DLGraph
MALWARE DETECTION
DL & GRAPH EMBEDDINGS

By Millis Sahar
Nov 2020

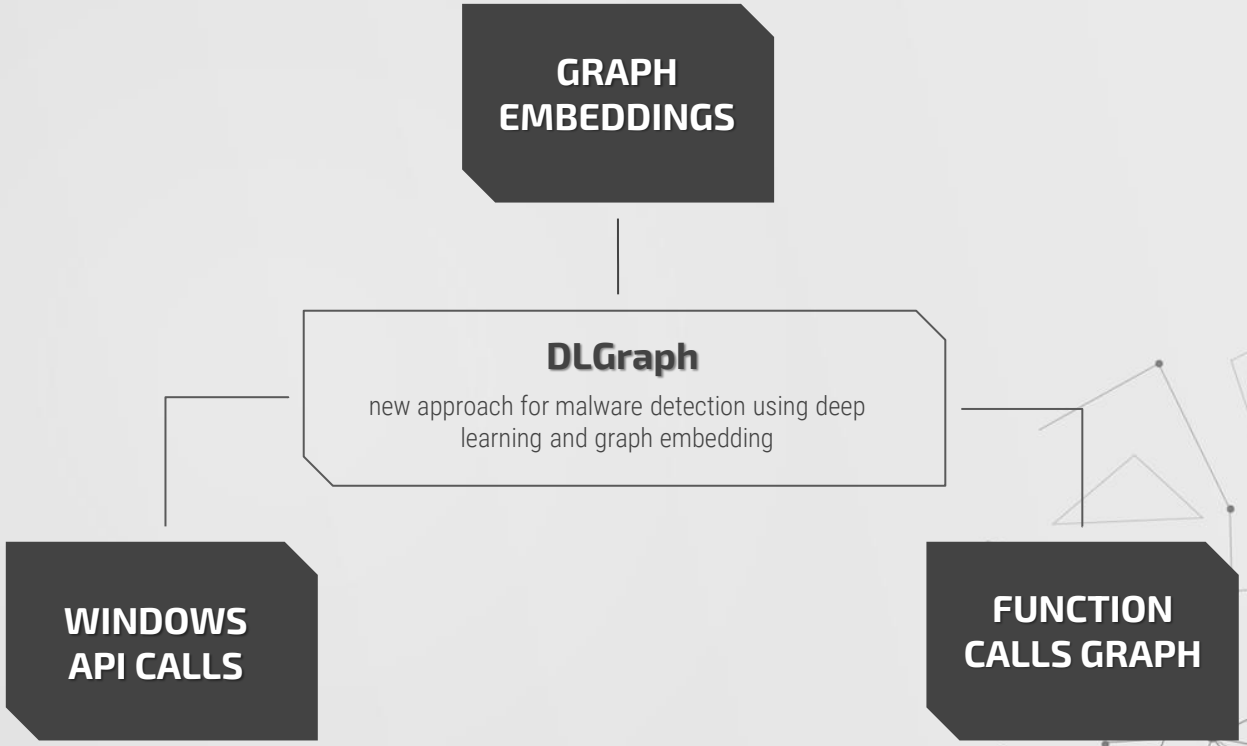
LinkedIn Medium GitHub

TABLE OF CONTENTS

01
ABSTRACT

DLGraph: Malware Detection Using Deep Learning and Graph Embedding | 2018
New Jersey Institute of Technology, USA
Professor Haodi Jiang

DL-GRAPH FOR MALWARE DETECTION



02

INTRODUCTION

The Problem, Challenge, Approch

THE PROBLEM

PROGRAM CLASSIFICATION

Given a program, they applied their Graphs Embedding's approach, then use DL to classify whether the given program is malware or not.

Based on different datasets demonstrate the effectiveness of the proposed approach and its superiority.

CHLLANGES

Why DL?

VOLUME

Malware population is rapidly growing.
It’s a race between antivirus software developers and malware producers

Why static?

BEHAVIOR ANALYSIS

Static vs Dynamic
Not by executing in a safe environment.
Waiting for the right time (disguise).

Why embeddings?

FEATURES

Malware detection features include string signatures, byte sequence n-grams, control flow graphs, and so on.

