Uma Plataforma para Correlação, Visualização e Extração de Tendências de Dados de Segurança

Fernanda Yukari Kawasaki
Departamento de Informática
Universidade Federal do Paraná
Curitiba, PR, Brasil
fyk18@inf.ufpr.br

Ulisses de Oliveira Penteado Junior
BluePex Tecnologia
Limeira, SP, Brasil
ulisses@bluepex.com.br

André Ricardo Abed Grégio
Departamento de Informática
Universidade Federal do Paraná
Curitiba, PR, Brasil
gregio@inf.ufpr.br

RESUMO

Digital data production increases on a daily basis due to the widespread use of Internet and interconnected computing devices. Acting security mechanisms may result in several types of information, if processed and correlated, since their output logs range from IP addresses/regions to attack events. Hence, the application of data science techniques is essential to extract knowledge and insights from this massive amount of data. In this article, we present a platform for cybersecurity data visualization in an effort to identify trends, associations and patterns, which enable better data-driven decisions. To prototype and test the proposed platform, we focused on endpoint logs provided from a cybersecurity company. The conclusions drawn from this study are that there is a substantial concentration of victims in urbanized areas, notably the state capitals, as well as a higher risk level for Server Operating Systems. Nevertheless, these results are still preliminary, considering the limitations of the dataset (few, specific samples from internal testing endpoints), but helped pave the way towards new models for further threat analysis.

KEYWORDS

Ciência de Dados, Segurança Computacional, Antivírus, *Malware*

1 INTRODUÇÃO

A popularização da Internet e a evolução da capacidade de processamento e armazenamento de dispositivos computacionais causaram um aumento massivo na quantidade de dados produzidos e em tráfego. Tais dados representam o conteúdo de comunicações e informações associadas, como data de criação, localização geográfica, origem e destino. A área de segurança da informação é fortemente dependente da existência de registros (*logs*) para se alcançar a descoberta de conhecimento, mas a quantidade deles dificulta a identificação de associações entre dados de diferentes fontes por analistas humanos. Há dispositivos centralizadores que proveem um ambiente integrado para coleta de *logs*, os UTMs (*Unified Threat Management*), que podem auxiliar nessa tarefa quando colocados em pontos estratégicos de uma rede para monitoração contínua (*endpoints*).

O uso de técnicas de visualização, em particular, é um elemento viabilizador para a interpretação dos dados de segurança, permitindo que um analista humano possa perceber eventuais padrões existentes nos *logs* coletados. Além disso, o processamento dos dados para torná-los apropriados para visualização gera um formato que facilita a aplicação de algoritmos que os correlacionem e explicitem *outliers* no meio da massa de atividades inofensivas ou comuns. Com isso, viabiliza-se também a extração de tendências presentes nesses dados, o que permite a detecção de eventos que podem estar perdidos e a rápida resposta a incidentes de segurança. Um exemplo disso seria a criação de regras de bloqueio dinâmicas ou listas de URLs de acordo com a necessidade do momento, como é o caso de ataques sazonais (*malware* sobre IRPF ou COVID-19) ou *sites* infectados provendo conteúdo malicioso. Neste artigo, apresenta-se uma plataforma para automatização das atividades de correlação e visualização de dados provenientes de dispositivos de segurança, a fim de auxiliar na identificação de tendências de ameaças e tomada de decisões. A principal contribuição esperada é mostrar como dados distintos podem ser combinados para se extrair tendências de ameaças e promover a consciência situacional de forma a fomentar a prevenção de ataques similares.

2 CONCEITOS/TRABALHOS RELACIONADOS

Dados de segurança podem ser provenientes de inúmeras fontes, sejam elas aplicações, mecanismos ou dispositivos. Por exemplo, um antivírus (AV) provê alertas (texto) de assinaturas ou heurísticas comportamentais, arquivos quarentenados (binários) ou informações gerais sobre uma infecção em formatos diversos (JSON, XML, IoC, tráfego de rede). Todos esses dados devem ser tratados e processados para geração de informação útil. Isso é possível através da aplicação do processo conhecido como ciência de dados [1]. Tal método é composto pelos seguintes passos, cujos detalhes são inspirados no CRIPS-DM [2]:

(1) **Coleta**, ou obtenção dos dados (neste artigo, provenientes de um UTM que monitora máquinas em rede e emite alertas sobre/de AVs;
(2) **Limpeza**, ou preparação dos dados para formatação e remoção de ruídos ou dados corrompidos/incompletos;
(3) **Análise exploratória**, ou descoberta da distribuição dos dados e tipo de informação eles podem prover;
(4) **Modelagem**, ou abstração que represente os dados em um formato de entrada para etapas futuras, como um classificador;
(5) **Implantação**, ou aplicação do modelo em produção a fim de se avaliá-lo contra dados em constante evolução, reais, não previamente rotulados.

