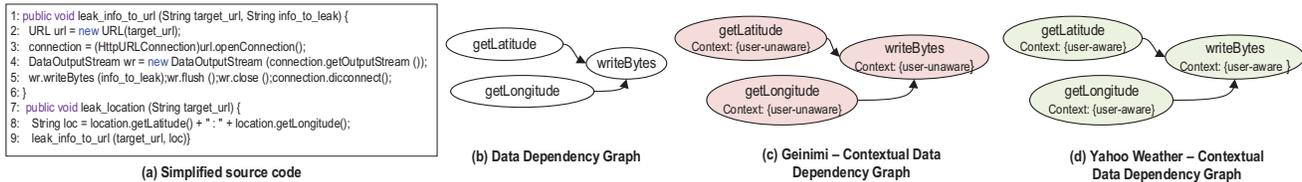


Fig. 1: Location information being leaked in *Geinimi* (malware) and *Yahoo Weather* (benign) apps. (a) code snippet corresponding to leaking location information in both the apps. (b) DDG corresponding to the location leak. (c) *Geinimi*'s DDG illustrating that it leaks information without the user's knowledge. (d) *Yahoo Weather*'s DDG illustrating that it leaks information with the user's knowledge.

For instance, recent research on malware analysis has revealed that besides capturing neighbourhood (i.e., structural) information from PRGs it is important to capture the context under which the neighbourhoods are reachable to accurately detect malicious neighbourhoods [2, 7] (explained in detail in §II). Many existing graph kernels such as [11] and [12] can capture and compare structural information from PRGs effectively. However, they are not designed to capture the reachability context, as it is a strong domain-specific requirement and hence fail to do so. To address this, we develop a novel graph kernel which is capable of capturing both the aforementioned types of information.

For similar domain-specific reasons, researchers from other fields such as computer vision [18], bio- and cheminformatics [15]–[17] have developed a number of kernels that specifically suit their applications. Despite graphs being natural representations of programs and amenable for various activities, the program analysis research community has not devoted significant attention to development of domain-specific graph kernels. We take the first step towards this, by developing a kernel on PRGs which specifically suits our task of malware detection.

Our approach. To improve the accuracy of malware detection process, we propose a method to enrich the feature space of a graph kernel that inherently captures structural information with contextual information. We apply this feature-enrichment idea on a state-of-the-art graph kernel, namely, Weisfeiler-Lehman kernel (WLK) [11] to obtain the Contextual Weisfeiler-Lehman kernel (CWLK). Specifically, CWLK associates to each sub-structure feature of WLK a piece of information about the context under which the sub-structure is reachable in the course of execution of the program. A sub-structure appearing in two different PRGs will match only if it is reachable under the same context in both PRGs. We show that for the malware detection problem, CWLK is more expressive and hence more accurate than WLK and other state-of-the-art kernels while maintaining comparable efficiency.

Experiments. Through our large-scale experiments with more than 50,000 Android apps, we demonstrate that CWLK outperforms two state-of-the-art graph kernels (including WLK) and three malware detection techniques by more than 5.27% and 4.87% F-measure, respectively, while maintaining high efficiency. This, in essence shows the significance of incorporating the contextual information along with structural information in the graph kernel while performing malware detection.

Contributions. The paper makes the following contributions: (1) We develop a graph kernel that captures both structural and contextual information from PRGs to perform accurate

and scalable malware detection (§IV). To the best of our knowledge, this is the first graph kernel specifically addressing a problem from the field of program analysis.

(2) Through large-scale experiments and comparative analysis, we show that the proposed kernel outperforms two state-of-the-art graph kernels and three malware detection solutions in terms of accuracy, while maintaining high efficiency (§V).

(3) We make an efficient implementation of the proposed kernel (along with the dataset information) publicly available¹.

II. BACKGROUND & MOTIVATION

In this section, we motivate the design of our kernel by describing why considering just the structural information from PRGs is insufficient to determine the maliciousness of a sample and how supplementing it with contextual information helps to increase detection accuracy. To this end we use a real-world Android malware² from the *Geinimi* family which steals users' private information. We contrast its behavior with that of a well-known benign app, *Yahoo Weather*.

***Geinimi*'s execution.** The app is launched through a background event such as receiving a SMS or call. Once launched, it reads the user's personal information such as geographic location and contacts and leaks the same to a remote server. The (simplified) malicious code portion pertaining to the location information leak is shown in Fig. 1 (a). The method *leak_location* reads the geographic location through *getLatitude* and *getLongitude* Application Programming Interfaces (APIs). Subsequently, it calls *leak_info_to_url* method to leak the location details (through *DataOutputStream.writeBytes*) to a specific server. The Data Dependency Graph (DDG) corresponding to the code snippet is shown in Fig. 1 (b). The nodes in DDG are labeled with the sensitive APIs that they invoke.

***Yahoo Weather*'s execution.** On the other hand, *Yahoo Weather* could be launched only by user's interaction with the device (e.g., by clicking the app's icon on the dash board). The app then reads the user's location and sends the same to its weather server to retrieve location-specific weather predictions. Hence, DDG portions of *Yahoo Weather* is same as that of *Geinimi*.

Contextual information. From the explanations above, it is clear that both the apps leak the same information in the same fashion. However, what makes *Geinimi* malicious is the fact that its leak happens without the user's consent. In other words, unlike *Yahoo Weather*, *Geinimi* leaks private information through an event which is not triggered by user's

¹<https://sites.google.com/site/cwlkernel>

²SHA256:05620032f3a2abd5e5ea482b5e5d5b8ff5faa8115019736013d87f442032b6be

interaction. We refer to this as a leak happening in *user-unaware* context. On the same lines, we refer to *Yahoo Weather's* leak as happening in *user-aware* context.