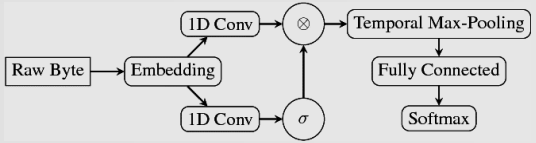
The signature-based approach usually cannot recognize new malware

03

RELATED WORK

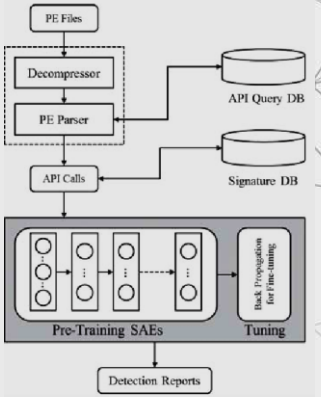
Other techniques, Unique idea

OTHER TECHNIQUES



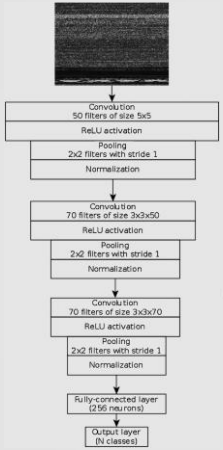
Edward Raff, 2017

Malware Detection by Eating a Whole EXE
Article



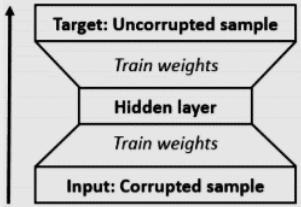
William Hardy, 2016

A Deep Learning Framework for Intelligent Malware Detection
DL4MD Article



Daniel Gibert, 2017

Convolutional neural networks for classification of malware assembly code
Article



