Métricas de Privacidade: Revisão da Literatura e Aplicação em Software Científico na Área de Redes de Computadores

Estudante: Jessica Yumi Nakano Sato
Orientador: Daniel Macêdo Batista

[1]Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo (USP)

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

Resumo. *Apesar de regulamentações recentes como a Lei Geral de Proteção de Dados (LGPD) e a General Data Protection Regulation (GDPR) terem exigido que desenvolvedores de software passem a se preocupar de forma mais severa com privacidade, recomendações relacionadas a esse assunto existem há muitos anos, por exemplo desde a proposta da P3P (Platform for Privacy Preferences Project) para a Internet em 2007. Apesar do entendimento, relativamente antigo, de que sistemas computacionais precisam garantir a privacidade do usuário, tem sido difícil encontrar trabalhos que avaliem essas regulamentações em um sistema existente, principalmente porque uma métrica de privacidade pode variar a depender do domínio da aplicação (Por exemplo, a privacidade de um algoritmo de roteamento pode ser calculada como* $1/n$*, onde* n *é a quantidade de roteadores por onde um fluxo de rede passa, enquanto que a privacidade de um sistema de aprendizado de máquina para detecção de intrusão pode ser calculada como* $a/b - 1$ *onde* a *é o tamanho total do fluxo e* b *é o tamanho das informações armazenadas para treinamento). Nesse trabalho, pretende-se estudar sobre privacidade e as métricas existentes para medi-la em sistemas de software resultantes de pesquisas científicas em redes de computadores, fornecendo uma base importante a ser aplicada nos futuros resultados científicos do projeto SMARTNESS, um projeto de pesquisa financiado pela FAPESP.*

1. Introdução e Justificativa

Segurança, Privacidade e Ética são essenciais hoje nas telecomunicações devido à proliferação de ataques DDoS massivos e à necessidade de adequação às recentes regulamentações relacionadas a proteção de dados pessoais, como a LGPD no Brasil e a GDPR na Europa. No projeto SMARTNESS, no qual este trabalho se insere, considera-se que os mecanismos propostos respeitarão essas regulamentações, garantindo a privacidade dos usuários no armazenamento de dados e na inspeção tanto de cabeçalhos quanto de carga útil dos pacotes de rede. Em particular, o *Research Strand* ADVENTURE (*Adaptive and Secure Network Applications over the Industrial Internet*) do projeto SMARTNESS, liderado pelo orientador deste Trabalho de Conclusão de Curso (TCC), considera que a Internet do Futuro exigirá que plataformas e dispositivos heterogêneos, e sob controle de diversos parceiros, potencialmente não confiáveis, precisarão interoperar sem risco para as garantias de privacidade dos usuários.

Do ponto de vista de quem possui os dados, seria importante que ele(a) tivesse uma gama de opções de ferramentas que garantisse a privacidade, por exemplo antes desses dados serem enviados para uma etapa de treinamento, supondo um sistema baseado em aprendizado de máquina. 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]. Entretanto, ao se buscar por métricas de privacidade na literatura, muito se encontra a respeito de métricas do ponto de vista da percepção que o usuário tem de um sistema, ou métricas bem específicas do domínio de aplicação. Por exemplo, em um algoritmo de roteamento ad-hoc, a métrica de privacidade pode ser calculada como $1/n$, onde n é a quantidade de roteadores por onde um fluxo de rede passa. Há trabalhos que discorrem sobre métricas de um modo geral, independente do domínio de aplicação. Porém, carece-se da aplicação dessas métricas gerais para atestar o aumento da privacidade em algum sistema que se proponha a isso. Considerando o escopo do projeto SMARTNESS, atestar o nível de privacidade de software científico resultante de pesquisas em redes de computadores é algo extremamente relevante.

2. Objetivo

Neste TCC, o objetivo é estudar sobre privacidade e as métricas existentes para medi-la em sistemas de software resultantes de pesquisas científicas em redes de computadores. Com o conhecimento das métricas existentes, pretende-se aplicá-las para medir o nível de privacidade de tais sistemas, principalmente daqueles que afirmam que melhoram a privacidade de algum cenário. O estudo de métricas existentes e a busca pelos sistemas de software ideais para o projeto já encontra-se em andamento desde o fim de 2022 (os trabalhos já realizados encontram-se no `Notion` cujo link se encontra no site da estudante). No processo de estudo desses sistemas, um resultado adicional será a compreensão de qual problema relacionado com privacidade os autores estão tentando resolver.

Esta pesquisa tem potencial de trazer benefícios para a comunidade científica de redes de computadores por gerar uma visão do nível quantitativo de privacidade dos sistemas de software resultantes das pesquisas que envolvem algum tipo de análise e manipulação de fluxos na Internet. Esse benefício poderá se estender para a população em geral caso algum dos sistemas analisados durante esse projeto de pesquisa venha a ser colocado em produção em sistemas computacionais com acesso a dados sensíveis.