Cabe ressaltar que os resultados apresentados neste trabalho abrangem os passos de 1 a 3 acima listados. A literatura traz diversas tentativas de se fazer correlação de dados e alcançar um estado de consciência situacional [3–5]. Entretanto, as abordagens são dependentes de diversos padrões de troca de dados de segurança

que podem não ser utilizados, ou estarem mal-configurados ou indisponíveis. Há também trabalhos que propõem novas formas de se visualizar dados de segurança [6], ou de avaliar *frameworks* de visualização [7].

3 PROJETO E IMPLEMENTAÇÃO

O protótipo da plataforma foi implementado em Python, sendo que as bibliotecas mais relevantes foram Pandas, Plotly e BeautifulSoup, que serviram de base para o processamento, a visualização e a obtenção de dados, respectivamente. Dentro da plataforma, foi adotado um processo simplificado de análise dos dados, que foram recebidos em formato JSON e SQL dumps e transformados em um *DataFrame*. O conteúdo deles compreende os seguintes campos: data e hora, versão do Windows, presença de licenciamento do sistema operacional, nome dos artefatos e seu nível de risco e o status do antivírus (instalado, habilitado e atualizado). Inicialmente, é feita uma breve análise para determinar a integridade dos dados e verificar características como consistência e completude (ex.: o rótulo "`EICAR-Test-File (not a virus)`" foi desconsiderado por se tratar de um arquivo de teste não relevante para a identificação de infecção em uma máquina).

Em seguida, o banco de dados Db-IP é utilizado no processo de georreferenciamento de endereços de IP. Dessa forma, é possível extrair as coordenadas e o nome da localização desde o nível federal até o municipal. Entretanto, o banco contém algumas informações imprecisas, principalmente os nomes das cidades. Como parte da análise é focada na incidência de ataques no Brasil, foi feito o uso de um *dataset* público para normalização dos nomes. Ademais, um processo semelhante é feito com o histórico de versões do Windows, que foi extraído por meio da técnica de *Web Scraping*. Após os passos acima, os dados estão prontos para serem analisados e plotados. Neste ponto, são utilizadas correlações e técnicas de visualização para identificar padrões. A dinâmica de funcionamento de um `ipynb` (modo como os componentes da plataforma foram prototipados) propicia a criação de esboços que podem ser alterados rapidamente.

O próximo passo é o aperfeiçoamento dos gráficos, a fim de garantir que eles representem os dados de forma verídica e legível. Com esse intuito, parâmetros como as cinco qualidades e boas visualizações [8] e o termo *chart junk* [9] foram considerados. Dentre os problemas encontrados até o momento, a falta de um padrão de nomenclatura para rotulação de códigos maliciosos se sobressai, pois dificulta a classificação destes em tipos e famílias. Uma ferramenta que contorna tal problema é o AVClass [10], que normaliza um rótulo com base em diversas entradas de AVs distintos. Contudo, a entrada esperada pelo algoritmo é um arquivo JSON proveniente da API *VirusTotal*. Por isso, a presença de mais de um rótulo e seu *hash* correspondente (no mínimo) é mandatória. A continuidade deste trabalho envolve modelar regras para criação de um normalizador de tipos/famílias de *malware* para quando se tem apenas um rótulo.

4 SOLUÇÃO PROPOSTA

Foram coletados 72.6 MB de dados em formato JSON, contendo 80778 entradas válidas que representam *logs* obtidos de AVs instalados em 8707 instituições únicas (determinadas pelo IP de seus respectivos *gateways*). Após processamento pelos módulos da plataforma, foram obtidos os seguintes resultados preliminares. Nas Figuras 1 e 2, observa-se incidência maior de ataques nas regiões cuja a taxa de urbanização é maior, em particular as capitais estaduais. Como referência, a Figura 3, que mostra a distribuição populacional urbana em 2010. Segundo a PNAD, em pesquisa realizada pelo IBGE entre 2017 e 2018, somente 46,5% dos domicílios rurais possuíam acesso à Internet, enquanto que para domicílios urbanos este valor era de 79,4% [11]. Dentre os motivos para a não utilização, o mais discrepante entre eles é a disponibilidade de acesso a este tipo de serviço: 20% e 1% para os rurais e urbanos, respectivamente.

