Um relatório do CDT | Research

Olhando de fora para dentro

Abordagens para a moderação de conteúdo em Sistemas com Criptografia de Ponta a Ponta

CENTER FOR DEMOCRACY & TECHNOLOGY
Autoria

INSTITUTO DE REFERÊNCIA EM INTERNET E SOCIEDADE
Tradução

Um relatório do CDT | Research

Olhando de fora para dentro

Abordagens para a moderação de conteúdo em Sistemas com Criptografia de Ponta a Ponta

Autoria (Em ordem alfabética de sobrenomes)

Seny Kamara
Professor Adjunto de Ciência da Computação na Brown University.

Mallory Knodel
Diretora de Tecnologia do CDT.

Emma Llansó
Diretora do Projeto Liberdade de Expressão do CDT.

Greg Nojeim
Conselheiro Jurídico Sênior e Co-Diretor de Projeto de Segurança e Vigilância do CDT.

Lucy Qin
Doutoranda em ciência da computação na Brown University.

Dhanaraj Thakur
Diretor de Pesquisa do CDT.

Caitlin Vogus
Caitlin Vogus é Subdiretora do Projeto Liberdade Expressão do CDT.

Com contribuições de
Samir Jain, DeVan Hankerson, Hannah Quay-de la Vallee, Air Goldberg e Tim Hoagland

Tradução

Paulo Renâ e Victor Vieira
Tradução, pesquisadores do IRIS

Luíza Brandão
Revisão, diretora do IRIS

Felipe Duarte
Design e diagramação, coordenador de comunicação do IRIS

Agosto de 2021

AGRADECIMENTOS

Agradecemos a Riana Pfefferkorn, Jonathan Lee e Beda Mohanty por seus comentários a uma versão anterior deste relatório. Agradecemos também aos especialistas das empresas do setor e a outros com quem conversamos, que ajudaram a informar nossa análise. Todas as visões neste relatório são do CDT.

Este trabalho foi viabilizado por meio de uma bolsa da Fundação John S. e James L. Knight

Citação sugerida: CENTER FOR DEMOCRACY & TECHNOLOGY (2022). Olhando de fora para dentro: abordagens para moderação de conteúdo em sistemas criptografados de ponta a ponta. Tradução: SANTARÉM, Paulo Renná da Silva. VIEIRA, Victor Barbieri Rodrigues. Instituto de Referência em Internet e Sociedade - IRIS-BH. XX de janeiro de 2022. https://bit.ly/3GKVY7e

O IRIS agradece ao CDT pela autorização de traduzir e divulgar, de forma gratuita, o relatório originalmente publicado em inglês.

O Center for Democracy & Technology (CDT - “Centro para Democracia e Tecnologia”, em inglês) é uma organização sem fins lucrativos isenta de tributos (nos termos da lei dos EUA 501, c, 3) que há 25 anos trabalha para promover os valores democráticos por meio da formulação de políticas e arquitetura de tecnologia. O CDT é sediado em Washington (DC) e possui um escritório europeu em Bruxelas, na Bélgica.

O Instituto de Referência em Internet e Sociedade é um centro de estudos independente fundado em Belo Horizonte, Minas Gerais. Sua missão é explorar, investigar e entender os desdobramentos da Internet sobre a sociedade contemporânea: seu desenvolvimento, suas dinâmicas, suas normas e seus padrões. Com atuação em distintas frentes, o IRIS se dedica a questões relacionadas à Internet e à sociedade, estabelecendo parcerias estratégicas na publicação de livros, estudos, policy papers, notas técnicas, podcasts e relatórios.

ÍNDICE

Introdução

Uma nova frente se abriu nas Guerras Criptográficas:[1] a moderação de conteúdo. Durante a década de 1990, os debates políticos nos EUA e na Europa sobre criptografia se concentraram nos benefícios e riscos do acesso público e estrangeiro à criptografia. As autoridades[2] em todo o mundo pressionaram por restrições ao desenvolvimento e exportação de tecnologias de criptografia, argumentando que um maior acesso público limitaria sua capacidade de monitorar as comunicações para combater o crime e proteger o público. No final, o poder público dos Estados Unidos decidiu contra tais restrições com uma mudança na política em 1999 (SWIRE & AHMAD, 2011) e outros países seguiram o exemplo.

