Modelagem das Áreas de Risco de Sistemas de Detecção de Intrusão para Cálculo de Métricas de Privacidade

Jessica Yumi Nakano Sato, Daniel Macêdo Batista

[1]Departamento de Ciência da Computação
Universidade de São Paulo (USP)

jyns1703@usp.br, batista@ime.usp.br

Abstract. *Although recent regulations require software developers to become severely concerned about privacy, recommendations related to this have been around for years. Despite the relatively old understanding that computational systems need to guarantee the user's privacy, it has been difficult to find works that evaluate the regulations in existing systems, mainly because privacy metrics vary with the application domain. This paper presents preliminary results from the modeling of risk areas aiming to measure privacy metrics of an IDS based on machine learning. It is shown that it was possible to adapt the principles of the literature to our domain.*

Resumo. *Embora regulamentações recentes exijam que desenvolvedores de software passem a se preocupar de forma severa com privacidade, recomendações relacionadas a isso existem há anos. Apesar do entendimento, relativamente antigo, de que sistemas computacionais precisam garantir a privacidade do usuário, tem sido difícil encontrar trabalhos que avaliem as regulamentações em sistemas existentes, principalmente porque métricas de privacidade variam com o domínio das aplicações. Este artigo apresenta resultados preliminares decorrentes da modelagem das áreas de risco visando a medição da privacidade de um IDS baseado em aprendizado de máquina. É mostrado que foi possível adaptar princípios da literatura para o nosso domínio.*

1. Introdução

Privacidade na Internet tem sido um assunto recorrente tanto na literatura científica quanto em literatura para o público geral, inclusive em mídias sociais online [Silva and França 2023]. Essa preocupação aumentou nos últimos anos principalmente por conta de regulamentações como a LGPD (Lei Geral de Proteção de Dados Pessoais) no Brasil e a GDPR (*General Data Protection Regulation*) na Europa [Iwaya et al. 2023]. Outro motivo que tem aumentado as discussões em torno de privacidade na Internet é o aumento da utilização de técnicas de inteligência artificial, principalmente aprendizado de máquina, para classificação de dados. Por exemplo, um sistema de detecção de intrusão (IDS – *Intrusion Detection System*) baseado em aprendizado de máquina, é uma ferramenta capaz de ler os pacotes entrantes da rede e identificar se eles trazem risco para ela ou não. Nele é necessário treinar o classificador com fluxos que sejam o mais próximo possível do real, para que ele tenha uma boa acurácia. Caso a detecção seja feita em um ambiente não confiável, isso pode levar ao vazamento de informações sigilosas contidas nesses pacotes [Mosaiyebzadeh et al. 2023].

Do ponto de vista de quem possui os dados, seria importante que ele(a) tivesse uma gama de ferramentas que garantisse a privacidade. Porém, mesmo que houvesse tal gama disponível, a tomada de decisão sobre qual delas usar dependeria de alguma comparação entre as opções, o que justifica a importância de uma métrica padronizada para julgar a garantia de privacidade dessas opções[1]. Há trabalhos que discorrem sobre métricas de um modo geral, independente do domínio. Porém, carece-se da aplicação dessas métricas gerais para atestar o aumento da privacidade em algum sistema que se proponha a isso.

Este artigo apresenta resultados preliminares de um projeto em andamento que visa medir a privacidade de IDSs baseados em aprendizado de máquina e que tenham sido construídos tendo privacidade como um dos objetivos. Ao término do projeto pretende-se avaliar, de forma quantitativa, se tais sistemas garantem a privacidade. Um passo inicial para essa quantificação é a modelagem das áreas de risco dos sistemas e tal passo, baseado na adaptação de metodologias existentes na literatura, é o que é apresentado neste artigo.

As demais seções deste artigo estão organizadas da seguinte forma: A Seção 2 apresenta trabalhos relacionados, destacando o principal usado como base. A Seção 3 descreve a metodologia que foi aplicada para a modelagem das áreas de risco de um IDS baseado em aprendizado de máquina. A Seção 4 conclui o artigo.