As explained in [7] and [2], in the case of Android apps, one could determine whether a PRG node is reachable under *user-aware* or *user-unaware* context by examining its entry point nodes. Following this procedure we add the context as an attribute to every DDG node. This context annotated DDG of *Geinimi* and *Yahoo Weather* are shown in Fig. 1 (c) and (d), respectively.

Requirements for effective detection. From the aforementioned example the two key requirements that makes a malware detection process effective can be identified:

(R1) Capturing structural information. Since malicious behaviors often span across multiple nodes in PRGs, just considering individual nodes (and their attributes) in isolation is not enough. Capturing the structural (i.e., neighborhood) information from PRGs is of paramount importance.

(R2) Capturing contextual information. Considering just the structural information without the context is not enough to determine whether a sensitive behavior is triggered with or without user's knowledge. For instance, if structural information alone is considered, the features of both *Geinimi* and *Yahoo Weather* apps become identical, thus making the latter a false positive. Hence, it is important for the detection process to capture the contextual information as well to make the detection process more accurate.

Many existing graph kernels could address the first requirement well. However, the second requirement which is more domain-specific makes the problem particularly challenging. To the best of our knowledge, none of the existing graph kernels support capturing this reachability context information along with structural information. Hence, this gives us a clear motivation to develop a new kernel that specifically addresses our two-fold requirement.

III. DEFINITIONS AND NOTATIONS

The formal definitions and notations that will be used throughout the paper are presented in this section.

Definition 1 (Program Representation Graph). $PRG = (N, E, \lambda, \xi)$ is a directed graph where N is a set of nodes and each node $n \in N$ denotes program entity such as a function or instruction. $E \subseteq (N \times N)$ is a set of edges and each edge $e(n_1, n_2) \in E$ denotes either control- or data-flow or dependency from n_1 to n_2 . λ is the set of labels that characterize the (security-sensitive) operations of a node and $\ell : N \rightarrow \lambda$, is a labeling function which assigns a label to each node. ξ is a set of events that denote the context of a node and $C : N \rightarrow \xi$, is a function which assigns the context to each node.

Definition 2 (Context). The context of a node $n \in N$ in the PRG of a program P is a set of attributes $\xi(n)$ that govern the reachability of n in the course of execution of P .

Examples of contexts. In the case of Windows executables, the *guard conditions* that govern the execution of a node could be considered as its context [8]. Unlike Windows (and other desktop OS) binaries, Android and iOS mobile apps typically have multiple entry points [2, 7]. Hence, in the case of such

mobile apps, besides guard conditions, the categories of entry points through which a node is reachable could also be considered as its context. Similar platform-specific constraints and observations could be considered while defining the contexts for executables of other platforms.

IV. CONTEXTUAL WEISFEILER-LEHMAN GRAPH KERNEL

In this section, we begin by explaining how the regular WLK can be applied to perform malware detection using PRGs and how it falls short. Subsequently, we introduce our CWLK and discuss how it addresses the shortcomings of WLK. Finally, we prove CWLK's semi-definitiveness and analyze its time complexity.

A. Regular Weisfeiler-Lehman Graph Kernel

WLK computes the similarities between graphs based on the 1-dimensional WL test of graph isomorphism [11].

WL test of isomorphism. Suppose we are to determine whether a given a pair of graphs G and G' are isomorphic. The WL test of isomorphism works by augmenting the node labels by the sorted set of labels of neighboring nodes. This process is referred to as *label-enrichment* and new labels are referred as *neighborhood labels*. Thus, in each iteration i of the WL algorithm, for each node $n \in N$, we get a new neighborhood label, $\lambda_i(n)$ that encompass the i^{th} degree neighborhood around n . $\lambda_i(n)$ could be optionally compressed using a hash function $f : \Sigma^* \rightarrow \Sigma$ such that $f(\lambda_i(n)) = f(\lambda_i(n'))$, iff $\lambda_i(n) = \lambda_i(n')$. To test graph isomorphism, the re-labeling process is repeated until the neighborhood label sets of G and G' differ, or the number of iterations reaches a specific threshold. Therefore, one iteration of WL relabeling is equivalent to a function $r((N, E, \lambda_i)) = (N, E, \lambda_{i+1})$ that transforms all graphs in the same manner.

Definition 3 (WL sequence). Define the WL graph at height i of the graph $G = (N, E, \lambda)$ as the graph $\mathcal{G}_i = (V, E, \lambda_i)$. The sequence of graphs

$$\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_h = (V, E, \lambda_0), (V, E, \lambda_1), \dots, (V, E, \lambda_h) \quad (1)$$

is called the WL sequence up to height h of G , where $\mathcal{G}_0 = G$ (i.e., $\lambda_0 = \lambda$) is the original graph and $\mathcal{G}_1 = r(\mathcal{G}_0)$ is the graph resulting from the first relabeling, and so on.

Definition 4 (WL kernel). Given a valid kernel $k(\cdot, \cdot)$ and the WL sequence of graph of a pair of graphs G and G' , the WL graph kernel with h iterations is defined as

$$k_{WL}^{(h)}(G, G') = k(\mathcal{G}_0, \mathcal{G}'_0) + \dots + k(\mathcal{G}_h, \mathcal{G}'_h) \quad (2)$$

where h is the number of WL iterations and $\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_h$ and $\mathcal{G}'_0, \mathcal{G}'_1, \dots, \mathcal{G}'_h$ are the WL sequences of G and G' , respectively. h is referred as *height of the kernel*.