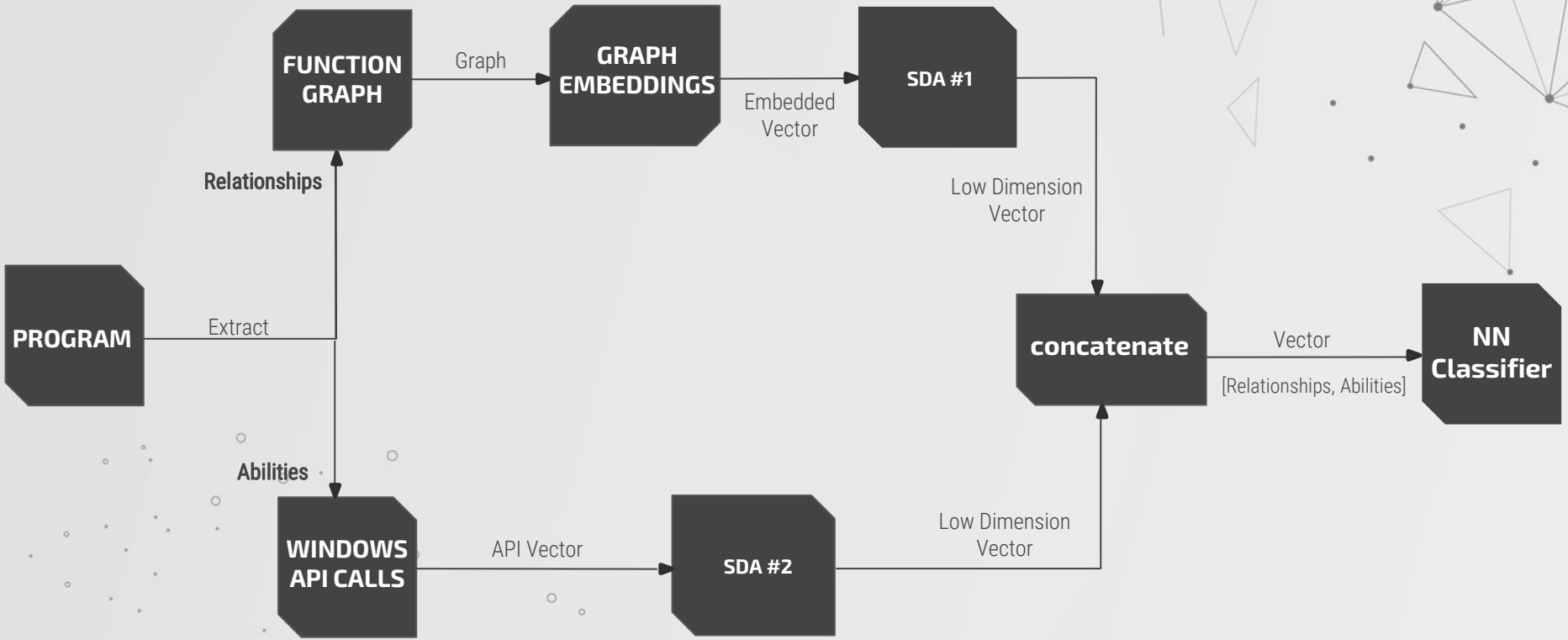
David & Netanyahu, 2015

Deep Learning for Automatic Malware Signature Generation and Classification
DeepSign Article

COMPARE SIMILAR TECHNIQUES

	WIN API CALLS	FUNC CALLS	STACKED AUTOENCODERS	CODE ANALYSIS	GRAPH EMBEDDINGS
DLGraph 2018	V 22k	V	V	Static	V
DL4MD 2016	V 10k	X	V	Static	X
DeepSign 2015	V 20k	V	V	Dynamic (Sandbox)	X

APPROCH


FUNCTION GRAPH
Graph
GRAPH EMBEDDINGS
Embedded Vector
SDA #1
Low Dimension Vector
Relationships
PROGRAM
Extract
concatenate
Vector
[Relationships, Abilities]
NN Classifier
Abilities
WINDOWS API CALLS
API Vector
SDA #2
Low Dimension Vector