À medida que bilhões de pessoas em todo o mundo começaram a usar serviços criptografados para proteger sua privacidade e dados ao se comunicarem com outras pessoas, as preocupações das autoridades ganharam destaque na última década. Em 2014, o então diretor do FBI argumentou que as comunicações criptografadas eram um obstáculo para a execução da lei (FEDERAL BUREAU OF INVESTIGATION, 2014). Uma declaração de 2020 dos governos dos EUA, Reino Unido, Canadá, Índia, Japão, Austrália e Nova Zelândia expressou preocupações semelhantes, pedindo maior acesso por parte das autoridades policiais às comunicações criptografadas (U.S. DEPARTMENT OF JUSTICE, 2020).

A disponibilidade de serviços de comunicação criptografados seguros é fundamental para a privacidade, a liberdade de expressão e a segurança do comércio online atual.

Declarações como essas tendem a enfocar a política de criptografia nas alegações das autoridades de que elas precisam ser capazes de acessar comunicações criptografadas (NATIONAL ACADEMIES OF SCIENCES, ENGINEERING AND MEDICINE, 2018). Mas a criptografia não é apenas uma questão de execução da lei. A disponibilidade de serviços de comunicação criptografados seguros é fundamental para a privacidade, a liberdade de expressão e a segurança do comércio online atual (THOMPSON & PARK, 2020).

Talvez reconhecendo a difícil batalha que eles enfrentam para minar uma parte tão crucial de nossa infraestrutura online, algumas autoridades oficiais[3] começaram a vincular a ameaça de conteúdo ilegal irrestrito online a preocupações sobre as práticas de moderação de conteúdo de grandes plataformas de

1 Ver SWIRE & AHMAD (2011) para uma descrição e história dos debates sobre políticas públicas que caracterizaram as Guerras Criptográficas.

2 Nota da Tradução 1: optou-se pela palavra em português "Autoridade" onde o texto original usa a expressão "*Law Enforcement and Intelligence Agencies*".

3 NdT. 2: "autoridades oficiais" se refere a "*law enforcement officials*" no original.

mídia social. Nos EUA, por exemplo, a proposta *EARN IT Act*[4] foi estruturada como um projeto de lei que estabeleceria as melhores práticas de moderação de conteúdo para o combate a material de abuso sexual de crianças (*CSAM*), mas o debate rapidamente passou a se concentrar nas implicações do projeto de lei para a criptografia de ponta a ponta (*E2EE*, iniciais de "*end-to-end encryption*"), com muitos comentaristas expressando a preocupação de que a abordagem do projeto foi desenhada para desencorajar os provedores a oferecerem serviços com *E2EE* ou para criar fortes incentivos a embutirem um mecanismo de acesso especial para as autoridades (MURDOCK, 2020; NEWMAN, 2020; RUANE, 2020).

A definição legal de material de abuso sexual de crianças (CSAM, iniciais de "child sexual abuse material", em inglês) varia de acordo coam a jurisdição e geralmente se refere à retratação ou representação de crianças envolvidas em atividade ou abuso sexual (INTERNATIONAL CENTRE FOR MISSING & EXPLOITED CHILDREN (ICMEC), 2018).

Mas qual é o efeito real da criptografia na moderação de conteúdo?

Neste artigo, avaliamos as propostas técnicas existentes para moderação de conteúdo em serviços com criptografia de ponta a ponta. Primeiro, explicamos as várias ferramentas disponíveis para a moderação de conteúdo, como são usadas e as diferentes fases do ciclo de moderação, incluindo a detecção de conteúdo indesejado. Em seguida, apresentamos uma definição de criptografia e criptografia de ponta a ponta, a qual inclui garantias de privacidade e segurança para os usuários finais, antes de avaliarmos as atuais propostas técnicas para a detecção de conteúdo indesejado em serviços com criptografia de ponta a ponta, que se contrapõem a essas garantias.

Descobrimos que as abordagens técnicas para denúncia por usuários e análises de metadados são as que aparentemente mais preservam as garantias de privacidade e segurança para os usuários finais. Ambas fornecem ferramentas eficazes que podem detectar quantidades significativas de diferentes tipos de conteúdo problemático em serviços com criptografia de ponta a ponta, incluindo mensagens abusivas e de assédio, *spam* (mensagem não solicitada), informação errada[5], desinformação e material de abuso sexual de crianças, embora mais pesquisas sejam necessárias para melhorar essas ferramentas e medir melhor sua eficácia. Por outro lado, descobrimos que outras técnicas que pretendem facilitar a detecção de conteúdo em sistemas com criptografia de ponta a ponta têm o efeito de minar as principais garantias de segurança dos sistemas.

