Detecção de URLs Maliciosas

JULIO RODRIGUES, Universidade Federal de São João del-Rei, Brasil

1 INTRODUÇÃO

De modo geral, o elo mais fraco na cibersegurança é o ser humano, mais especificamente o usuário. URLs são vias rápidas e diretas para possibilitar ações maliciosas na internet, bastando no máximo alguns cliques para efetivar um ataque. Por isto, o potencial de infecção por este meio é exponencial, embora a efetividade desses ataques dependa de características adicionais da URL e das técnicas de camuflagem utilizadas.

Em 2019, o Brasil estava no Top 15 entre os países com o maior número de vítimas de ataques com URLs maliciosas [1]. Atualmente, estas estatísticas possivelmente podem apresentar-se bem piores, devido ao período transcorrido da pandemia de *Covid-19*, com a elevação no uso da tecnologia de modo geral. É um campo em que existe uma constante evolução dos dois lados conflitantes. O lado de combate, aprimorando as técnicas para detecção, e o lado de ataque, evoluindo as técnicas de camuflagem das URLs.

No caso do *phishing*, por exemplo, principalmente nos casos de ataques por email, geralmente a intenção é sempre provocar alguma urgência no assunto, instigando o acesso da vítima à página sem nenhuma checagem. Por isto, são utilizadas técnicas de engenharia social na tentativa de assemelhar ao máximo URLs maliciosas com URLs seguras. É basicamente uma verdadeira guerra entre os lados, alternando entre invenção de novas técnicas e detecção das novas técnicas, isto tudo na esperança de um dia não existirem mais opções viáveis para formular ataques deste tipo.

2 MOTIVAÇÃO

Em relação à motivação para execução deste trabalho, até o momento existe uma pergunta central a qual deseja-se responder:

- RQ1: *Quais são as principais características que definem a natureza de uma URL?*

O intuito de trabalhar sobre esta questão única é entender:

(1) *"Em que aspectos cada URL se destaca?"*;
(2) *"Que informações são suficientes para distinção entre URLs seguras e maliciosas?"*;
(3) *"O que separa **phishing** de **malware**?"*.

Estas são as questões as quais se dará maior atenção no desenvolvimento deste trabalho, e as quais se espera responder, cumprindo ao menos os objetivos iniciais definidos.

Manuscript submitted to ACM

3 CRONOGRAMA

Nesta seção, serão citados os trabalhos relacionados que servirão como base inicial para os primeiros passos no desenvolvimento deste trabalho. Também será apresentado um cronograma de desenvolvimento provisório, o qual possivelmente poderá sofrer alterações diversas, dependendo das direções seguidas (ou eventuais atrasos) ao longo da produção do trabalho.

3.1 Trabalhos Relacionados

Inicialmente, foram selecionados três artigos para auxiliar no desenvolvimento. Dois deles [3] [4] estão diretamente relacionados com o problema à ser tratado neste trabalho: detecção de URLs maliciosas. O segundo [2] está relacionado à uma abordagem que se pretende avaliar em estágios mais avançados do desenvolvimento: extração de *meta-features*.

3.2 Desenvolvimento

Para primeira apresentação parcial, pretende-se agregar mais bases de dados e continuar a análise exploratória de forma mais extensa. Já no intervalo entre as entregas parciais, os esforços seriam concentrados na criação de novos atributos. No entanto, o foco não seria na parte léxica da URL, e sim, no conteúdo, tentando extrair informação útil do código HTML da página, ou até mesmo de dados relacionados à rede. O objetivo neste ponto, é encontrar atributos relevantes que possam definir de forma mais clara as classes.

Para a parcial 2, cessada a etapa de construção dos atributos, iniciaria-se o processo de análise e seleção dos mesmos, identificando a relevância de cada um no(s) modelo(s). Em seguida, seriam selecionados um conjunto de algoritmos de *machine learning* para executar os treinamentos iniciais, e realizar um comparativo com outros trabalhos da literatura, destacando similaridades, limitações, pontos positivos e negativos.

Por fim, a última etapa seria dedicada para a exploração da extração de *meta-features*, processo o qual ainda se faz um tanto quanto nebuloso no estágio de proposta deste trabalho. No entanto, indenpendentemente dos avanços e resultados obtidos com as *meta-features*, nesta etapa se dará a construção do modelo final, para então validar os resultados e finalizar a escrita do artigo, concluindo o trabalho. Todas as etapas de desevolvimento estão sumarizadas na Tabela 1.

Tabela 1. Cronograma de Desenvolvimento

Data	Atividade	Desenvolvimento
25/05/2023	Parcial I	Análise das Bases
06/06/2023	—	Criação de Atributos
13/06/2023	Parcial II	Comparativo com Trabalhos
15/06/2023	—	*Meta-features*
27/06/2023	Final	Modelo Final e Artigo

REFERÊNCIAS

[1] Abranet. 2019. Relatório aponta que cada URL maliciosa no Brasil afeta 18 usuários. https://www.abranet.org.br/Noticias/Relatorio-aponta-que-cada-URL-maliciosa-no-Brasil-afeta-18-usuarios-2585.html?UserActiveTemplate=site.

[2] Edesio Alcobaça, Felipe Siqueira, Adriano Rivolli, Luís P. F. Garcia, Jefferson T. Oliva, and André C. P. L. F. De Carvalho. 2020. MFE: Towards Reproducible Meta-Feature Extraction. *J. Mach. Learn. Res.* 21, 1, Article 111 (jan 2020), 5 pages.

[3] Taeri Kim, Noseong Park, Jiwon Hong, and Sang-Wook Kim. 2022. Phishing URL Detection: A Network-Based Approach Robust to Evasion. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Los Angeles, CA, USA) *(CCS '22)*. Association for Computing Machinery, New York, NY, USA, 1769–1782. https://doi.org/10.1145/3548606.3560615

[4] Ch Rupa, Gautam Srivastava, Sweta Bhattacharya, Praveen Reddy, and Thippa Reddy Gadekallu. 2021. A Machine Learning Driven Threat Intelligence System for Malicious URL Detection. In *Proceedings of the 16th International Conference on Availability, Reliability and Security* (Vienna, Austria) *(ARES 21)*. Association for Computing Machinery, New York, NY, USA, Article 154, 7 pages. https://doi.org/10.1145/3465481.3470029

Received 09 Maio 2023