Ademais, na figura 2 é possível perceber um padrão semanal. Durante os dias úteis, a quantidade de detecções alcança seus picos (dias em que a maioria das pessoas trabalha), enquanto nos finais de semana a tendência é de queda. Nota-se também a presença de outliers, que podem ser utilizados para evidenciar falhas. Na semana do dia primeiro de Fevereiro, por exemplo, a quantidade de artefatos de classe Trojan detectados é bastante divergente da maioria dos dias, pois ocorreu um problema no servidor durante esse intervalo de tempo.

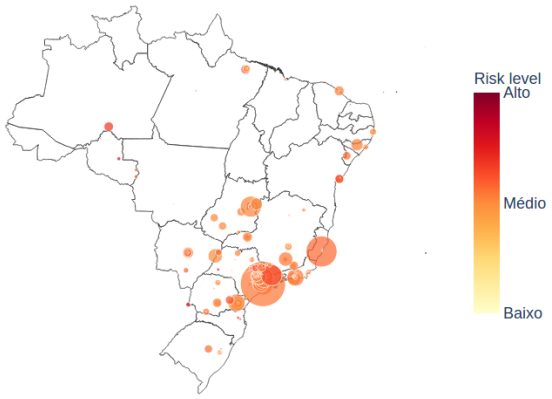

Figura 1: Incidência de artefatos e média de nível de risco representados por círculos e escala de cores, respectivamente.

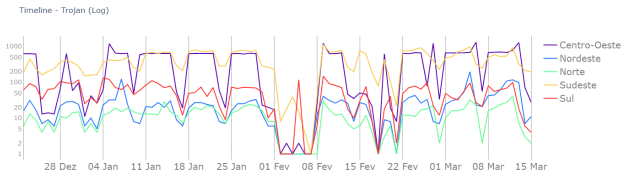

Figura 2: Linha do tempo com distribuição de artefatos da classe Trojan por região

Na Figura 4, nota-se que os maiores níveis de risco são encontrados em versões do Windows para servidores, já que estes possuem maior estabilidade, como tempo ligado e melhor internet de banda larga. Esse ambiente é propício e atrativo para invasores, que podem usar o servidor como base para outros ataques, e até mesmo

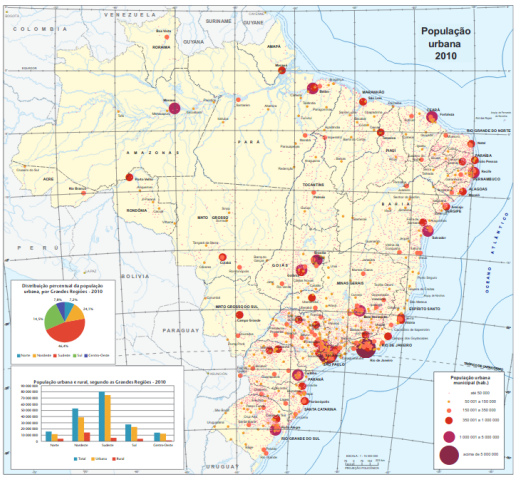

Figura 3: Distribuição da população urbana em 2010 [12]

como fonte de armazenamento de malwares e de conteúdos ilegais. Por fim, a figura 5 representa um grafo, onde verifica-se as relações da classe e suas subdivisões.

Apesar das análises anteriores, é importante ressaltar que estes resultados são preliminares. O volume de dados utilizado é pequeno, e assim, não é possível fazer inferências fidedignas. Além disso, o espaço amostral utilizado é limitado, pois possui somente dados referentes aos ataques e, por isso, só seria possível fazer afirmações sobre a população de vítimas destes. Vale ressaltar também que os dados são referentes aos clientes de uma empresa de cybersegurança, e portanto podem diferir da distribuição real.

