FACULDADE DE ENGENHARIA DE COMPUTAÇÃO

PROJETO FINAL I e II

PLANO DE TRABALHO

SFAnalytics

Lucas Carvalho Roncoroni

Edmar Roberto Santana de Rezende

15/04/2017

INTRODUÇÃO

Atualmente, 32% dos computadores no mundo estão infectados por algum tipo de malware (TopTenReviews, 2014).

A identificação de um malware, ou programa malicioso, na maioria dos casos, é feita por um antivírus. Os antivírus trabalham com um sistema de detecção por assinatura.

O que é uma assinatura? Uma assinatura é uma sequência de bytes que identifica um malware. Para gerar uma assinatura é necessário que um programa suspeito tenha sido classificado como malware.

A forma mais comum de determinar se um programa sem nenhuma assinatura é ou não um malware, é verificando sua funcionalidade, através de uma análise de dados. Para tanto utilizam-se certos parâmetros, como código fonte, strings encontradas, dlls usadas, uso de memória, dentre outros, extraídos através de ferramentas específicas para essa finalidade. Para o analista, que se proponha a tal tarefa, é essencial justificar a classificação encontrada.

Outra forma habitual de se realizar essa identificação é através da utilização de técnicas de aprendizagem de máquina, onde com uma base inicial, torna-se possível a classificação de novos programas. O problema com essa abordagem é que o analista fica sem saber como foi feita a classificação, não conseguindo justificar o resultado. Por isso essa abordagem é pouco utilizada.

CARACTERIZAÇÃO DE PROBLEMAS E OBJETIVO (S)

Quase um milhão de malwares são criados todos os dias (CNN, 2015), hackers estão custando entre U$345 e U$545 bilhões anualmente para usuários e empresas (U.S. News, 2014). Por isso o desenvolvimento de ferramentas que ajudem um analista a identificar novas ameaças é de extrema importância.

Devido a esse cenário, urge que se desenvolvam ferramentas mais simples que auxiliem o analista em seu trabalho de detecção de novas ameaças. Seria interessante uma ferramenta com a utilização de uma interface gráfica, além de possibilitar o uso da classificação com aprendizado de máquina.

Sendo assim este trabalho propõe a elaboração de um artefato que extraia parâmetros de um programa suspeito, através da análise estática, apresente os parâmetros extraídos utilizando uma interface gráfica, e mostre como foi feita a classificação do programa pelo algoritmo de aprendizagem de máquina.

PLANO DE AVALIAÇÃO DO TRABALHO

Será feita a aprendizagem com malwares para treino; após o treino, alguns especialistas farão a análise de um malware utilizando o artefato proposto. Após classificarem o malware, eles irão responder a um questionário de avaliação da ferramenta.

O trabalho será bem-sucedido, se a avaliação dos especialistas sobre a ferramenta, utilizando o questionário for positiva.

PROPOSTA DO ARTEFATO

O artefato deste trabalho consiste em um classificador de arquivos em maliciosos ou não através de aprendizado supervisionado de máquina com a frequência de cada instrução do código objeto, dlls utilizadas pelo programa e strings dentro do programa.

O diagrama de arquitetura do artefato é apresentado na Figura 1.

