FACULDADE DE ENGENHARIA DE COMPUTAÇÃO

PROJETO FINAL I e II

PLANO DE TRABALHO

SFAnalytics

Lucas Carvalho Roncoroni

Edmar Roberto Santana de Rezende

16/05/2017

INTRODUÇÃO

Quase um milhão de malwares são criados todos os dias (CNN, 2015) e ataques de 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 profissionais de segurança da informação a identificar novas ameaças é de extrema importância.

Malwares são programas maliciosos que danificam e, ou, executam ações indesejadas em um computador. A identificação de um malware, quando existe uma assinatura, é feita por um antivírus. No caso das novas ameaças, é necessário que sua identificação seja feita por aqueles profissionais, só depois disso pode ser gerada uma assinatura.

CARACTERIZAÇÃO DE PROBLEMAS E OBJETIVO (S)

Para a identificação de novos malwares, especialistas utilizam ferramentas de análise de malware. Estas ferramentas extraem informações do programa que ajudem um profissional da área de segurança a classificar um programa em maliciosos ou não.

O problema é que, estas ferramentas, em sua grande maioria, não possuem interface gráfica, tem saídas de dados muito extensas e nem sempre fica claro o que está sendo extraído do programa, o que leva o especialista a gastar muito tempo aprendendo como utilizar a ferramenta.

Sendo assim, este trabalho tem como objetivo facilitar a identificação de programas maliciosos com uma ferramenta de análise que possua interface gráfica e extraia dados de um executável, apresentando-os de forma a facilitar a análise.

PLANO DE AVALIAÇÃO DO TRABALHO

Para a avaliação da ferramenta é necessário que um profissional da área submeta algumas amostras para familiarizar-se, e ver seu funcionamento.

Após o uso da ferramenta o profissional responderá um questionário, onde ele apontará pontos positivos e negativos da ferramenta, dizendo como a ferramenta se compara com outras que já existem que ele já tenha utilizado.

O projeto será considerado bem-sucedido se obtiver uma avaliação positiva do profissional de segurança da informação, de acordo com suas respostas ao questionário.

PROPOSTA DO ARTEFATO

O artefato deste trabalho consiste em um classificador de arquivos do tipo PE, executáveis do sistema operacional Windows, da arquitetura x86, em maliciosos ou não através do aprendizado supervisionado de máquina com features de opcodes, dlls e strings do programa.

O artefato terá duas telas de submissão, Upload para classificação e Upload para aprendizagem. Na primeira um executável é submetido e dele serão extraídos dados, alguns desses dados serão transformados em features para classificar o executável, todos os dados extraídos serão apresentados em uma interface gráfica, junto com sua classificação. Na segunda, será feito um upload com duas pastas, uma com malwares, outra, com programas não maliciosos; após o upload será realizada a aprendizagem com estes executáveis, a aprendizagem terá como resultado, uma árvore de decisão, que servirá para classificar os malwares submetidos no Upload para classificação.

Após a o Upload para classificação, a Visualização das características do executável é acionada, mostrando características extraídas deste e sua classificação. Nesta tela também haverá uma opção para ver regras extraídas da árvore de decisão. Também haverá a opção de ver alguns executáveis na base que contenham características parecidas com este que foi classificado; está opção acionará uma pesquisa na base por um executável parecido.