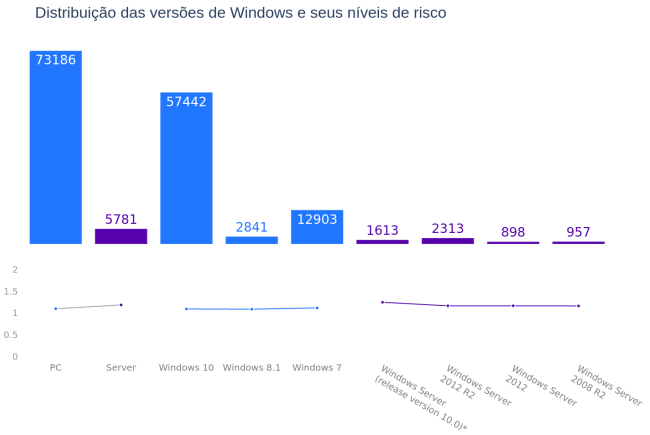

Figura 4: Distribuição dos artefatos entre as versões de Windows e as respectivas médias de nível de risco.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, foram expostos resultados de uma plataforma para visualização de tendências de dados de segurança, mais especificamente providos por AVs ligados a um UTM. Tal plataforma encontra-se em desenvolvimento, mas testes preliminares com um conjunto limitado de dados reais de organizações monitoradas mostram seu potencial de descoberta de tendências. A comparação entre abordagens da literatura e esta proposta, bem como a validação da plataforma de visualização e continuidade das etapas faltantes do processo de ciência de dados são trabalhos futuros.

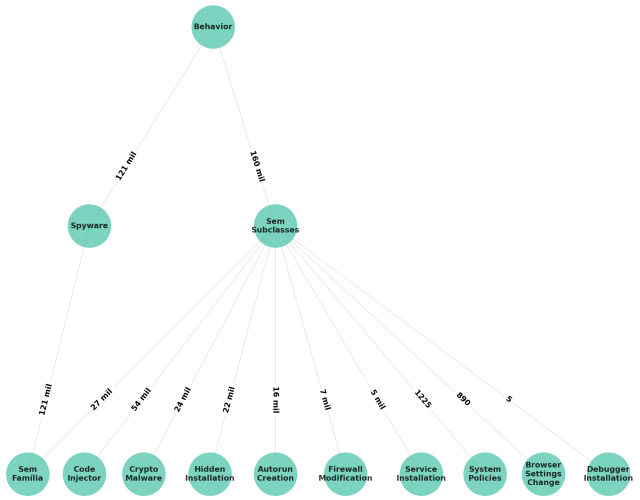

Figura 5: Grafo das relações entre a classe Behavior, suas subclasses e famílias.

REFERÊNCIAS

[1] Joshua Saxe and Hillary Sanders. *Malware Data Science.* No Starch Press, 2018.
[2] Chanin Nantasenamat. The data science process: A visual guide to standard procedures in data science, Jul. 2020. URL https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b.
[3] A. Atifi and E. Bou-Harb. On correlating network traffic for cyber threat intelligence: A bloom filter approach. In *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 384–389, 2017. doi: 10.1109/IWCMC.2017.7986317.
[4] X. Jin, B. Cui, J. Yang, and Z. Cheng. An adaptive analysis framework for correlating cyber-security-related data. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 915–919, 2018. doi: 10.1109/AINA.2018.00134.
[5] G. Settanni, Y. Shovgenya, F. Skopik, R. Graf, M. Wurzenberger, and R. Fiedler. Correlating cyber incident information to establish situational awareness in critical infrastructures. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, pages 78–81, 2016. doi: 10.1109/PST.2016.7906940.
[6] C. N. Adams and D. H. Snider. Effective data visualization in cybersecurity. In *SoutheastCon 2018*, pages 1–8, 2018. doi: 10.1109/SECON.2018.8479113.
[7] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. An evaluation framework for network security visualizations. *Computers Security*, 84:70 – 92, 2019. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2019.03.005. URL http://www.sciencedirect.com/science/article/pii/S0167404818308952.
[8] Alberto Cairo. *The Truthful Art: Data, Charts, and Maps for Communication.* New Riders, 2016.
[9] Edward Tufte. *The Visual Display of Quantitative Information.* Graphics Pr, 1983.
[10] Silvia Sebastián and Juan Caballero. Avclass2: Massive malware tag extraction from AV labels. In *ACSAC '20: Annual Computer Security Applications Conference, Virtual Event / Austin, TX, USA, 7-11 December, 2020*, pages 42–53. ACM, 2020. doi: 10.1145/3427228.3427261. URL https://doi.org/10.1145/3427228.3427261.
[11] IBGE. Pesquisa nacional por amostra de domicílios (pnad), Dezembro 2018. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv101705_informativo.pdf.
[12] IBGE. *Atlas do censo demográfico 2010.* 2013. ISBN 9788524042812. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv64529_cap6.pdf.