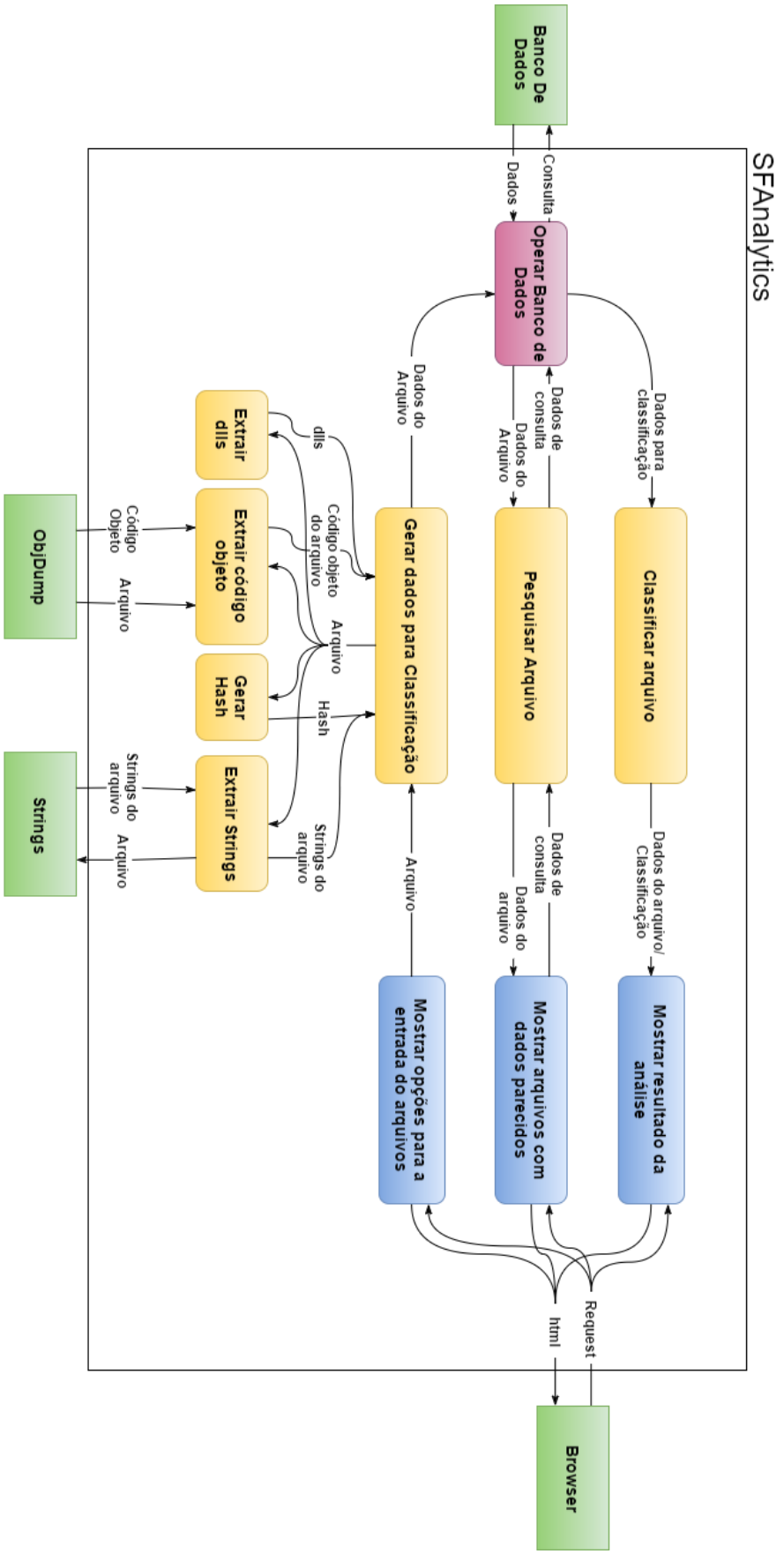

Figura 1. Diagrama de Arquitetura.

TRABALHOS RELACIONADOS

Trabalho	Análise estática	Análise dinâmica	Interface gráfica	Aprendizado de máquina	Descrições em alto nível
VxStream	X	X	X		X
Malwr	X	X	X		
SFAnalytics	X		X	X	X

MÉTODO DE DESENVOLVIMENTO

O método escolhido foi o Scrum, pois é um método bem flexível, os Sprints são definidos durante o projeto, e o que é feito em cada Sprint varia conforme as necessidades do cliente.

CRONOGRAMA

Identificação da Atividade	**Descrição**	**Duração**	
		Início	**Fim**
A1	Gerenciar o TCC	23/2/15	07/12/15
A2	Definição do projeto	23/02/17	24/03/17
A3	Definição das tecnologias utilizadas no projeto	07/03/17	31/03/17
A4	Definição dos dados utilizados para AM	07/03/17	31/03/17
A5	Realização do diagrama de arquitetura	22/03/17	24/05/17
A6	Realização do plano de trabalho	22/03/17	16/06/17
A7	Realização do diagrama de fluxo de dados	22/03/17	10/04/17
A8	Realização do diagrama de classes	25/03/17	19/04/17
A9	Implementação do artefato	03/04/17	13/10/17
A10	Validação com o cliente da GUI	12/04/17	29/05/17
A11	Montagem da Base de Dados	12/06/17	30/06/17
A12	Definição do algoritmo de aprendizagem	03/07/17	16/08/17
A13	Inserção do algoritmo de aprendizagem	03/07/17	16/08/17
A14	Avaliar e Validar o Trabalho	18/08/17	18/10/17
A15	Escrever monografia	10/10/17	27/11/17
A16	Preparar defesa do TCC	10/10/17	27/11/17

DISTRIBUIÇÃO DE ATIVIDADES

Identificação da Atividade	**Primeiro Semestre Mês/Semana**																				
	Fev				**Mar**					**Abr**				**Mai**				**Jun**				
				1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
A1				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
A2				X	X	X	X	X														
A3						X	X	X	X													
A4								X	X	X												
A5								X	X	X	X	X	X	X	X	X						
A6								X	X	X	X	X	X	X	X	X	X	X	X			
A7								X	X	X	X											
A8								X	X	X	X	X										
A9										X	X	X	X	X	X	X	X	X	X	X	X	X
A10											X	X	X	X	X	X	X					
A11																		X	X	X	X	X