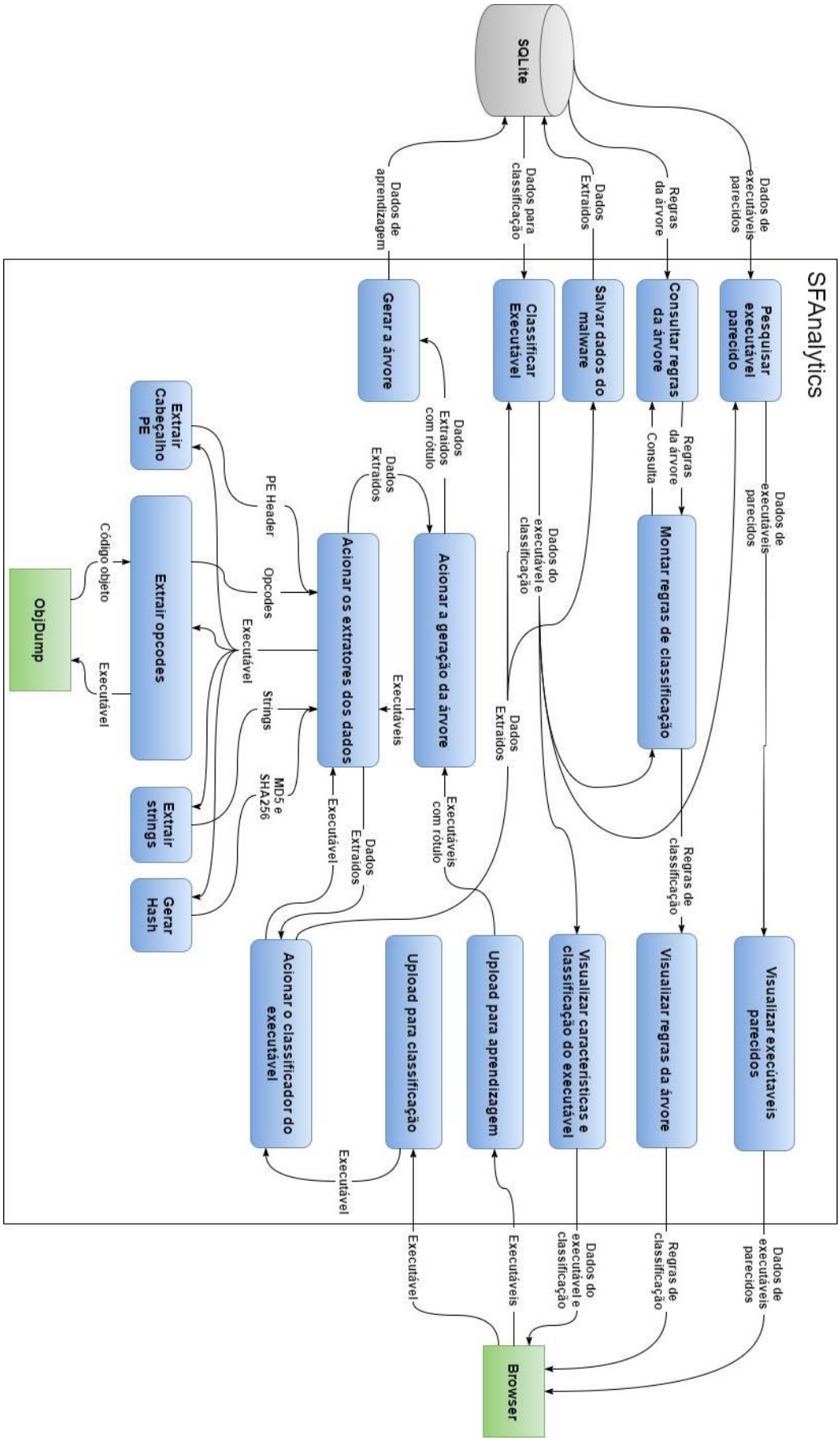

Figura 1 – Diagrama de Arquitetura. Fonte: O próprio autor.

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

- *Malwr*: é um serviço gratuito para análise de malware, onde qualquer arquivo pode ser submetido para análise. A ferramenta faz a execução do malware, mostra informações de rede, modificações no registro e submete o malware a alguns antivírus.
- *VxStream*: é uma ferramenta criada pela Payload Security para análise de um malware. A ferramenta também aceita arquivos pdf, e faz descrições em alto nível do arquivo analisado. Algumas de suas funcionalidades só estão disponíveis na versão paga.

MÉTODO DE DESENVOLVIMENTO

O método de desenvolvimento escolhido foi o Scrum. O Scrum consiste em Times Scrum e seus papéis eventos e artefatos (SCRUMGUIDE, 2016). O Time Scrum é composto por Product Owner, Development Team e Scrum Master (SCRUMGUIDE, 2016). O Scrum prescreve quatro eventos formais: Sprint Planning, Daily Scrum, Sprint Review e Sprint Retrospective (SCRUMGUIDE, 2016), como mostra a figura 2. O Scrum define três artefatos, o Product Backlog, Sprint Backlog e o increment (SCRUMGUIDE, 2016), mostrados também na figura 2, sendo increment o incremento.