2. Trabalhos Relacionados

[Agarwal 2016] discute sobre avaliação de impacto de privacidade, ou PIA (*Privacy Impact Assessments*), que é um processo que ajuda organizações a identificarem e gerenciarem os riscos de privacidade decorrentes de um novo projeto ou política. Argumenta-se que a PIA deve ser fácil e rápida, pois deve ser feita continuamente ao longo do processo de desenvolvimento de um sistema. Entretanto, a maioria das avaliações sobre privacidade resultam em longos relatórios complexos de analisar e comparar. Nesse sentido, o autor propõe uma nova métrica quantitativa para suprir as deficiências das métricas existentes.

[Wagner and Eckhoff 2018] avaliam as métricas disponíveis para medir PETs (*Privacy Enhancing Technologies*) e definem diversos conceitos relacionados, como violação de privacidade, as métricas propriamente ditas e domínios de privacidade. Oitenta métricas são revisadas no trabalho.

A pesquisa relatada em [Kioskli et al. 2022] foi utilizada como base para este artigo. Os autores modelam a segurança e a privacidade de um Laboratório Vivo (*Living Lab*) considerando os principais desafios e ataques desse contexto. Para fazer a modelagem do sistema, a metodologia *Secure Tropos* [Mouratidis and Giorgini 2007] foi utilizada. Essa metodologia foca principalmente em descrever o sistema baseando-se em agentes, ações e objetivos. Há três níveis de abstração da análise: **Visão Organizacional**, onde é representada a estrutura organizacional em que os Laboratórios vivos estão inseridos, dando ênfase a seus principais atuadores (agentes), objetivos e recursos utilizados; **Visão Mapeadora de Dados**, onde são representadas as ações executadas sobre os recursos; e por fim, **Visão *Privacy by Design (PbD)***, onde são modelados os principais riscos e restrições sobre/dos recursos. Essa visão é importante para guiar o desenvolvimento do sistema, uma vez que segue os princípios de PbD e ajuda a identificar quais medidas

[1]Inclusive, nos artigos 35 e 36 da GDPR exige-se que, em operações onde haja riscos à privacidade, esses sejam mensurados mesmo antes deles ocorrerem

e mecanismos devem ser tomados para garantir a privacidade dos usuários relacionados ao laboratório. Essas três visões são geradas de forma sequencial. Para cada visão, um diagrama baseado na visão anterior é criado, podendo ganhar ou perder alguns de seus elementos de forma que os principais aspectos de análise sejam mantidos em foco.

3. Resultados Preliminares

O principal trabalho realizado até o momento diz respeito à aplicação dos princípios da metodologia Secure Tropos proposta em [Mouratidis and Giorgini 2007], usando como exemplo a aplicação descrita em [Kioskli et al. 2022]. No nosso caso, o sistema considerado foi um IDS baseado em aprendizado de máquina para detecção de anomalias. Vale ressaltar que as metodologias anteriores tiveram que ser alteradas, uma vez que são mais comumente usadas durante a concepção do sistema, e não durante a análise posterior. As principais características das metodologias que foram utilizadas nesse projeto de análise foram: diagramas em diferentes níveis de abstração, preocupação com as diretrizes do *Privacy by Design* e identificação de pontos de risco de privacidade dentro do sistema. O nosso objetivo é modelar o tipo de IDS em questão para que possamos aplicar métricas de privacidade já existentes e que no entanto são restritas a áreas específicas de um sistema. Uma vez analisados os riscos de privacidade e as métricas mais comumente usadas para medi-los, podemos mensurar o IDS como um todo, por exemplo agregando as medidas parciais por meio de uma média ponderada.

Neste estudo, embora os IDSs já tenham sido construídos e possam não ter sido elaborados seguindo diretrizes do *Privacy by Design*, queremos analisar quão bem eles lidam com os aspectos de risco mapeados. Dessa forma, criamos diagramas que mostram os possíveis pontos de risco e restrições dos recursos para então analisarmos individualmente como esses pontos estão sendo tratados no sistema já pronto.