Intuitively, WLK counts the common neighborhood labels in two graphs. Hence we have $k_{WL}^{(h)}(G, G') = |\{\lambda_i(n), \lambda_i(n')\}|$, iff $f(\lambda_i(n)) = f(\lambda_i(n'))$ for $i \in \{1, \dots, h\}$, $n \in N, n' \in N'$, where f is injective and the sets $\{f(\lambda_i(n)) | n \in N \cup N'\}$ and $\{f(\lambda_j(n)) | n \in N \cup N'\}$ are disjoint for all $i \neq j$.

Example & WLK's shortcoming. We now apply WLK on the real-world examples discussed in §II to see if it distinguishes malicious and benign neighborhoods clearly, facilitating accurate detection. For the ease of illustration,

Algorithm 1 CWLK - Contextual re-labeling

Input:

$G = G_0 = (N, E, \lambda_0, \xi)$ — PRG with set of nodes (N), set of edges (E) and set of node labels (λ_0) and context for each node (ξ)
 h — number of iterations

Output:

$\{\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_h\}$ - contextual WL sequence of height h

```
1: procedure CONTEXTUAL RE-LABEL( $G, h$ )
2:   for  $i = 0$  to  $h$  do
3:     for all  $n \in N$  do
4:        $\sigma_i(n) \leftarrow \emptyset$ 
5:       if  $i = 0$  then
6:         for  $c \in \xi(n)$  do
7:            $\sigma_i(n) \leftarrow \sigma_i(n) \cup c \oplus \lambda_0(n)$ 
8:         end for
9:       else
10:         $\mathcal{N}(n) \leftarrow \{m \mid (n, m) \in E\}$ 
11:         $M_i(n) \leftarrow \{\lambda_{i-1}(m) \mid m \in \mathcal{N}(n)\}$ 
12:         $\lambda_i(n) \leftarrow \lambda_{i-1}(n) \oplus \text{sort}(M_i(n))$ 
13:        for  $c \in \xi(n)$  do
14:           $\sigma_i(n) \leftarrow \sigma_i(n) \cup c \oplus \lambda_i(n)$ 
15:        end for
16:        end if
17:         $\sigma_i(n) \leftarrow \text{join}(\sigma_i(n))$ 
18:         $\gamma_i(n) \leftarrow f(\sigma_i(n))$ 
19:      end for
20:       $\mathcal{G}_i \leftarrow (N, E, \gamma_i)$ 
21:    end for
22:    return  $\{\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_h\}$ 
23: end procedure
```

the label compression step is avoided. Applying WLK on the DDG for both *Geinimi* and *Yahoo Weather* apps, shown in Fig. 1 (b), for the node `getLatitude`, for heights $h = 0, 1$, we get the neighborhood labels `getLatitude` and `getLatitude, writeBytes`, respectively. Clearly, WLK captures the neighborhood around the node `getLatitude`, incrementally in every iteration of h . In fact, neighborhood label for $h = 1$ captures that another sensitive node, `writeBytes` lies in the neighborhood of `getLatitude`, which highlights a possible privacy leak. However, WLK does not capture whether the neighborhood involved in this leak is reached in *user-aware* or *unaware* context. This is precisely what we address through our CWLK.

B. Contextual Weisfeiler-Lehman graph Kernel

The goal of CWLK is to capture not only neighborhoods around the node, but also to include the contexts in which each of the neighborhoods is reachable in the PRG. To this end, we modify the re-labeling step of WLK so as to accommodate the context of every neighborhood. We refer to this process as *contextual-relabeling* and the sequence of graphs thus obtained as *contextual WL sequence*.

Contextual re-labeling. Specifically, CWLK performs one additional step in the re-labeling process which is to attach the contexts of every node to its neighborhood label in every iteration. This in effect, indicates the contexts under which a particular neighborhood is reachable. The label thus obtained is referred to as *contextual neighborhood label*. The contextual relabeling process is presented in detail in Algorithm 1.

The inputs to the algorithm are PRG, G and the degree of neighbourhoods to be considered for re-labeling, h . The output

is the sequence of contextual WL graphs, $\{\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_h\} = \{(N, E, \gamma_0), (N, E, \gamma_1), \dots, (N, E, \gamma_h)\}$, where $\gamma_1, \dots, \gamma_h$ are constructed using the contextual relabeling procedure.

For the initial iteration $i = 0$, no neighborhood information needs to be considered. Hence the contextual neighborhood label $\gamma_0(n)$ for all nodes $n \in N$ is obtained by justing prefixing the contexts to the original node labels and compressing the same (lines 6-8,17-18). For $i > 0$, the following procedure is used for contextual re-labeling. Firstly, for a node $n \in N$, all of its neighboring nodes are obtained and stored in $\mathcal{N}(n)$ (line 10). For each node $m \in \mathcal{N}(n)$ the neighborhood label up to degree $i - 1$ is obtained and stored in multiset $M_i(n)$ (line 11). $\lambda_{i-1}(n)$, neighborhood label of n till degree $i - 1$ is concatenated to the sorted value of $M_i(n)$ to obtain the current neighborhood label, $\lambda_i(n)$ (line 12). Finally the current neighborhood label is prefixed with the contexts of node n to obtain the string $\sigma_i(n)$ which is then compressed using the function f to obtain the contextual neighborhood label, $\gamma_i(n)$ (lines 13-15,17-18).