04

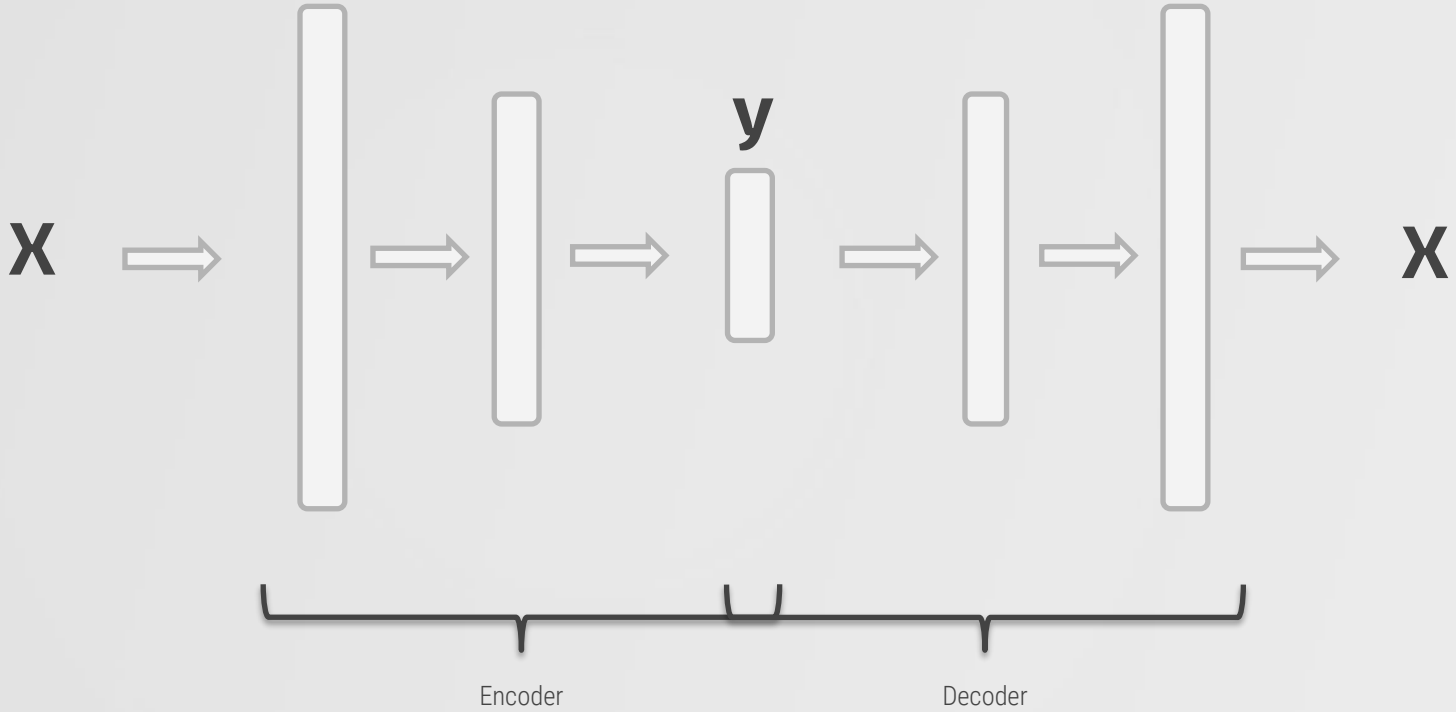
SDA MODEL

Stacked Denoising Autoencoder

STACKED AUTOENCODER

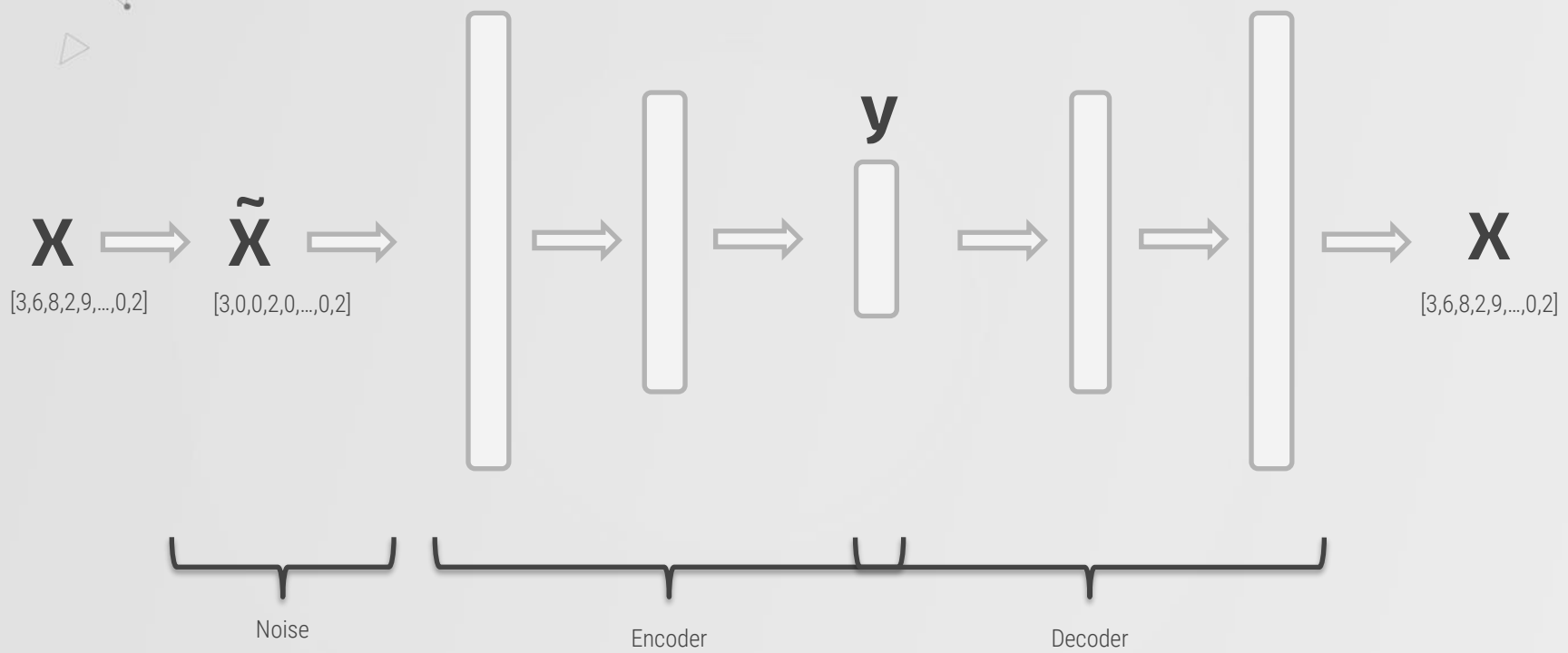
WHY?