Todas as propostas técnicas atuais que revisamos se concentram na detecção de conteúdo, que é apenas uma parte do processo de moderação de conteúdo. Assim, pode haver

4 NdT. 3: Do inglês "*Eliminating Abusive and Rampant Neglect of Interactive Technologies*". Tradução livre: "Lei de Eliminação da Negligência Desenfreada e Abusiva de Tecnologias Interativas"

5 NdT. 4: "informação errada" aqui se refere ao original "*misinformation*", em distinção ao termo "desinformação", correspondente a "*disinformation*".

outras abordagens úteis e eficazes de moderação para combater o abuso em sistemas com criptografia de ponta a ponta, incluindo educação do usuário sobre as políticas aplicáveis, *design* aprimorado para encorajar denúncias por usuário e consistência das decisões de moderação. Essas abordagens podem oferecer importantes caminhos potenciais para os pesquisadores avançarem a partir de nossa análise.

Compreendendo a moderação de conteúdo

Moderação de Conteúdo se refere ao conjunto de políticas, sistemas e ferramentas que os intermediários de conteúdo gerado pelo usuário usam para decidir qual conteúdo gerado pelo usuário ou contas publicar, remover ou gerenciar de alguma forma (BLOCH-WEHBA, 2020; ver também GRIMMELMANN, 2015; KLONICK, 2018). Neste artigo, nos concentramos principalmente nas decisões de moderação por provedores de hospedagem, embora observemos a longa pressão sobre os mecanismos de pesquisa, provedores de mensagens, provedores de nomes de domínio, provedores de acesso e outros intermediários técnicos para que se envolvam na moderação.

Os provedores de hospedagem de conteúdo[6] podem moderar tanto o conteúdo que é ilegal quanto o conteúdo que, embora legal, viola seus termos de serviço ou outras regras. As estruturas de responsabilidade geralmente distinguem entre os sistemas que um provedor de hospedagem possui para responder a conteúdo ilegal e aqueles implementados para lidar com conteúdo que viola seus próprios termos de serviço. No entanto, na prática, os provedores de hospedagem removem quantidades substanciais de conteúdo supostamente ilegal como violação de seus termos de serviço (KLONICK, 2018). Este documento examina os processos que os provedores de hospedagem podem usar para agir contra o conteúdo gerado pelo usuário ou suas contas, independentemente do motivo.

Os provedores de hospedagem adotam uma variedade de abordagens para moderação de conteúdo. Alguns usam sistemas automatizados para filtrar o conteúdo gerado pelo usuário no *upload*, por exemplo, para detectar possível violação de direitos autorais ou material de abuso sexual de crianças antes da publicação; outros principalmente revisam e moderam o conteúdo depois da postagem. Alguns agem de forma reativa,

A distinção entre conteúdo que é ilegal e conteúdo que viola os termos de serviço de um provedor de hospedagem é importante; os regimes legais que exigem a remoção de conteúdo ilegal devem garantir que os tribunais, e não os intermediários, sejam responsáveis por determinar que o conteúdo é ilegal antes que ele seja removido. Consulte Princípios de Manila sobre responsabilidade de intermediário (última visita em 30 de março de 2021), https://manilaprinciples.org/pt-br.html.

6 Ndt. 5: Considerando o instituto jurídico "provedor de aplicações" constante do Marco Civil da Internet (Lei nº 12.965/2014), optou-se por privilegiar o termo "provedor de hospedagem", como correspondente aos originais "*host*", literalmente "hospedeiro"; e "*hosting intermediaries*", "intermediários de hospedagem".

revisando e moderando o conteúdo somente depois de ele ser denunciado como condenável; outros procuram conteúdo de forma proativa para moderar (KLONICK, 2018). Alguns contam com revisão manual por humanos para moderar o conteúdo, enquanto outros contam com processos automatizados.[7] Os provedores de hospedagem podem usar uma combinação de revisão humana e automatizada de maneiras reativas e proativas (BLOCH-WEHBA, 2020). Mas muitos serviços online, especialmente serviços menores, continuam a depender de uma revisão reativa posterior à publicação do conteúdo que é denunciado à provedora do serviço por um usuário ou outro terceiro.