Definition 5 (CWL kernel). Given a valid kernel $k(\cdot, \cdot)$ and the CWL sequence of graph of a pair of graphs G and G' , the contextual WL graph kernel with h iterations is defined as

$$k_{WL}^{(h)}(G, G') = k(\mathcal{G}_0, \mathcal{G}'_0) + \dots + k(\mathcal{G}_h, \mathcal{G}'_h) \quad (3)$$

where h is the number of CWL iterations and $\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_h$ and $\mathcal{G}'_0, \mathcal{G}'_1, \dots, \mathcal{G}'_h$ are the CWL sequences of G and G' , respectively.

Intuitively, CWLK counts the common contextual neighborhood labels in two graphs. Hence we have $k_{CWL}^{(h)}(G, G') = |(\sigma_i(n), \sigma_i(n'))|$, iff $f(\sigma_i(n)) = f(\sigma_i(n'))$ for $i \in \{1, \dots, h\}, n \in N, n' \in N'$.

Example. We now apply CWLK on the apps in our example to show how it overcomes WLK's shortcomings. The contextual neighborhood labels σ (without compression) of the node `getLatitude` in *Geinimi* app for heights $h = 0, 1$ are, **user-unaware** \oplus `getLatitude` and **user-unaware** \oplus `getLatitude, writeBytes`, respectively. For the same node in *Yahoo Weather* the contextual neighborhood labels are **user-aware** \oplus `getLatitude` and **user-aware** \oplus `getLatitude, writeBytes`. Hence, it is evident that the CWLK's contextual relabeling provides a means to clearly distinguish malicious PRG neighborhoods from the benign ones. This is achieved by complementing the structural information with contextual information. Therefore, unlike WLK, CWLK based classification does not detect *Yahoo Weather* as a false positive. This example clearly establishes the suitability of CWLK for the malware detection task.

We now prove CWLK's positive definiteness and also analyze its time complexity.

Theorem 1. CWLK is positive definite.

Proof. Let us define a mapping ϕ that counts the occurrences of a particular contextual neighborhood label sequence σ in G (generated in h iterations of Algorithm 1). Let $\phi_\sigma^{(h)}(G)$ denote the number of occurrences of σ in G , and analogously

$\phi_\sigma^{(h)}(G')$ for G' . Then,

$$k_\sigma^{(h)}(G, G') = \phi_\sigma^{(h)}(G), \phi_\sigma^{(h)}(G') = |\{(\sigma_i(n), \sigma_i(n'))\}| \quad (4)$$

$$\sigma_i(n) = \sigma_i(n'), i \in \{1, \dots, h\}, n \in N, n' \in N'\}$$

Summing over all σ from the vocabulary Σ^* , we get

$$k_{CWL}^{(h)}(G, G') = \sum_{\sigma \in \Sigma^*} k_\sigma^{(h)}(G, G') = \sum_{\sigma \in \Sigma^*} \phi_\sigma^{(h)}(G) \phi_\sigma^{(h)}(G')$$

$$= |\{(\sigma_i(n), \sigma_i(n')) | \sigma_i(n) = \sigma_i(n'), i \in \{1, \dots, h\}, n \in N, n' \in N'\}|$$

$$= |\{(\sigma_i(n), \sigma_i(n')) | f(\sigma_i(n)) = f(\sigma_i(n')),$$

$$i \in \{1, \dots, h\}, n \in N, n' \in N'\}| \quad (5)$$

where the last equality follows from the fact that f is injective. As $f(\sigma) \neq f(\sigma')$ if $\sigma \neq \sigma'$, the string σ corresponds to exactly one contextual neighborhood label and $k_{CWL}^{(h)}$ defines a kernel with corresponding feature map $\phi_{CWL}^{(h)}$, such that

$$\phi_{CWL}^{(h)} = (\phi_\sigma^{(h)}(G))_{\sigma \in \Sigma^*} \quad (6)$$

Complexity. The runtime complexity of CWLK with h iterations on a graph with n nodes and e edges is $O(he)$ (assuming that $e > n$) which is same as that of WLK. More specifically, the neighborhood label computation with sorting operations (lines 10-12 of Algorithm 1) take $O(e)$ time for one iteration and the same for h iterations take $O(he)$. The inclusion of context (lines 6-8,13-15), does not incur additional overhead as $e \gg |\xi|$. Hence the final time complexity remains as $O(he)$. For a detailed derivation and analysis of the time complexity of WLK, we refer the reader to [11].

Efficient computation of CWLK on K graphs. When computing CWLK on K graphs to obtain $K \times K$ kernel matrix, a naïve approach would involve K^2 comparisons, resulting a time complexity of $O(K^2he)$. However, as mentioned in [11], a Bag-of-Features (BoF) model based optimization could be performed to arrive the kernel matrix in $O(Khe + K^2hn)$ time. This optimized computation involves the following steps: (1) A vocabulary Σ of all the contextual neighbourhood labels of nodes across the K graphs is obtained in $O(Khe)$ time. This facilitates representing each of the K graphs as feature vectors of $|\Sigma|$ dimensions. (2) Subsequently, $K \times K$ kernel matrix can be computed by multiplying these vectors in $O(K^2hn)$ time. In summary, CWLK has the same efficiency as that of WLK and supports explicit feature vector representations of PRGs.