Generalization | Dimensionality Reduction | Feature Extraction | Noise reduction



DENOISING STACKED AUTOENCODER

WHY?
Generalize better & Faster



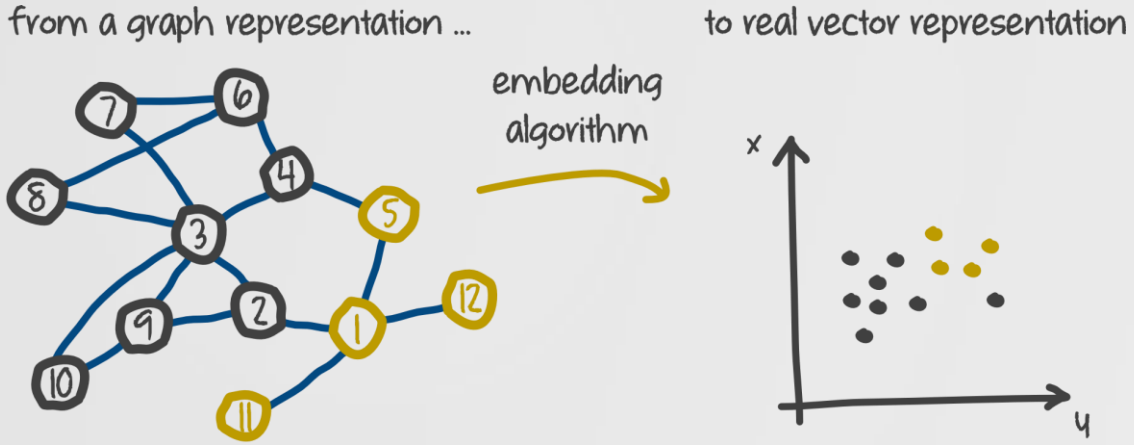
05

GRAPH EMBEDDINGS

Datasets, Performance Measure, Experimental Results

IT'S NOT WHAT YOU KNOW...
IT'S WHO YOU KNOW!

A way of mapping something into a fixed length vector
that captures key features while reducing the dimensionality



EMBEDDINGS MOTIVATION

How to represent graphs in a mathematical way?
How similar two graphs are?
Similar meaning?

Unknown Features

Ability/Will/May download files

May to monitor PC

May send information to Microsoft

May send information to “Hackers”

Chrome	Edge	Power Point	Malware
0.99	0.96	0.13	0.36
0.24	0.31	0.02	0.84
0.02	0.89	0.76	0.08
0.3	0.23	0.08	0.63

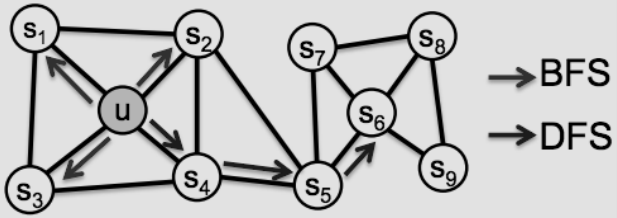
BIAS 2-ORDER RANDOM WALKS

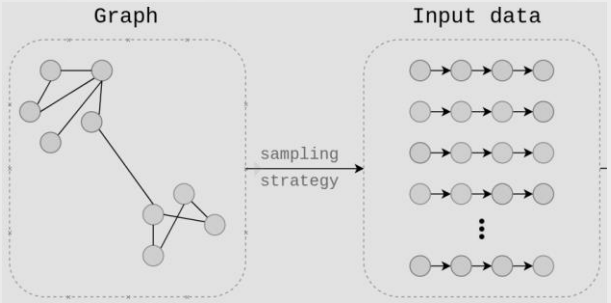
ATTRIBUTES WEIGHTS AND MORE

Node2Vec

Random walks Algorithm to generate vector representations of nodes on a graph, and learns low-dimensional representations for nodes