Além disso, diferentes provedores de hospedagem exercem diferentes níveis de controle sobre a moderação de conteúdo devido às especificidades do design de seu site, modelo de negócios, capacidade de incorporar o contexto local em suas avaliações e outras considerações (CAPLAN, 2018). Muita atenção tem sido dada aos provedores de hospedagem – como Facebook, Twitter e YouTube – que estão diretamente envolvidos na moderação de conteúdo e tomam decisões de moderação de forma centralizada. Esses provedores de hospedagem podem escrever políticas e regras externas e internas sobre o conteúdo permitido em seus sites, usar funcionários ou contratados terceirizados para revisar o conteúdo e tomar decisões de moderação de conteúdo, e empregar equipes para analisarem os recursos do usuário contra as decisões de moderação de conteúdo (GILLESPIE, 2012; KLONICK, 2018).

No entanto, outros provedores de hospedagem contam com a comunidade ou moderação distribuída, na qual os próprios usuários moderam o conteúdo com pouco ou nenhum envolvimento do provedor de hospedagem. Por exemplo, Reddit, Wikipedia, Slashdot e Discord definem políticas básicas para conteúdo, enquanto contam com voluntários para definir regras adicionais, tomar decisões de moderação de conteúdo ou ambos (CAPLAN, 2018; GRIMMELMANN, 2015; LAMPE & RESNICK, 2004; SWARTZ, 2006). Alguns provedores de hospedagem combinam os dois métodos, empregando a tomada de decisão central para certas decisões de moderação de conteúdo e moderação da comunidade para outras. Por exemplo, o serviço de transmissão de vídeo Twitch descreve sua abordagem de moderação como “uma abordagem em camadas para a segurança - uma que combina os esforços do Twitch (por meio de ferramentas e pessoal) e de membros da comunidade, trabalhando juntos” (TWITCH, 2021).

7 Para um exame aprofundado das várias técnicas que podem ser usadas para analisar conteúdo gerado pelo usuário, consulte SHENKMAN *et al.* (2021).

A moderação de conteúdo é um processo muito mais complexo do que simplesmente tomar decisões binárias para remover ou permitir conteúdo gerado pelo usuário em um serviço.

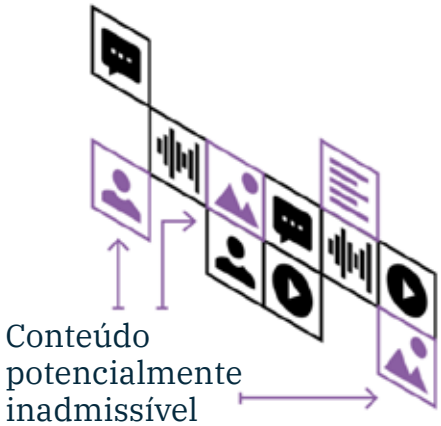
Fases da moderação de conteúdo

É útil pensar na moderação de conteúdo como ocorrendo em seis fases: definição, detecção, avaliação, intervenção, recurso e educação. Além disso, a moderação de conteúdo é um processo iterativo; essas fases estão inter-relacionadas e cada fase pode acontecer várias vezes e em uma ordem diferente da descrita abaixo.

A. Definição

Na fase de **definição**, os provedores de hospedagem ou outros determinam qual conteúdo gerado pelo usuário é ou não permitido no serviço. Isso envolve a definição de conteúdo e comportamento inadmissíveis e a descrição de tal conteúdo e comportamento para outras pessoas, tanto externamente aos usuários quanto internamente. Os provedores de hospedagem podem definir e comunicar o conteúdo permitido e não permitido em seus termos de serviço ou diretrizes da comunidade, mas as regras também podem ser definidas e comunicadas de outras maneiras. Por exemplo, as regras dos fóruns *subreddit* no Reddit são comumente exibidas para usuários no próprio *subreddit*, com frases curtas para identificar o tópico da regra e uma breve explicação (FIESLER *et al.*, 2018).

B. Detecção



A **detecção** é como os provedores de hospedagem ou outros moderadores identificam o conteúdo gerado pelo usuário que pode violar suas políticas ou a legislação (para obter mais informações sobre detecção, consulte "Detecção de conteúdo em ambientes com criptografia ponta a ponta," p. 16). Os provedores de hospedagem se envolvem em uma variedade de métodos a fim de detectar conteúdo para moderação e podem usar vários métodos simultaneamente. Conforme descrito acima, a detecção pode ocorrer em diferentes momentos – seja "antes que o conteúdo seja realmente publicado no site, em uma moderação *ex ante*, ou depois que o conteúdo ter sido publicado, em uma moderação *ex post*" (KLONICK, 2018, p. 1635). A detecção *ex post* pode ser reativa, na qual os moderadores contam com os usuários ou terceiros para "chamar sua atenção para o conteúdo", como por meio de sinalização; ou proativa, "na qual as equipes de moderadores procuram ativamente o conteúdo publicado para remoção" (KLONICK, 2018, p. 1635). Em seus esforços de detecção, os provedores de hospedagem podem confiar no conteúdo que os usuários sobem, bem como nos metadados associados a esse conteúdo, como informações da conta, endereço IP, volume/frequência de postagem e outros sinais.