A12																	X	X	X	X	X
A13																					
A14																					
A15																					
A16																					

Identificação da Atividade	**Segundo Semestre Mês/Semana**																				
	Jul				**Ago**				**Set**					**Out**				**Nov**			
	20	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
A1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
A2																					
A3																					
A4																					
A5																					
A6																					
A7																					
A8																					
A9	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X						
A10																					
A11																					
A12	X	X	X	X																	
A13	X	X	X	X	X	X	X														
A14							X	X	X	X	X	X	X	X	X	X					
A15														X	X	X	X	X	X	X	X
A16														X	X	X	X	X	X	X	X

RESULTADOS ESPERADOS

Identificação do Resultado	**Descrição**	**Identificação da Atividade**
R1	Plano de trabalho	A1
R2	Relatório de Atividades	A2
R3	Diagrama de fluxo de dados	A3
R4	Diagrama de classes	A4
R5	Artefato computacional	A5
R6	Base de dados	A6

RECURSOS MATERIAIS

Recursos de hardware:

Notebook.

Recursos de software:

Objdump;

Strings;

Visual Studio.

UTILIZAÇÃO DOS RECURSOS MATERIAIS

Dia	Segunda-feira	Terça-feira	Quarta-feira	Quinta-feira	Sexta-feira	Sábado	Domingo
Horário	17h-23h	17h-23h	20h-23h	17h-23h	17h-23h	14h-18h	13h-17h
Recurso	Notebook	Notebook	Notebook	Notebook	Notebook	Notebook	Notebook

GRAU DE DIFICULDADE – ASPECTOS DE INOVAÇÃO E APRIMORAMENTO

Inovação	**Grau de dificuldade**
Mostrar como foi feita a classificação	Alto

Mostrar como foi feita a classificação – O artefato deve mostrar o que o algoritmo de aprendizado de máquina aprendeu para fazer a classificação. O grau é alto devido à falta de conhecimento do autor sobre como fazer isso e também interfere na decisão do algoritmo utilizado no projeto.

Aprimoramento	**Grau de dificuldade**
Django	Médio
Objdump	Baixo
Python	Médio
Strings	Baixo

Django – Framework para a realização do sistema. O grau é médio devido a falta de conhecimento do autor do framework.

Objdump – Ferramenta para extração do código objeto de um executável. O grau é baixo pois a ferramenta não é difícil de ser utilizada, a dificuldade é na criação de uma interface dela com o artefato.

Python – Linguagem de programação utilizada no projeto. O grau é médio pela falta de conhecimento aprofundado do autor nessa linguagem.

Strings – Ferramenta para extração do código objeto de um executável. O grau é baixo pois a ferramenta não é difícil de ser utilizada, a dificuldade é na criação de uma interface dela com o artefato.

ANÁLISE DE RISCOS

Entrada	Probabilidade	Risco	Alternativa
Base de dados	Baixa	Médio	Gerar base a partir de Malwares de forma automatizada
Objdump	Baixa	Leve	Existem outras ferramentas que fazem a mesma coisa.
Django	Alta	Médio	Pedir ajuda ao orientador, por ter mais conhecimento sobre o framework.
Notebook do autor	Média	Médio	Usar computadores da faculdade para implementar o projeto.

OUTRAS OBSERVAÇÕES

Para o controle de versionamento e backup tem sido usados o Google Drive e o github.

REFERÊNCIAS

TopTenReviews, *How Infected Are We?* Disponível em: <http://www.toptenreviews.com/software/security/best-antivirus-software/how-infected-are-we.html>. Acesso em 9 de abril de 2017.

CNN, *Nearly 1 million new malware threats released every day.* Disponível em: <https://www.usnews.com/news/articles/2014/06/09/study-hackers-cost-more-than-445-billion-annually>. Acesso em 8 de abril de 2017.

U.S. News, *Study: Hackers Cost More Than $445 Billion Annually.* Disponível em: <http://money.cnn.com/2015/04/14/technology/security/cyber-attack-hacks-security/>. Acesso em 8 de abril de 2017.

DEFINIÇÕES E ABREVIATURAS

Artefato Computacional – sistema de *software* ou de *hardware*, ou ainda uma combinação dos dois, que será desenvolvido com vistas à solução de um ou mais problemas identificados em um ambiente de interesse.

Dll – Biblioteca de linkagem dinâmica.

Malware – Programa que danifica ou faz ações indesejadas no computador.

Relatório de Atividades – conjunto de lançamentos de eventos que ocorrem no decorrer do TCC, sempre que ocorrer: término previsto, atraso, antecipação ou cancelamento, considerando o início e o fim de uma atividade. Um lançamento é constituído: da identificação da atividade, sua descrição, sua data de início e sua data de fim, conforme proposto no Cronograma. Segue o status (término conforme cronograma, atraso, antecipação ou cancelamento). Caso o término não seja o esperado, devem ser incluídos: justificativa (o porquê do evento); encaminhamento (alteração do cronograma – pode ser apenas a proposta de uma nova data de fim, por conta de um atraso, ou o cancelamento da atividade); e consequência (análise e alteração das atividades ainda não encerradas por conta do encaminhamento decidido). Esses lançamentos serão úteis para a escrita da monografia.