BFS
DFS



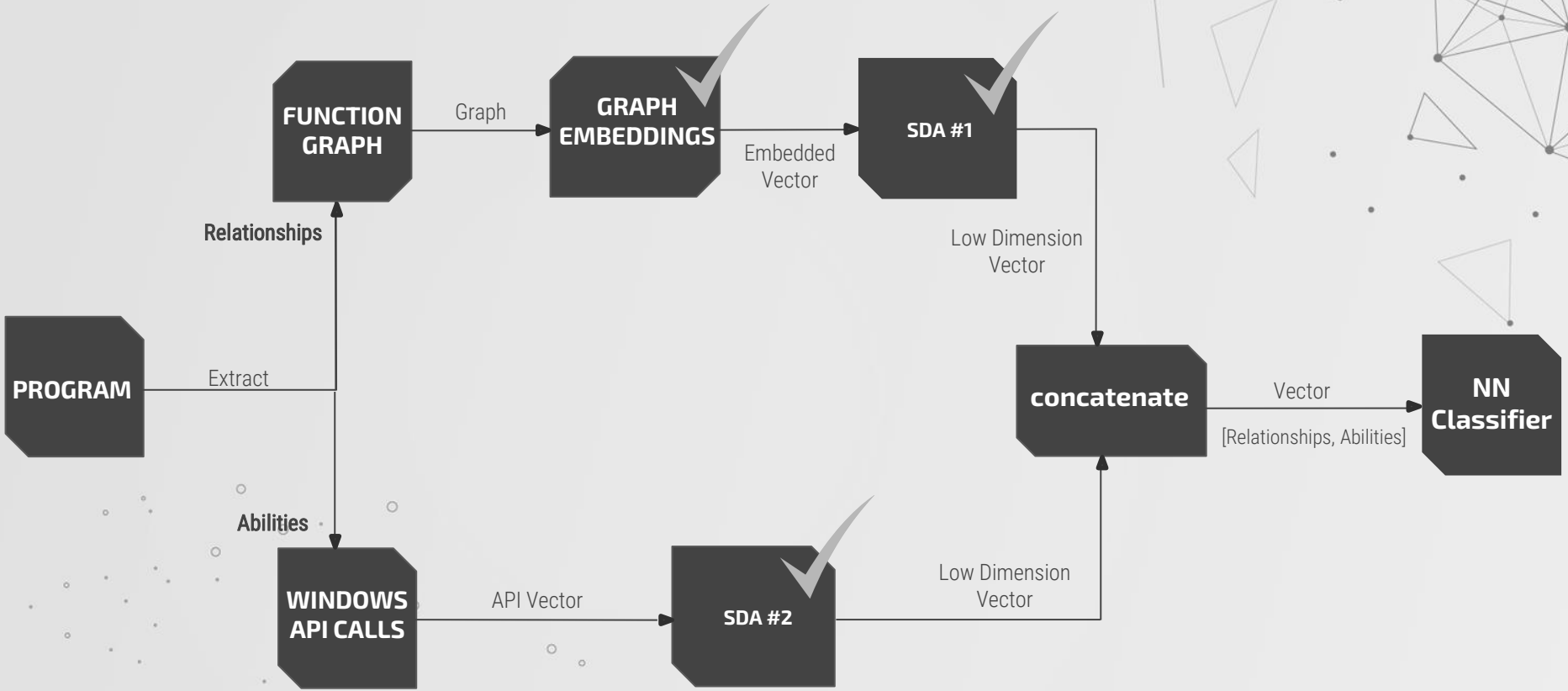


06

PROPOSED APPROACH

Windows API Calls
Function-Call Graph
Dual Stacked Denoising Autoencoders

APPROCH


FUNCTION GRAPH
Graph
GRAPH EMBEDDINGS
Embedded Vector
SDA #1
Low Dimension Vector
Relationships
PROGRAM
Extract
concatenate
Vector
[Relationships, Abilities]
NN Classifier
Abilities
WINDOWS API CALLS
API Vector
SDA #2
Low Dimension Vector

WINDOWS API CALLS

Abilities

CODE

Extract API calls

API CALLS

Count calls per occurrence

VECTOR

22K Boolean dimension vector

Possible WIN API CALLS:	Progrem WIN API CALLS:	Vector for WIN API CALLS:
API1,API2,API3,API4,API5	API1, API5	[1,0,0,0,1]

FUNCTION CALL GRAPH

Relationships

A function-call graph (FCG) shows the calling **relationships** between subroutines or functions in a computer program.