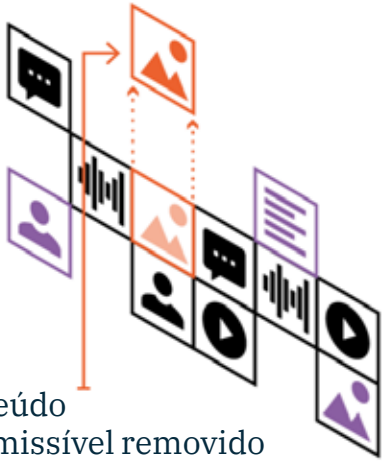
Os esforços para detectar "comportamentos inautênticos coordenados" em serviços de mídias sociais, por exemplo, basearam-se principalmente no uso de metadados (FRANÇOIS, 2020).

C. Avaliação

Avaliação de conteúdo potencialmente inadmissível

Durante a fase de **avaliação**, o conteúdo gerado pelo usuário é examinado para determinar se ele viola as políticas do provedor de hospedagem ou é potencialmente uma violação de uma legislação relevante. A avaliação pode ser feita por humanos, automaticamente, ou por meio de uma combinação de revisão automatizada e humana. Por exemplo, uma imagem que foi relatada por um usuário pode ser examinada por um moderador humano ou analisada por um programa de correspondência de *hash* (consulte a página 22 abaixo) para determinar se a imagem corresponde ao conteúdo que o provedor já sabe que deseja bloquear. Quando os provedores de hospedagem empregam a filtragem de conteúdo automatizada para bloquear o conteúdo no *upload*, as etapas separadas de detecção, avaliação e intervenção podem se resumir em uma única etapa.

D. Intervenção



Conteúdo inadmissível removido

Intervenção é a ação que um moderador realiza contra o conteúdo gerado pelo usuário que ele determina que viola uma política de conteúdo ou lei. Embora a remoção de conteúdo seja um método de intervenção possível, há uma grande variedade de outras ações que os moderadores podem tomar contra o conteúdo violador (Goldman, 2021). Por exemplo, os moderadores podem: adicionar um aviso antes que os usuários possam acessar o conteúdo ou um discurso contraposto, como uma checagem de fatos; desabilitar comentários do usuário ou outros recursos em uma postagem; diminuir a disponibilidade de algumas ou todas as postagens de um usuário, como por meio de banimentos velados[8], rebaixando a visibilidade do conteúdo nos resultados de pesquisa ou restringindo o encaminhamento ou compartilhamento de postagens; impor remédios monetários, como desmonetizar conteúdo; ou suspender ou desativar a conta de um usuário (GOLDMAN, 2021; MASNICK, 2018).

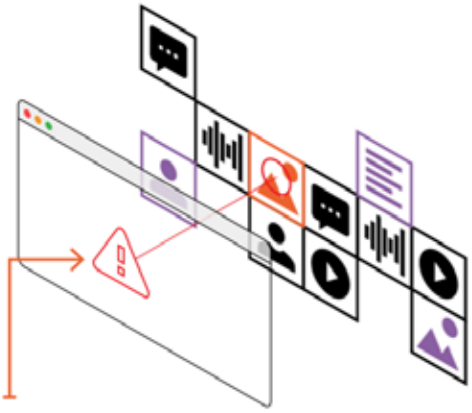
8 NdT. 6: "banimentos velados" se refere ao original "*shadowban*".

E. Recurso

Conteúdo Inadmissível considerado admissível e substituído

Após a intervenção, alguns provedores de hospedagem permitem aos usuários o **recurso** ou outra forma de buscar a revisão das decisões de moderação de conteúdo que os usuários acreditam estarem erradas (KLONICK, 2020). Os erros são inevitáveis na moderação de conteúdo e, dada a quantidade de conteúdo que alguns provedores de hospedagem moderam, mesmo aqueles com uma taxa muito alta de precisão na moderação de conteúdo ainda tomarão milhares ou milhões de decisões erradas[9]. Consequentemente, os recursos são uma parte essencial da moderação de conteúdo. Além de permitir recursos do usuário, as políticas de moderação de conteúdo devem reconhecer a inevitabilidade de erros e criar oportunidades para revisar e refinar as próprias políticas de moderação de conteúdo e as ferramentas usadas para implementá-las.