Relation to other spatial contextual kernels. Two recently proposed graph kernels [13] and [14], consider incorporating the spatial context information to neighborhood subgraph features. They define *context of a subgraph feature as another subgraph appearing in its vicinity*. As mentioned earlier, in our malware detection problem we refer to *attributes of a node which determine its reachability as its context*. This *reachability context* is different from *spatial context* discussed in [13] and [14]. Hence CWLK is consummately different from these two kernels.

TABLE I: Composition of dataset

| Portion of dataset | Dataset source | # of samples |
|--------------------|------------------------------|--------------|
| Malware portion | DREBIN [4], Virus Share [19] | 29877 |
| Benign portion | Google Play [20] | 25000 |

TABLE II: Dataset Statistics

| CG (Avg \pm Std) | | ICFG (Avg \pm Std) | |
|--------------------|-----------------|----------------------|-------------------|
| # of Nodes | # of Edges | # of Nodes | # of Edges |
| 1556 \pm 998 | 3327 \pm 1803 | 15323 \pm 9844 | 22745 \pm 20922 |

V. EVALUATION

We conducted large scale experiments involving more than 50,000 Android apps from two real-world malware datasets to evaluate the accuracy and efficiency of CWLK. We compare CWLK's performance against that of two state-of-the-art kernels and three Android malware detection solutions.

A. Datasets

DREBIN [4] provides a collection on 5,560 Android malware apps collected from 2010 to 2012. More recently, Virusshare [19] released a collection of 24,317 malware apps collected from 2010 to 2014. We combined these two datasets and use them in our evaluation. For the benign portion of the dataset, we collected 25,000 benign top-selling apps from Google Play store [20] that were released around the same time. Thus, our dataset contains a total of 54,877 apps. The composition of our dataset is presented in Table I.

Graph Representations considered. As mentioned in §I the proposed CWLK could be applied on any type of PRGs to perform malware detection. In our evaluations, we experimented with two types of PRGs namely, call graphs (CGs) and inter-procedural control-flow graphs (ICFGs). The nodes of a CG represent the methods present in an app; its directed edges represent the *calling/called relations* between the methods. The nodes of an ICFG represent the individual instructions present in those methods; its edges represent the *control-flows* among those instructions. These two types of PRGs are chosen since they capture program semantics at different levels of granularities. The statistics on the average number of nodes and edges of the PRGs in our dataset are presented in Table II. Details on construction of these PRGs are presented later in §V-B.

Training and Test sets. 60% of the samples were randomly chosen from the datasets and used for training the classifier and the remaining 40% samples are used to test their performances. The classifiers' hyper-parameters are determined on the training set using 5-fold cross-validation, whereas the test set is only used for determining the final detection performance. We repeat this procedure 5 times and average the results.

B. Implementation & Comparative Analysis

PRG construction & Context Identification. Both CG and ICFG of the apps in our datasets are constructed through static analysis using *Soot* [21], a well-known Android static analysis workbench. The nodes are labeled with the security-sensitive APIs they access and are annotated with context information. Nodes that do not access any sensitive APIs are removed. We use the category of entry points of each of the nodes in these graphs as their contexts. The procedure proposed in DroidSIFT [7] is used to identify and categorize each entry point as being in 'user-aware' or 'user-unaware' context.

Comparison with Graph Kernels. CWLK's accuracy and efficiency is compared against those of WLK [11] and Neighborhood Hash Graph Kernel (NHGK) [12]. Since these kernels cannot capture context information, we use the comparative

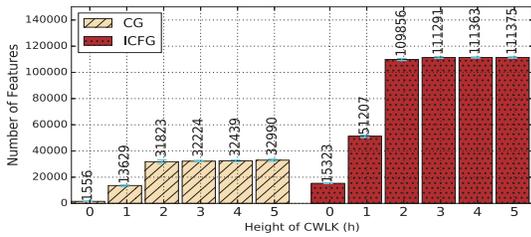


Fig. 2: No. of features obtained for various heights of CWLK.

analysis against them to ascertain whether including context information significantly improves the accuracy without affecting the efficiency. We implemented all these kernels in about 2,170 lines of Python code. For all the kernels, BoF model based implementation similar to the one discussed in §IV-B is used to obtain explicit feature vector representation of samples.

Comparison with Malware Detection Solutions. Also, our approach is compared against three light-weight state-of-the-art ML based Android malware detection solutions, namely, DREBIN [4], Allix *et al.* [5] and ADAGIO [3]. To this end, we re-implemented DREBIN and Allix *et al.*'s approaches through consultations with the authors. For ADAGIO, an open-source implementation provided by the authors is used.

Since the malware detection accuracy of these solutions predominantly depend on the features they use, we briefly introduce them here. DREBIN [4] considers features such as sensitive Android APIs, permissions and components used by apps. Allix *et al.* [5] constructs the CFGs of individual methods and represent them as signature strings which are subsequently used as features. ADAGIO [3] constructs CGs and uses byte-code instructions to assign labels to nodes. NHGK [12] is used to extract CG neighborhoods as features and a histogram-intersection (HI) kernel SVM is trained to detect malware. Due to severe scalability issues (explained in §V-D3) ADAGIO is ran only once on our dataset. All other techniques are ran 5 times and the average results are reported.

C. Experimental Design

Research Questions. Through our evaluations, we seek to address the following research questions: (1) Does including the context information in PRGs through CWLK significantly improve the malware detection accuracy? (2) Does capturing context information incur significant computation overhead to adversely affect the efficiency? (3) Does our context-based detection approach outperform existing malware detection solutions? Two separate experiments are designed to address these questions.