It can be generated from a binary executable through static analysis of the executable code with a disassembly tool.

They used IDA Pro.

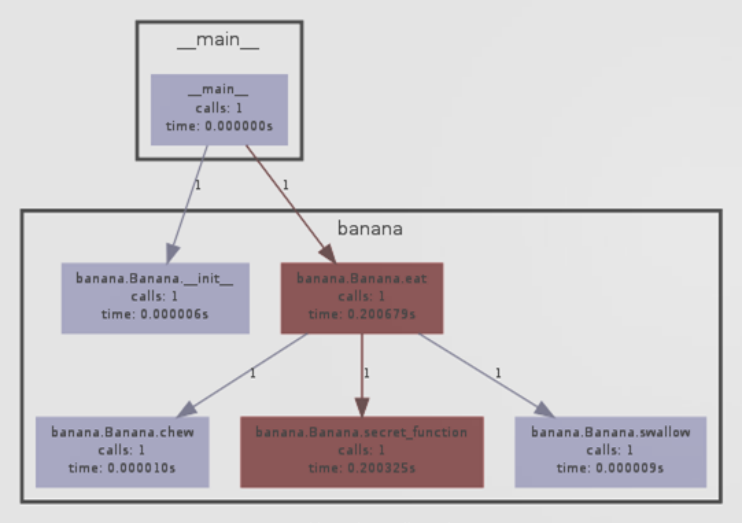
Egypt ncc KcacheGrind Graphviz CodeViz

EXAMPLE

In general, a function-call graph is a directed graph, in which each node represents either a local function implemented by the program designer, or an external system or library function.

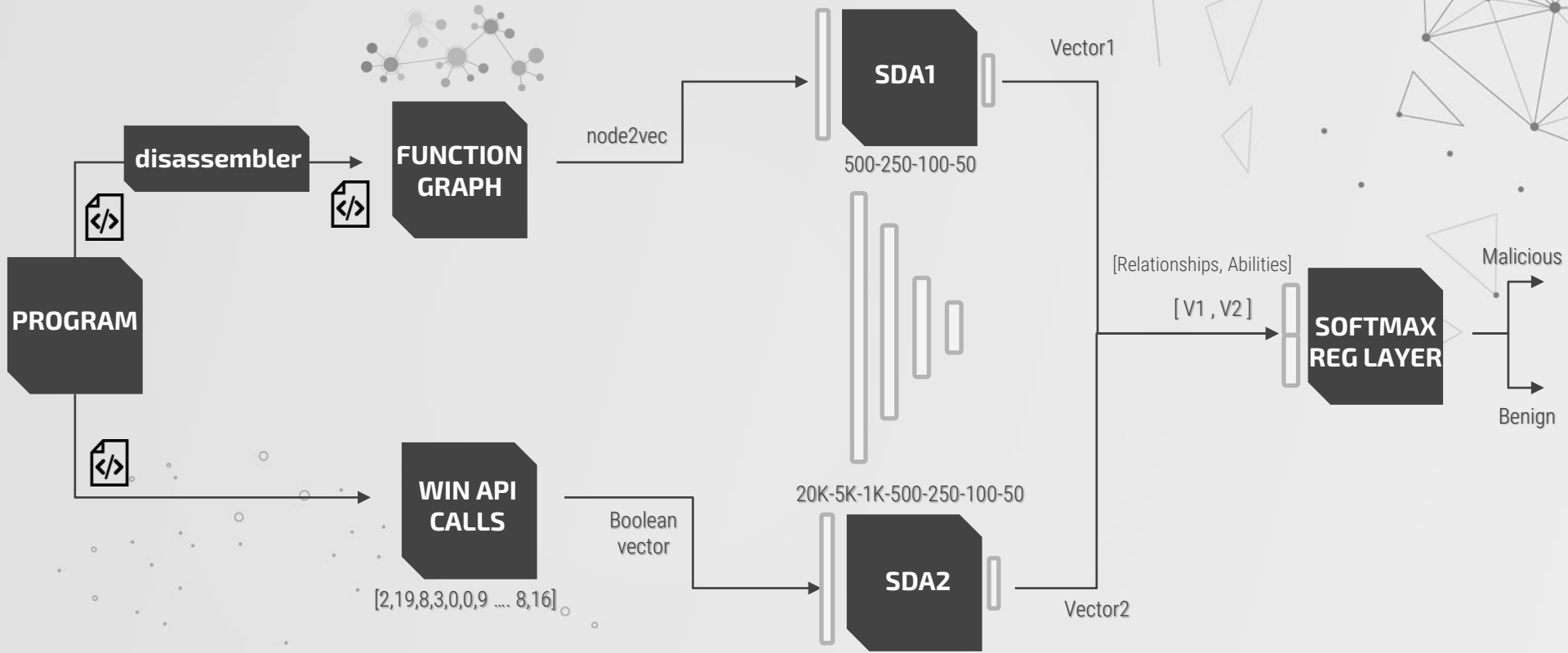
The directed edges in the function-call graph represent the caller-callee relationships between the functions (nodes)

```
class Banana:

    def __init__(self):
        pass

    def eat(self):
        self.secret_function()
        self.chew()
        self.swallow()

    def secret_function(self):
        time.sleep(0.2)

    def chew(self):
        pass

    def swallow(self):
        pass
```




```
from pycallgraph import PyCallGraph
from pycallgraph.output import GraphvizOutput
from banana import Banana

graphviz = GraphvizOutput(
                    input_file='Banana.py',
                    output_file='banana.png')
```