F. Educação



Usuário informado sobre conteúdo admissível e inadmissível

Por fim, os provedores de hospedagem podem **educar** os usuários sobre suas políticas de moderação de conteúdo e as maneiras como as políticas são aplicadas. A fase de educação pode assumir várias formas. Os mais básicos são os termos de serviço da aplicação, as Diretrizes da comunidade e outras informações ao usuário sobre as políticas do site. Os moderadores também podem educar os usuários sobre o conteúdo permissível e inadmissível, "elogiando o bom comportamento e criticando o mau ou de outra forma fornecendo explicações aos usuários quando seu conteúdo é moderado" (GRIMMELMANN, 2015, pp. 61–63). Isso pode incluir as notificações enviadas aos usuários quando são informados de que uma ação foi tomada contra seu conteúdo ou quando um recurso apresentado por eles foi negado. A educação é um componente crítico da moderação de conteúdo, especialmente em resposta a violações de boa fé das políticas de moderação de conteúdo. Os usuários devem entender quais tipos de conteúdo são e quais não são permitidos para que o processo de moderação de conteúdo funcione de forma eficaz (JHAVER *et al.*, 2019).

A discussão anterior sobre as fases inter-relacionadas de moderação de conteúdo são relevantes para todos os tipos de configurações e contaram com exemplos de ambientes de texto simples (ou não

9 Por exemplo, Klonick (2020) explica que no segundo e terceiro trimestres de 2019, menos de 1% das 4,8 bilhões de conteúdos que o Facebook removeu do site foram recorridas, resultando na restauração de 10,1 milhões conteúdos (pág. 2433-34).

criptografados) em particular. Como observamos anteriormente, nosso objetivo aqui é entender as implicações da criptografia nessas fases. Para fazer isso, a próxima seção explica o que queremos dizer com criptografia antes de passar para nossa análise de propostas que buscam permitir algumas formas de moderação de conteúdo em ambientes com criptografia de ponta a ponta.

Compreendendo a criptografia de ponta a ponta

A criptografia é crucial para a proteção da privacidade e da segurança do usuário final, bem como a promoção da liberdade de expressão online (KAYE, 2015). No entanto, os desafios permanecem para garantir que os usuários finais entendam exatamente o que é criptografia, incluindo especificamente criptografia ponta a ponta (Bai et al., 2020). Os formuladores de políticas também encontram muitos mitos sobre a criptografia ponta a ponta (Global Encryption Coalition, 2020).

Começamos nossa análise definindo primeiro a criptografia com referência às suas principais características de encriptação: assinatura e cifragem (Uhlig et al., 2021). Um esquema de cifragem toma como entrada uma chave e uma mensagem e produz um texto cifrado (uma versão criptografada da mensagem). Os esquemas de cifragem variam e podem oferecer propriedades diferentes. Por exemplo, um esquema de cifragem pode usar uma chave para empacotar os dados de uma forma que garanta autenticação (saber quem os enviou), confidencialidade (apenas o destinatário pode abrí-los) e integridade (a mensagem não foi adulterada). Dada a chave, o texto cifrado pode ser decifrado para revelar a mensagem. A criptografia raramente é usada isolada e geralmente é uma parte de um sistema maior, como um aplicativo de mensagens ou serviço de armazenamento em nuvem. Em tais sistemas, a criptografia é usada para garantir a autenticidade, confidencialidade e integridade dos dados.

Como exatamente a criptografia é implementada no sistema depende de quem obtém acesso aos dados e como. Como a criptografia só pode garantir a confidencialidade entre as partes que compartilham uma chave criptográfica, uma das considerações mais importantes ao implantar a criptografia é como gerenciar o compartilhamento de chaves. Essa troca compartilhada, juntamente com a confidencialidade da criptografia, permite que os projetistas do sistema protejam os dados em trânsito, enquanto estão em repouso e entre os terminais.

Criptografia em trânsito significa que a cifragem é implantada

para proteger os dados enquanto estão sendo comunicados por uma rede pelo provedor de serviços. Mais precisamente, significa que o transporte de dados é cifrado entre nós autenticados pelo canal de comunicação e oculto de todos os outros nós que podem estar retransmitindo os dados. Esse é o caso, por exemplo, quando um navegador da web criptografa o tráfego da web de um usuário antes de enviá-lo a um servidor da web. Aqui, as chaves de decifragem são conhecidas apenas pelo servidor da web.