Experiment E1. In order to evaluate the first two questions, the following experiment is conducted: First, the CGs of all the training set apps are constructed. Then, the WLK, NHGK and CWLK kernels are applied on the CGs to obtain respective kernel matrices. Subsequently, a SVM classifier is trained with each of these kernels to detect malicious apps. Finally, the test set apps are subjected to the aforementioned kernel computation processes and are used to evaluate the models. The same procedure is repeated with ICFGs.

Experiment E2. In order to address question (3), we subject DREBIN [4], Allix *et al.* [5] and ADAGIO [3] solutions to

the same training and test sets. We compare them against the best performing model that uses CWLK (obtained from experiment E1).

Setting the parameter h : We experimented with different kernel heights $h = 0$ to 5 for CWLK (see eq. (3)). The average number of contextual neighborhood features (from 5 runs of E1) for different values of h on the two PRGs is reported in Fig. 2. It is evident that the number of features does not increase significantly after $h = 2$ on both PRGs. This is because we have removed nodes that do not access sensitive APIs which affects the connectivity and restricts the neighborhood sizes. In other words, we seldom have neighborhoods around nodes that span beyond degree 2. Similar trend is observed in WLK and NHGK features. Hence we restrict the height h to be 0, 1 and 2 for all three kernels. Thus, for each kernel applied on each PRG, we have three SVM classifiers (one for each h). Therefore we have a total of 18 malware detection models under comparison (see Table III for details).

Evaluation metrics. Standard evaluation metrics such as Precision, Recall and F-measure are used to determine the effectiveness of malware detection [5]. Efficiency is determined in terms of training and testing durations.

Evaluation Setup. All the experiments were conducted on Intel Xeon Hexa-core E5-2640 processor (2.50 GHz) with 32 GB RAM running Ubuntu 14.04.

D. Results and Discussions

1) Impact of context information on F-measure: We compare the Precision, Recall and F-measure of the malware detection process through CWLK with those of WLK and NHGK. This is to ascertain whether incorporating the context information in PRGs boosts the effectiveness of detection. These results for the 18 models are presented in Table III, from which the following inferences are drawn:

- At the outset, two general observations are made: (1) All models perform better on ICFGs compared to CGs. This is because, ICFG is a more fine-grained representation of programs than CG, which enables it to capture program semantics at a finer level and thereby boosting the detecting accuracy. Hence, we conclude that in our experiments ICFG is a more effective representation for malware detection. (2) Considering larger neighborhoods helps capturing the structural information better which in turn reflects in better performances. This is evident as Precision, Recall and F-measure values get better with increasing values of h for all the three kernels. However, this observation may not hold for large values of h , as nodes that are far apart will be considered for neighborhood re-labeling, leading to a noisy re-labeling process.
- It is clear that CWLK outperforms both WLK and NHGK in terms of F-measure consistently on both CG and ICFG representations.
- *Since the only difference between WLK and CWLK is the latter's capability to capture the context information, evidently this is the reason for CWLK's superior performance.*
- Also, CWLK achieves better Precision than WLK in all the experiments, consistently. This indicates that CWLK suffers

TABLE III: Comparison of CWLK against state-of-the-art Graph Kernels (metrics expressed as % values) - avg. over 5 runs

| PRG Type | Kernel | $h = 0$ | | | $h = 1$ | | | $h = 2$ | | |
|----------|-----------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| | | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| CG | WLK [11] | 71.21 (± 1.15) | 80.93 (± 1.78) | 75.75 (± 1.28) | 85.00 (± 1.59) | 85.93 (± 2.10) | 85.46 (± 1.39) | 87.28 (± 1.02) | 86.15 (± 0.78) | 86.71 (± 0.88) |
| | NHGK [12] | 63.05 (± 3.11) | 60.14 (± 2.04) | 61.56 (± 2.56) | 67.18 (± 1.76) | 69.02 (± 2.01) | 68.08 (± 1.88) | 74.17 (± 1.11) | 70.22 (± 1.69) | 72.14 (± 1.88) |
| | CWLK | 77.29 (± 1.75) | 79.85 (± 0.83) | 78.54 (± 0.97) | 88.50 (± 0.95) | 87.98 (± 1.41) | 88.23 (± 1.10) | 89.19 (± 0.83) | 90.10 (± 0.96) | 89.64 (± 0.94) |
| ICFG | WLK [11] | 85.49 (± 2.59) | 81.66 (± 3.71) | 83.53 (± 2.19) | 90.17 (± 1.92) | 87.48 (± 1.20) | 88.8 (± 1.37) | 93.09 (± 0.81) | 88.16 (± 0.77) | 90.55 (± 0.80) |
| | NHGK [12] | 68.68 (± 4.10) | 70.73 (± 2.77) | 69.68 (± 3.20) | 75.78 (± 2.18) | 81.01 (± 2.64) | 78.30 (± 2.06) | 78.11 (± 1.07) | 82.28 (± 1.50) | 80.14 (± 1.29) |
| | CWLK | 89.74 (± 2.33) | 81.25 (± 3.12) | 85.28 (± 2.59) | 96.77 (± 1.63) | 92.08 (± 1.20) | 94.36 (± 1.15) | 97.15 (± 0.28) | 94.53 (± 0.75) | 95.82 (± 0.66) |

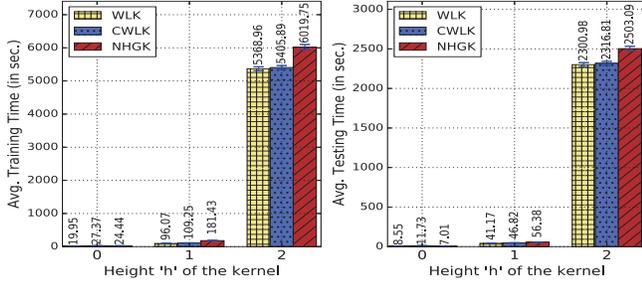


Fig. 3: Training and Testing durations for different values of h on ICFG representation (avg. over 5 runs)

lesser false positives than WLK. This reduction in false positives is a direct result of capturing context information which helps to precisely distinguish malicious neighborhoods in the PRGs from the benign ones (as discussed in §IV).