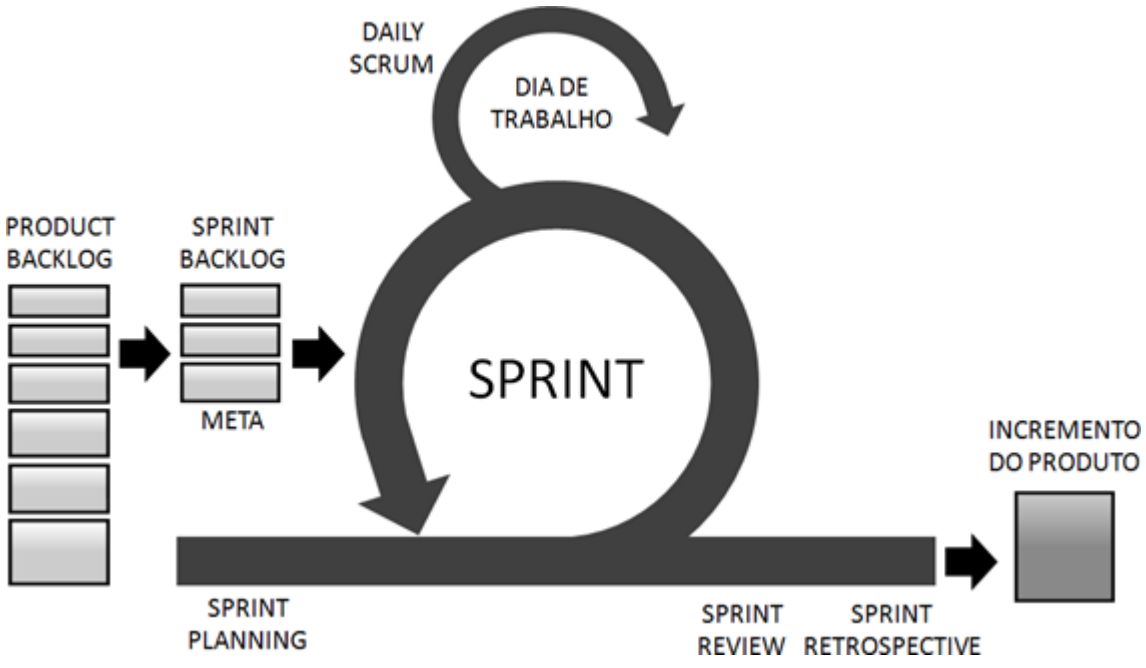

Figura 2 – Mecânica e ciclo do Scrum. Fonte: DEVMEDIA (2015).

O Product Backlog é uma lista que contém tudo o que o produto deverá ter (DEVMEDIA, 2015). No Scrum o Product Backlog é responsabilidade do Product Owner. Neste projeto o cliente fará o papel de Product Owner, mas não será responsável por fazer o Product Backlog.

O Sprint Backlog é um apanhado de itens selecionados do Product Backlog (SCRUMGUIDE, 2016), ele serve como um guia para o Development Team, ou time de desenvolvimento, para saberem o que deve ser feito durante o Sprint. Neste projeto o Development Team será somente o autor do projeto.

O Scrum Master, deve garantir o progresso do projeto (DEVMEDIA, 2015), mantendo a comunicação, monitorando o trabalho e organizando as reuniões. Neste projeto, este papel será desempenhado como um conjunto entre orientador, coorientador, cliente e autor do projeto. Orientador, coorientador e cliente sempre acompanharão o projeto, auxiliando, para o progresso do projeto, sendo que qualquer um deles pode marcar reuniões se achar necessário.

O Scrum define o Sprint Planning, a Daily Scrum, o Sprint Review e o Sprint Retrospective como os únicos eventos, ou reuniões, durante o desenvolvimento do projeto. Neste projeto haverá apenas uma reunião fixa a cada duas semanas, que desempenhará o mesmo papel que as reuniões de Sprint Planning, Sprint Review e Sprint Retrospective.

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	Realização do diagrama de arquitetura	22/03/17	24/05/17
A5	Realização do plano de trabalho	22/03/17	16/06/17
A6	Escolha do algoritmo de aprendizagem	22/03/17	05/05/17
A7	Escolha das features de aprendizagem	30/03/17	05/05/17
A8	Instalação das ferramentas para implementação do projeto	23/04/17	25/04/17
A9	Implementação do upload para classificação	25/04/17	05/05/17
A10	Validação da tela de upload com o cliente	05/05/17	07/05/17
A11	Implementação da extração de opcodes	07/05/17	13/05/17
A12	Implementação da extração de strings	13/05/17	15/05/17
A13	Implementação da geração de Hashs	15/05/17	16/05/17
A14	Implementação da extração do cabeçalho PE	16/05/17	22/05/17
A15	Implementar a geração de features	22/05/17	01/06/17
A16	Fazer a apresentação da banca	01/06/17	09/06/17
A17	Implementação da tela de visualização de características do executável	01/06/17	09/06/17
A18	Validação da tela de características do executável com o cliente	09/06/17	12/06/17
A19	Revisão da apresentação para a banca	12/06/17	23/06/17

A20	Revisão do artefato para a banca	12/06/17	23/06/17
A21	Implementação do upload para aprendizagem	31/06/17	10/07/17
A22	Implementação da árvore de decisão	10/07/17	20/07/17
A23	Implementação da classificação do executável	20/07/17	30/07/17
A24	Implementação da extração de regras da árvore	31/07/17	20/08/17
A25	Implementação da visualização das regras da árvore	20/08/17	31/08/17
A26	Validação da visualização das regras da árvore	20/08/17	31/08/17
A27	Implementação da pesquisa de executáveis parecidos	01/09/17	20/09/17
A28	Implementação da visualização de executáveis parecidos	20/09/17	30/09/17
A29	Validação da visualização de executáveis parecidos	20/09/17	30/09/17
A30	Montar questionário de avaliação do artefato	30/09/17	18/10/17
A31	Avaliar e validar TCC	30/09/17	18/10/17
A32	Escrever monografia	10/10/17	27/11/17
A33	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	X	X	X	X	X	X	X					
A5								X	X	X	X	X	X	X	X	X	X	X	X	X		
A6								X	X	X	X	X	X	X								
A7									X	X	X	X	X	X								
A8													X									
A9													X	X								
A10														X	X							
A11															X							
A12															X	X						
A13																X						
A14																X	X					
A15																	X	X				
A16																		X	X			
A17																		X	X			
A18																			X	X		
A19																				X	X	
A20																				X	X	
A21																						X
A22																						
A23																						
A24																						
A25																						
A26																						
A27																						
A28																						
A29																						
A30																						
A31																						
A32																						
A33																						