O primeiro diagrama, na Figura 1, mostra o fluxo dos dados que são utilizados pelo IDS e/ou que passaram para dentro da rede interna. Em um IDS baseado em aprendizado de máquina, teremos dois momentos: (i) treino do modelo e (ii) treino dos parâmetros do modelo para a rede local. Do lado direito da figura mostramos a etapa (i) em que dados de treino externo (fora da rede local) serão usados para treinar o modelo de detecção de anomalia. Essa etapa é geralmente feita por alguém fora da rede local, por exemplo pelo fornecedor do IDS, e conta com diversas etapas de treinamento e validação. Os dados de treinamento usados costumam ser públicos, podendo ser reais ou gerados artificialmente. Após o treino do modelo será feita a etapa (ii) para que o sistema possa ser implantado. Assim, dados da rede externa que desejam entrar para rede interna são coletados e usados para treinamento dos parâmetros do modelo. Após o treinamento, o IDS está pronto para ser usado encontrando possíveis anomalias que desejam adentrar a rede interna. O IDS pode também aprender de forma retroativa: frequentemente os dados entrantes serão coletados e usados para re-treinar os parâmetros do modelo.

Ao construir o diagrama da Figura 1 percebemos que o IDS lê uma ampla quantidade de pacotes entrantes da rede, possibilitando que diversas informações pessoais e confidencias dos usuários sejam acessadas. Tal situação pode trazer sérios riscos à privacidade do usuário. Pensando nessas questões e nos princípios defendidos pela LGPD em relação ao tratamento de dados as seguintes perguntas surgem:

Todos os dados coletados são de fato necessários ao funcionamento do IDS?

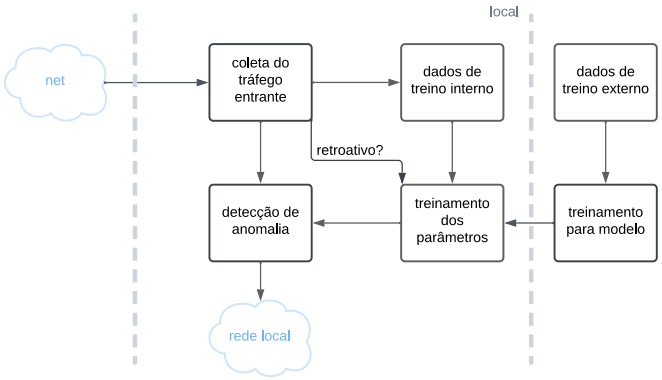

Figura 1. Fluxo dos dados no IDS

O ideal é que apenas os dados usados pelo IDS fossem coletados, e, mais do que isso, que o IDS requisitasse o mínimo de dados possíveis para fazer a classificação, já que, além de dados maliciosos, ele também têm acesso a dados normais. **Como identificar que dentro do IDS não está havendo uma utilização indevida (*re-purpose*) dos dados?** Essa pergunta foge do escopo do trabalho. Entretanto, algumas medidas para limitar a utilização imprópria dos dados podem ser discutidas, por exemplo possibilitar a análise do código e ter uma documentação ampla. **No caso de aprendizagem retroativa, o armazenamento dos dados poderia trazer mais risco à privacidade?** Se por um lado o armazenamento de dados pessoais poderia criar mais um ponto de risco à privacidade, por outro existem algumas técnicas de aprendizado online em que os dados ficam apenas provisoriamente armazenados [Arbex et al. 2021]. **Qual o impacto da privacidade dos dados de treino externos no sistema final?** Embora os dados de treino externo, a princípio, não digam respeito aos usuários da rede local e portanto não impactariam na privacidade dos dados deles, a etapa de treinamento do modelo está dentro da cadeia de produção do IDS e portanto deve ser considerada. Esse estudo e discussão no entanto também estão fora do escopo dessa pesquisa.