- NHGK consistently fails to produce better results than WLK and CWLK. This is because, the node labeling and hashing technique adapted in NHGK causes collisions among informative and non-informative subgraph features. Such collisions could be avoided in WLK and CWLK through label compression (as explained in §IV), making them more expressive and accurate. For further details on this limitation of NHGK, we refer the reader to the original work at [12] (particularly Section 5 where the authors discuss this limitation in detail).

Since CWLK uses both contextual and structural information, it is important to analyze the contribution of each of these types of information to its performance. Capturing only structural information is equivalent to using WLK. Hence from rows 1 and 4 of Table III (ignoring columns for $h = 0$), it is evident that structural information alone could provide a minimum of 85.46% and an average of 87.88% F-measure across both PRGs. Similarly, the contribution of contextual information alone is ascertained using CWLK and setting $h = 0$ to be a minimum of 78.54% and an average of 81.91% F-measure. Finally, the effect of using both types of information is ascertained by using CWLK and setting $h > 0$ to be a minimum of 88.23% and an average of 92.02% F-measure across both PRGs. This clearly conveys that structural information is primary for performing effective malware detection and contextual information complements it, thereby helping to improve the accuracy. CWLK attains superior performance by capturing both these types of information.

2) **Impact of context information on efficiency:** We compare the training and testing durations of classifiers using the three kernels under study from experiment E1. This is to infer whether CWLK achieves higher accuracy at the cost of higher computation time. As discussed earlier in §IV, CWLK’s time complexity is same as that of WLK. Hence, we expect

TABLE IV: Comparison of CWLK based malware detection against the state-of-the-art detectors (avg. over 5 runs)

| Method | Precision | Recall | F-measure |
|-------------------------|-----------------------------|-----------------------------|-----------------------------|
| DREBIN [4] | 97.02 (± 0.93) | 85.60 (± 1.03) | 90.95 (± 0.71) |
| Allix <i>et al.</i> [5] | 88.24 (± 0.74) | 86.29 (± 1.66) | 87.25 (± 0.54) |
| ADAGIO [3] | 92.18 | 87.32 | 89.68 |
| CWLK | 97.15 (± 0.28) | 94.53 (± 0.75) | 95.82 (± 0.66) |

negligible difference in efficiency.

Experimental Setting. Due to space constraints, the training and testing times of the *worst-case experimental setting* involving samples with maximum number of nodes and edges are only reported. This setting corresponds to using ICFG representations of apps. The average training and testing durations in this experimental setting for various values of h over 5 runs of experiment E1 are presented in Fig. 3, from which the following inferences are drawn.

- The training and testing durations of CWLK and WLK are almost same for all values of h . This is due to the fact that adding the context information which is the only additional operation in CWLK results in negligible difference on the total kernel computation, training and testing durations.
- The training and testing durations increase with the height of the kernel h for all the kernels in a similar fashion. This is mainly due to the fact that both WLK and CWLK condense the neighborhood labels in the same way which is very similar to that of NHGK. Also, for all these kernels, there is a huge increase in training and testing durations when h increases from 1 to 2 which does not reflect proportionally in terms of F-measure. Hence, one may choose to perform malware detection in a much scalable and reasonably accurate manner by setting $h = 1$ (i.e., considering only degree-1 neighbors).
- It is noted that CWLK along with the other two kernels in Fig. 3, shows high efficiency as it operates in linear time on the density of PRGs. In particular, it is more efficient than the classic walk-, tree- and path-based graph kernels (discussed in [10]) and is suitable for large-scale malware detection.

Summary. From experiment E1, we conclude that CWLK through its virtue of capturing both structural and contextual information, significantly improves upon the accuracy of WLK for the malware detection task without hurting the efficiency.

3) **CWLK Vs. state-of-the-art malware detectors:** We now compare CWLK based detection with the state-of-the-art Android malware detection solutions to study whether contextual PRG neighborhoods makes good features for malware detection, through experiment E2. For CWLK based detection, ICFG representation with $h = 2$ is used as it offers the best performance. The Precision, Recall and F-measures of each of these methods are reported in Table IV. The following observations are made from the table:

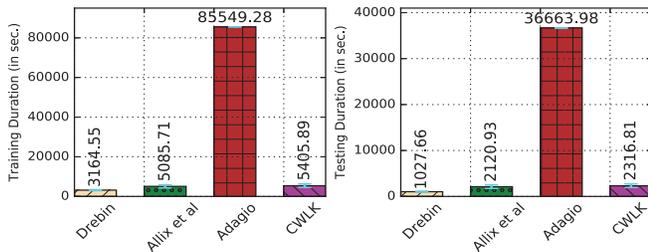


Fig. 4: Comparison on Efficiency of CWLK with state-of-the-art malware detectors (avg. over 5 runs)