DL ARCHITECTURE



07

EXPERIMENTS & RESULTS

Datasets, Performance Measure, Experimental Results

DATASETS

Collection & Labelling

VirusShare

Malware samples, security researchers

Microsoft

Malware Classification Challenge
2015, 35GB, 20K samples

KafanBBS

Famous Internet security forum in China

Windows 7

*.exe files, *.dll files and *.sys

VirusTotal

60 antivirus scanners

MALWARE

"This adware program shows ads as you browse the web. It can also redirect your search engine results, monitor your PC, download applications, and send information to hackers"

Lollipop

"The Kelihos botnet, also known as Hlux, is a botnet mainly involved in spamming and the theft of bitcoins"

Kelihos-ver3

COLLECTED DATA

	TYPE	COUNT	DS#1	DS#2	DS#3
Lollipop	malicious	2,434	V		V
Kelihos-ver3	malicious	2,584		V	V
***.exe files**	benign	631	V	V	V
***.dll files**	benign	1,178	V	V	V
***.sys files**	benign	368	V	V	V

Total of 2,177 benign samples

In each dataset

80% of the data - training
20% of the data - testing

EXPERIMENTAL RESULTS

	DL4MD	DLGraph	Improve by
Dataset #1	0.9838	**0.9914**	0.0076
Dataset #2	0.9912	**0.9936**	0.0024
Dataset #3	0.9875	**0.9931**	0.0056

* Accuracy Score
** TP is a malicious program

08

CONCLUSIONS

FUTURE WORK & REFERENCES

FUTURE WORK

SOPHISTICATED ARCHITECTURES

FUNCTION CALL GRAPH

Windows API calls

MALWARE IMAGES (CNNs)

PROGRAM BEHAVIOR

multiclass classification

RESOURCES

(not in article)

Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion
By Pascal Vincent, Hugo Larochelle and more
https://dl.acm.org/doi/10.5555/1756006.1953039

Relational inductive biases, deep learning, and graph networks
By DeepMind
https://arxiv.org/pdf/1806.01261

DeepWalk: Online Learning of Social Representations
By Bryan Perozzi, Rami Al-Rfou, Steven Skiena
https://arxiv.org/pdf/1403.6652

Anonymous Walk Embeddings
By Sergey Ivanov, Evgeny Burnaev
https://arxiv.org/pdf/1805.11921

Graph Nets library | https://github.com/deepmind/graph_nets

GraphSAGE: Inductive Representation Learning on Large Graphs | http://snap.stanford.edu/graphsage

DeepWalk: Implementing Graph Embeddings in Neo4j | https://neo4j.com/blog/deepwalk-implementing-graph-embeddings-in-neo4j

THANKS

Does anyone have any questions?

SAHAR.MILIS@GMAIL.COM

https://www.linkedin.com/in/sahar-millis
https://medium.com/@sahar.millis
https://github.com/saharmilis