Um sistema, serviço ou aplicativo conta com criptografia de ponta a ponta se as chaves usadas para cifrar e decifrar os dados são conhecidas apenas pelo remetente e pelos destinatários autorizados desses dados.

A criptografia em repouso se refere ao uso de criptografia simétrica para proteger os dados enquanto eles são armazenados. Mais especificamente, significa que um usuário emprega uma chave que ele pode ou não compartilhar com outras pessoas, de modo que os dados sejam decifrados quando desejam usá-los e cifrados quando não estão em uso. Como exemplo, considere um serviço de armazenamento em nuvem que faz backup e armazena dados do usuário. Nesse serviço, os dados são enviados do dispositivo do usuário para o provedor de nuvem – geralmente com criptografia em trânsito – e, então, criptografados pelo provedor para serem armazenados. Se os centros de dados do provedor forem violados (virtualmente ou fisicamente), o acesso aos dados de armazenamento seria negado sem a chave criptográfica simétrica que apenas o usuário possui. Em tal implantação, outro usuário ou o provedor de nuvem poderia ser autorizado a obter acesso aos dados ao receber o conhecimento da chave, mas o ideal é que não seja. Embora a criptografia em trânsito e a criptografia em repouso sejam essenciais para a construção de um sistema seguro, elas protegem os dados nesses contextos específicos. Para proteger os dados que são trocados entre dois ou mais usuários que não são provedores de serviços intermediários, e de forma que os provedores de serviços intermediários não possam testemunhar a conversa, outra arquitetura é necessária.

Essa arquitetura para implementar criptografia que protege os dados do usuário em todos os momentos e contra qualquer invasor no caminho, é a criptografia ponta a ponta (E2EE, do inglês end-to-end encryption). Aqui, os dados são cifrados no dispositivo do usuário e só podem ser decifrados por usuários autorizados que trocaram chaves entre si. Como esses usuários são os únicos com conhecimento das chaves de decifragem, os dados em sistemas com criptografia de ponta a ponta são confidenciais para esses usuários e ninguém mais, nem mesmo o intermediário provedor de serviço (KNODEL et al., 2021).

Em resumo, um sistema, serviço ou aplicativo conta com criptografia de ponta a ponta se as chaves usadas para cifrar e decifrar dados são conhecidas apenas pelo remetente e pelos destinatários autorizados desses dados. Especificamente, isso implica que as partes intermediárias que roteiam, armazenam, fazem backup e processam os dados criptografados não têm

acesso às chaves e, portanto, não podem aprender nenhuma informação sobre os dados.

Exemplos de serviços que contêm criptografia de ponta a ponta

Na prática, a criptografia de ponta a ponta é usada de várias maneiras e todas elas atraíram a preocupação das autoridades e outros, em termos de detecção de conteúdo. Abaixo estão as descrições de alguns serviços comuns e as garantias de privacidade que a criptografia de ponta a ponta deve lhes fornecer:

Armazenamento. Isto é um serviço de armazenamento em nuvem que armazena arquivos ou fotos criptografados de ponta a ponta. Ou seja, os dados são primeiro criptografados em uma chave conhecida apenas pelo usuário e, em seguida, armazenados na nuvem. Os exemplos incluem serviços como o sistema de arquivos Keybase e o aplicativo de fotos Pixek. Nesta configuração, a E2EE é usada para garantir que o usuário seja a única pessoa que pode acessar os dados. Outros exemplos, como o Dropbox, criptografam arquivos e fotos do usuário em repouso, mas o serviço mantém o acesso à chave.

Mensagens. Uma troca de mensagens criptografada é uma conversa entre duas ou mais pessoas em um aplicativo de mensagens com criptografia de ponta a ponta. Aqui, as mensagens são criptografadas usando chaves conhecidas apenas pelos participantes da conversa. Isso inclui aplicativos de mensagens como WhatsApp e Signal. Aqui, a E2EE é usada para tornar a conversa confidencial, no sentido de que apenas os participantes autenticados na conversa podem acessar as mensagens.

Email. Isto é um serviço de e-mail, ou ferramenta, que permite aos usuários enviar e receber e-mails com criptografia de ponta a ponta. Como nas mensagens criptografadas, as chaves são conhecidas apenas pelo remetente e destinatários, o que garante que nenhum terceiro – incluindo o provedor do serviço de e-mail – possa ver o conteúdo do e-mail.