A Figura 2 mostra um dos diagramas baseados na metodologia Secure Tropos. Nele evidenciamos agentes, seus objetivos e os recursos disponíveis. Na legenda, o agente referencia a pessoa que deseja alcançar um certo objetivo, esse objetivo é alcançado fazendo operações que podem levar à execução de uma ação sobre um recurso. No caso de um IDS, temos que um dos agentes é o administrador da rede (à esquerda na figura) que, desejando melhorar a segurança, implanta e gerencia o IDS. Para isso ele deve treinar os parâmetros do modelo do IDS para sua rede local, colocá-lo em uso e, posteriormente, fazer uma análise de desempenho. Assim ele entra em contato com os dados entrantes da rede, tendo de armazená-los para treinamento e usá-los no IDS. Outro recurso acessado pelo administrador são os dispositivos em que o IDS será instalado. À direita da figura temos o agente treinador do modelo, que tem por objetivo achar um bom modelo de detecção de intrusão que seja capaz de detectar com precisão as anomalias. Para isso ele deve treinar e testar o modelo utilizando dados reais e/ou artificiais.

O estudo dos recursos na Figura 2 e como eles são utilizados é importante na análise de privacidade de um sistema. Em especial, no caso dos dados de um IDS, teremos que uma parte dos dados utilizados contém informações pessoais dos usuários da rede

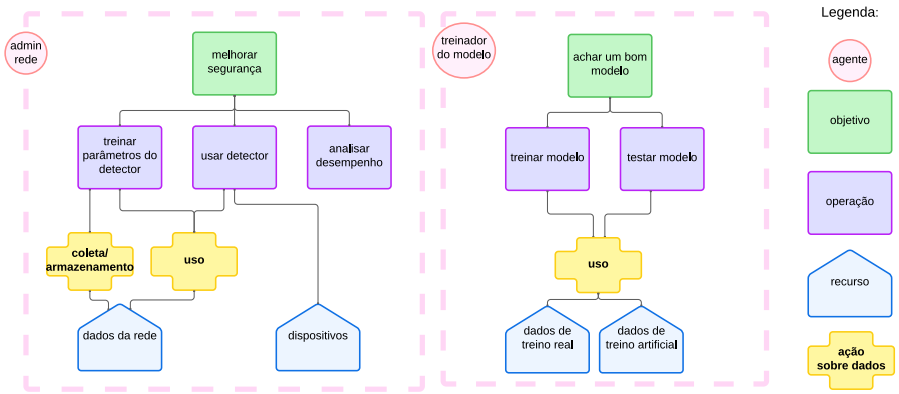

Figura 2. Visão de dados

[Elias et al. 2022] e portanto faz-se necessário compreender cada área que tem contato com os recursos para que seja feita a análise do risco relacionado a cada um deles.

Por fim, o diagrama da Figura 3 expõe riscos e restrições relacionados a cada recurso. Riscos são acontecimentos desencadeados por ações externas ao sistema e que trazem algum efeito prejudicial para os usuários. Restrições são características dos recursos internos do sistema que garantirão que estes não sejam mal utilizados. Neste caso, temos que os dados entrantes da rede e os dados reais de treinamento possuem riscos semelhantes: serem usados fora do propósito a que foram coletados e terem informações pessoais expostas. As restrições ligadas aos dados querem garantir que a privacidade dos indivíduos, cujas informações estão dentro dos dados, seja mantida, dessa forma os dados devem ser anônimos, confidenciais, conterem apenas as informações que serão usadas e serem usadas somente para isso. No caso dos dados de treinamento, a preocupação excede as capacidades de interferência do IDS mas ainda é preciso cuidado e atenção sobre os dados que são utilizados em todos os estágios de desenvolvimento de um detector.