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																					
A10																					
A11																					
A12																					
A13																					
A14																					
A15																					
A16																					
A17																					
A18																					
A19																					
A20																					
A21	X	X																			
A22		X	X																		
A23			X	X																	
A24				X	X	X	X														
A25							X	X													
A26							X	X													
A27									X	X	X	X									
A28												X	X								
A29													X								
A30													X	X	X	X					
A31													X	X	X	X					
A32															X	X	X	X	X	X	X
A33															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	Módulo de upload do executável para classificação	A9
R4	Módulo de extração de opcodes	A11
R5	Módulo de extração de strings	A12
R6	Módulo da geração de hashs	A13
R7	Módulo de extração do cabeçalho PE	A14
R8	Apresentação da banca	A16
R9	Módulo de visualização de características do executável	A17
R10	Módulo de upload para aprendizagem	A18
R11	Módulo da árvore de decisão	A22
R12	Módulo de classificação do executável	A23
R13	Módulo de extração de regras da árvore	A24

R14	Módulo de visualização de regras da árvore	A25
R15	Módulo de pesquisa de executáveis parecidos	A27
R16	Módulo de visualização de executáveis parecidos	A27
R17	Questionário de avaliação do artefato	
R18	Monografia	A31
R19	Defesa do TCC	A32

RECURSOS MATERIAIS

Recursos de hardware:

Notebook.

Recursos de software:

Atom;

GitHub;

Google Drive;

Objdump.

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**
Extrair conhecimento adquiridos	Alto

Extrair conhecimento adquiridos – O artefato deve mostrar os conhecimentos adquiridos pelo algoritmo de aprendizado de máquina para fazer a classificar um programa. 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
pefile	Alto

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.

pefile – Biblioteca em Python que extrai informação do cabeçalho PE. O grau é considerado alto devido à grande quantidade de informações no cabeçalho, o que demanda estudo extenso sobre o cabeçalho e sobre como conseguir estas informações com a biblioteca.

ANÁLISE DE RISCOS

Entrada	Probabilidade	Risco	Alternativa
Objdump	Baixa	Leve	Existem outras ferramentas que fazem a mesma coisa.
GitHub	Baixa	Leve	Escolher outra plataforma para backup e versionamento.
Google Drive	Baixa	Leve	Escolher outra plataforma para backup e versionamento.
Django	Alta	Médio	Pedir ajuda ao orientador, por ter mais conhecimento sobre o framework.
Notebook do autor	Média	Médio	Fazer backups durante todo o projeto.

OUTRAS OBSERVAÇÕES

Para o controle de versionamento e backup estão sendo usados o github para armazenamento do artefato e o Google Drive para o armazenamento de documentos.

No caso dos documentos, o Google Drive tem uma ferramenta no Windows, que após instalada, cria uma réplica local do que está no servidor. A ferramenta monitora as cópias, atualizando automaticamente o servidor com uma nova versão do arquivo se ele for modificado. Sendo assim não é necessária nenhuma rotina para manter o servidor atualizado, a própria ferramenta cuida disso.

No caso do github, as atualizações do servidor são feitas manualmente, fica sendo papel do usuário realizar a atualização dos dados do servidor. Todos os dias que houverem modificações no código, deve ser feito pelo menos uma atualização ao servidor, para diminuir os riscos de perda de modificações feitas localmente.

REFERÊNCIAS

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.

Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais, 11, 2011, Brasília. SBSeg 2011 Brasília: Sociedade Brasileira de Computação, 2011. 280 p.

SCRUMGUIDES. *The Scrum Guide*. Disponível em: <http://www.scrumguides.org/docs/scrumguide/v2016/2016-Scrum-Guide-US.pdf#zoom=100>. Acesso em 18 de abril de 2017.

DEVMEDIA. *Introdução ao Scrum*. Disponível em: <http://www.devmedia.com.br/introducao-ao-scrum/33724>. Acesso em 18 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.
Feature – propriedade mensurável de um indivíduo sendo observado.

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

PE – Portable Executable

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.