- Clearly, CWLK based malware detection outperforms all the compared solutions in terms of F-measure. In particular, our approach outperforms the best performing technique (i.e., DREBIN) by 4.87% F-measure. In terms of Precision, our approach outperforms ADAGIO and Allix *et al.*'s methods and is comparable to DREBIN. In terms of Recall, ours outperforms other methods.
 - Out of the methods compared, DREBIN does not use both structural and contextual features. ADAGIO and Allix *et al.*'s approaches use structural information but not contextual information. This reveals that capturing both these types of information is the reason for our approach's superior performance, reinforcing our findings from experiment E1.
- We now compare the efficiency of CWLK based detection against that of state-of-the-art malware detectors. It is noted that these techniques use different features and classifiers and hence a wide variation in training and testing durations is expected. The results of this comparison is presented in Fig. 4, from which the following observations are made:
- DREBIN being a light-weight non PRG based approach it has significantly higher efficiency than all other methods, including ours.
 - Allix *et al.*'s method is similar to ours in terms of using PRG based features. Hence our efficiency is comparable to this method.
 - ADAGIO uses NHGK and HI kernel SVM in the primal formulation. Hence it takes a prohibitively long time for training and testing. Our method is far more efficient than ADAGIO.

In conclusion, our method's efficiency is comparable to that of other PRG based methods, far better than heavy-weight approaches and inferior to non PRG based light-weight methods. **Summary.** From experiment E2, we conclude that when compared to state-of-the-art malware detectors, CWLK produces considerably higher accuracy with a practically tractable efficiency, making it suitable for large-scale real-world malware detection.

VI. CONCLUSION & FUTURE WORK

In this paper, we present CWLK, a novel graph kernel that facilitates detecting malware using PRGs. Unlike the existing kernels which capture only the security-sensitive neighborhoods in PRGs, CWLK captures these neighborhoods along with the context under which they are reachable. This makes CWLK more expressive and in turn more accurate than existing kernels. Besides expressiveness, CWLK has two specific advantages: (1) shows high efficiency, (2) supports building

explicit feature vector representations of PRGs. CWLK is evaluated on a large-scale experiment with more than 50,000 Android apps, and is found to outperform two state-of-the-art graph kernels and three malware detection techniques in terms of F-measure, while maintaining comparable efficiency.

Future work. In our future work, we plan to investigate incorporating contextual information in other sub-structure based graph kernels such as [12] and [14] and subsequently, study their suitability for performing malware detection.

Implementation & Dataset. We provide an efficient implementation of CWLK and information on the datasets used within this work at: <https://sites.google.com/site/cwlkernel>

VII. ACKNOWLEDGMENT

We thank the authors of [4] and [5], for their suggestions that helped us re-implement their methods.

REFERENCES

- Kaspersky 2014 Annual Threat Report. URL: https://securelist.com/files/2015/12/Kaspersky-Security-Bulletin-2015_FINAL_EN.pdf
- Yang, Wei, et al. "Appcontext: Differentiating malicious and benign mobile app behaviors using context." Proc. of the International Conference on Software Engineering (ICSE), 2015.
- Gascon, Hugo, et al. "Structural detection of android malware using embedded call graphs." Proceedings of the 2013 ACM workshop on Artificial intelligence and security. ACM, 2013.
- Arp, Daniel, et al. "Drebin: Effective and explainable detection of android malware in your pocket." Proceedings of the Annual Symposium on Network and Distributed System Security (NDSS), 2014.
- Allix, Kevin, et al. "Empirical assessment of machine learning-based malware detectors for Android." Empirical Software Engineering (2014): 1-29.
- Sahs, Justin, and Latifur Khan. "A machine learning approach to android malware detection." Intelligence and Security Informatics Conference (EISIC), 2012 European. IEEE, 2012.
- Zhang, Mu, et al. "Semantics-aware Android malware classification using weighted contextual API dependency graphs." Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014.
- Fredrikson, Matt, et al. "Synthesizing near-optimal malware specifications from suspicious behaviors." Security and Privacy (SP), 2010 IEEE Symposium on. IEEE, 2010.
- Anderson, Blake, et al. "Graph-based malware detection using dynamic analysis." Journal in Computer Virology 7.4 (2011): 247-258.
- Vishwanathan, S. Vichy N., et al. "Graph kernels." The Journal of Machine Learning Research 11 (2010): 1201-1242.
- Shervashidze, Nino, et al. "Weisfeiler-lehman graph kernels." The Journal of Machine Learning Research 12 (2011): 2539-2561.
- Hido, Shohei, and Hisashi Kashima. "A linear-time graph kernel." Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. IEEE, 2009.
- Navarin, N., Sperduti, A., & Tesselli, R. (2015, November). Extending local features with contextual information in graph kernels. In Neural Information Processing (pp. 271-279). Springer International Publishing.
- Costa, Fabrizio, and Kurt De Grave. "Fast neighborhood subgraph pairwise distance kernel." Proceedings of the 26th International Conference on Machine Learning. Omnipress, 2010.
- Borgwardt, Karsten M., et al. "Protein function prediction via graph kernels." Bioinformatics 21.suppl 1 (2005): i47-i56.
- Fröhlich, Holger, Jörg K. Wegner, and Andreas Zell. "Assignment kernels for chemical compounds." Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. Vol. 2. IEEE, 2005.
- Mahé, Pierre, and Jean-Philippe Vert. "Graph kernels based on tree patterns for molecules." Machine learning 75.1 (2009): 3-35.
- Harchaoui, Zäid, and Francis Bach. "Image classification with segmentation graph kernels." Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.
- Virus Share malware dataset. URL: <http://virusshare.com>
- Google Play Store. URL: <https://play.google.com/store>
- Soot framework. URL: <http://sable.github.io/soot>