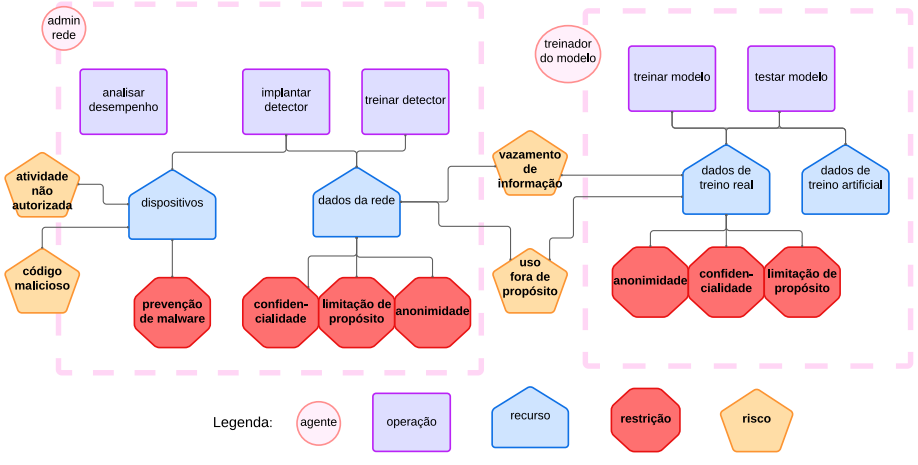

Figura 3. Visão de *Privacy by Design*

4. Conclusão e Próximos Passos

Os diagramas construídos contribuem para identificar os principais aspectos de privacidade que devem ser considerados ao desenvolver um IDS. Além disso, eles pontuam em

que momento os riscos surgem e quais as restrições que se deve ter sobre os dados para que a ocorrência deles possa ser diminuída. Para calcular uma métrica de privacidade para o sistema, tais características devem ser levadas em consideração e portanto esses diagramas contribuirão para a formulação dessas métricas. Como próximo passo, é necessário estimar quanto de cada restrição é cumprida pelo sistema. De acordo com essa medida, um valor geral para o sistema pode ser calculado. Para isso, no entanto, também será necessário saber quão importante é cada restrição no sistema geral e achar pesos que balanceiem bem essa relação.

Agradecimentos

Esta pesquisa é parte do INCT da Internet do Futuro para Cidades Inteligentes, financiado por CNPq (proc. 465446/2014-0), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e FAPESP (procs. 14/50937-1 e 15/24485-9). Também é parte do projeto FAPESP proc. 21/06995-0.

Referências

[Agarwal 2016] Agarwal, S. (2016). *Developing a Structured Metric to Measure Privacy Risk in Privacy Impact Assessments*, pages 141–155. Springer International Publishing, Cham.

[Arbex et al. 2021] Arbex, G. V., Machado, K. G., Nogueira, M., Batista, D. M., and Hirata, R. (2021). IoT DDoS Detection Based on Stream Learning. In *12th NoF*.

[Elias et al. 2022] Elias, E. M. d., Carriel, V. S., De Oliveira, G. W., Dos Santos, A. L., Nogueira, M., Junior, R. H., and Batista, D. M. (2022). A Hybrid CNN-LSTM Model for IIoT Edge Privacy-Aware Intrusion Detection. In *14th IEEE LATINCOM*.

[Iwaya et al. 2023] Iwaya, L. H., Babar, M. A., and Rashid, A. (2023). Privacy Engineering in the Wild: Understanding the Practitioners' Mindset, Organisational Aspects, and Current Practices. *IEEE Transactions on Software Engineering*, pages 1–26.

[Kioskli et al. 2022] Kioskli, K., Dellagiacoma, D., Fotis, T., and Mouratidis, H. (2022). The Supply Chain of a Living Lab: Modelling Security, Privacy, and Vulnerability Issues alongside with their Impact and Potential Mitigation Strategies. *JoWUA*, 13(2):147–182.

[Mosaiyebzadeh et al. 2023] Mosaiyebzadeh, F., Pouriyeh, S., Parizi, R. M., Sheng, Q. Z., Han, M., Zhao, L., Sannino, G., Ranieri, C. M., Ueyama, J., and Batista, D. M. (2023). Privacy-Enhancing Technologies in Federated Learning for the Internet of Healthcare Things: A Survey. *Electronics*, 12(12).

[Mouratidis and Giorgini 2007] Mouratidis, H. and Giorgini, P. (2007). Secure Tropos: A Security-Oriented Extension of the Tropos Methodology. *IJSEKE*, 17.

[Silva and França 2023] Silva, L. and França, R. (2023). Educação para a Cidadania Digital: Um mapeamento sobre as práticas de ensino para promover a segurança e a privacidade de dados. In *XXXI WEI*, pages 533–544. SBC.

[Wagner and Eckhoff 2018] Wagner, I. and Eckhoff, D. (2018). Technical Privacy Metrics. *ACM Computing Surveys*, 51(3):1–38.