3. Plano de trabalho e metodologia

O desenvolvimento do trabalho envolverá três etapas bem definidas:

1. Revisão bibliográfica do tópico de métricas de privacidade;
2. Busca por sistemas de software científico resultantes de pesquisa científica em redes de computadores que tenham por objetivo aumentar a privacidade;
3. Definição das métricas de privacidade ideais para avaliar o resultado da etapa anterior e aplicação de tais métricas para atestar o aumento da privacidade em sistemas de software que se proponham a alcançar tal aumento.

As etapas 1 e 2 encontram-se em andamento desde o final do ano de 2022 (o orientador e a estudante têm realizado reuniões quinzenais regulares desde então) e têm usado como principal fonte de busca as seguintes bases de dados de artigos científicos: Google Scholar, IEEEXplore, ACM Portal e a biblioteca online SOL da Sociedade Brasileira de Computação. Material a respeito do assunto também tem sido encontrado no formato de vídeo, por isso a plataforma de compartilhamento de vídeo YouTube também será considerada como fonte de informação (nesse caso, a autoria do conteúdo em vídeo será levado em conta ao selecionar algum conteúdo

[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

como relevante). Abaixo segue um breve resumo de alguns dos trabalhos encontrados durante a execução da Etapa 1:

- *Styx: Design and Evaluation of a New Privacy Risk Communication Method for Smartphones* - `https://dx.doi.org/10.1007/978-3-642-55415-5_10`: Foi argumentado que os usuários não têm meios fáceis de saberem sobre os dados que os aplicativos de smartphones coletam e quais as informações que podem ser retiradas deles. Os autores desenvolveram uma aplicação que consegue informar os usuários de forma clara e simples os dados que estão sendo coletados, quais os possíveis riscos de privacidade que o aplicativo traz, as inferências que podem ser feitas sobre o usuário e um sumário simples de ser interpretado sobre questões de privacidade;
- *Developing a Structured Metric to Measure Privacy Risk in Privacy Impact Assessments* - `https://dx.doi.org/10.1007/978-3-319-41763-9_10`: O trabalho 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 de todo o processo de desenvolvimento de um sistema. Entretanto, a maioria das avaliações sobre privacidade resultam em longos relatórios que são complexos de analisar e comparar. Nesse sentido, propõe-se uma nova métrica quantitativa para suprir as deficiências das demais métricas existentes;
- *Technical Privacy Metrics: A Systematic Survey* - `https://doi.org/10.1145/3168389`: Neste trabalho são avaliadas as métricas disponíveis para medir PETs (*Privacy Enhancing Technologies*) e são definidos diversos conceitos relacionados, como violação de privacidade, as métricas propriamente ditas e domínios de privacidade. 80 métricas são revisadas no trabalho;
- *Privacidade e Monitoramento: Uma perspectiva LGPD e GDPR | Alessandra Monteiro* - `https://youtu.be/-yC-8lP4aFk`: São explicados alguns conceitos sobre privacidade e como o projeto de um sistema deve ser feito para garantir a privacidade dos usuários e ser concordante a lei. Assuntos como *Privacy by Design*, *Security by Design* e LGPD são tratados.

As seguintes tarefas foram planejadas para serem realizadas em 8 meses:

- T1: Continuar a leitura de trabalhos relacionados com métricas de privacidade;
- T2: Continuar a leitura de artigos científicos na área de redes de computadores que apresentem algum software que tenha por objetivo aumentar a privacidade do usuário final (O software deve estar disponível publicamente);
- T3: Selecionar ao menos dois sistemas de software encontrados na tarefa T2 para servirem de estudo de caso da aplicação de métricas de privacidade;
- T4: Selecionar as melhores métricas de privacidade a serem aplicadas nos sistemas de software selecionados na tarefa T3;
- T5: Conduzir ajustes finos nas métricas da tarefa T4;
- T6: Escrita da monografia.

A Tabela 1 apresenta a distribuição das tarefas dentro do período deste projeto.

Para o desenvolvimento dessa pesquisa, tudo o que for potencialmente produzido como tecnologia durante as tarefas T4 e T5 será baseado em e disponibilizado como software livre para facilitar a disseminação dos resultados encontrados. Serão utilizadas as boas práticas de desenvolvimento de Software Livre e de Métodos Ágeis, apoiando-se na experiência de 20

Tarefas	Maio e Junho	Julho e Agosto	Setembro e Outubro	Novembro e Dezembro
T1	X			
T2	X	X		
T3		X		
T4		X	X	
T5			X	X
T6			X	X

Tabela 1. Cronograma.

anos do Grupo de Sistemas de Software do IME-USP. O código será gerenciado por meio de repositório Git e hospedado em ambiente colaborativo. Boa parte dos resultados deste projeto serão resumos de trabalhos existentes e relatórios sobre boas práticas referentes à privacidade e às métricas de privacidade. Esses conteúdos em formato de texto serão disponibilizados na página que a estudante tem mantido sobre a pesquisa no ambiente compartilhado `Notion` desde o começo de 2023.

Serão realizadas reuniões quinzenais presenciais com o orientador, mantendo a rotina que vem sendo seguida desde o final do ano de 2022. Isso permitirá um acompanhamento constante do progresso da estudante. Além das reuniões quinzenais, o relato das atividades também será realizado semanalmente pela estudante na página do projeto no `Notion` e no site do TCC.