Detecção de conteúdo em ambientes com criptografia ponta a ponta

Com a expansão da gama de serviços com criptografia de ponta a ponta disponíveis, as preocupações das autoridades sobre "ficarem no escuro" voltaram ao primeiro plano dos debates políticos em todo o mundo.[10] Um novo ângulo dessas deliberações é o foco na moderação de conteúdo, levantando questões sobre se e como os provedores de serviços com *E2EE* podem ou devem identificar proativamente conteúdo problemático, responder a usos abusivos de seus sistemas e implementar ordens legais para bloquear conteúdo.[11]

À medida que os legisladores se envolvem nesses debates importantes sobre como, por exemplo, interromper a disseminação de material de abuso sexual de crianças ou abordar o abuso de serviços com *E2EE* por organizações terroristas, é importante que eles entendam que as questões de como melhor lidar com o conteúdo problemático em ambientes com criptografia de ponta a ponta geralmente são específicos para como a criptografia é implementada. Também é essencial centrar as garantias de privacidade e segurança que desenvolvedores de sistema e usuários esperam da criptografia ponta a ponta, ou seja, que apenas o remetente e os destinatários autorizados tenham acesso às chaves de cifragem

10 Em um ambiente de texto simples, as autoridades ou outros órgãos estatais frequentemente obtêm acesso ao conteúdo das comunicações privadas para investigar crimes e proteger a segurança nacional, com ou sem a cooperação do provedor de serviços. Com a criptografia de ponta a ponta, isso normalmente não é possível e alguns países pediram medidas que dariam às autoridades um "acesso excepcional" às comunicações com *E2EE* (NATIONAL ACADEMIES OF SCIENCES, ENGINEERING AND MEDICINE, 2018). Essas propostas de "acesso excepcional" dariam às autoridades o acesso às chaves usadas para decifrar os dados usando um sistema de custódia de chaves, ou permitiriam "portas clandestinas" de criptografia modificando deliberadamente o esquema de criptografia para permitir o acesso de terceiros (por exemplo, pelo provedor de serviços em resposta a um processo legal). Mais especificamente, onde os prestadores de serviços permitem tais formas de acesso excepcional, isso significa que, por definição, não estamos mais a falar de um sistema com criptografia de ponta a ponta. Isso também significa que outros atores (e potenciais adversários), além das autoridades, podem obter acesso ao conteúdo (ABELSON *et al.*, 2015). Em geral, a criação de portas clandestinas criará intencionalmente vulnerabilidades em um sistema, aumentando os riscos para todos os usuários (GLOBAL ENCRYPTION COALITION, 2020).

11 Ndt. 7: A fim de destacar o caráter de excepcionalidade, optou-se por "portas clandestinas" em referência ao original "*backdoors*", em da tradução literal "porta de trás" ou mesmo do comum "porta dos fundos". Para saber mais sobre a proposta de tradução, ver: RENÁ, Paulo. Blog "Portas clandestinas": uma tradução mais precisa para debatermos backdoors em criptografia. Disponível em: <https://irisbh.com.br/portas-clandestinas-uma-traducao-mais-precisa-para-debatermos-backdoors-em-criptografia/>. Acesso em: 31 de jan. de 2022.

e decifragem e aos dados, e, portanto, os intermediários não tenham. Com isso em mente, revisamos uma série de propostas técnicas emergentes de pesquisas em ciência da computação e criptografia que buscam possibilitar a detecção de conteúdo em serviços com *E2EE*.[12]

Uma observação inicial sobre essas propostas é que elas se concentram na detecção de conteúdo: como um provedor de serviços pode identificar que algum subconjunto de dados criptografados é um conteúdo problemático?

Conforme discutido anteriormente, a detecção é apenas uma fase da moderação de conteúdo. Além disso, os tipos de conteúdo de interesse – normalmente conteúdo "prejudicial", ilegal ou indesejado, tal como propaganda terrorista, material de abuso sexual de crianças, informação errada e desinformação ou spam – não têm características técnicas exclusivas que os tornem facilmente distinguíveis de tipos mais inócuos de conteúdo (ou seja, uma imagem é uma imagem independentemente de seu conteúdo). Assim, o que muitas vezes é identificado como um debate sobre a moderação de conteúdo indesejado em serviços com criptografia de ponta a ponta é, na verdade, uma discussão sobre a detecção de (qualquer) conteúdo sob criptografia de ponta a ponta.

Para a maioria dos tipos de conteúdo abusivo, ilegal ou de outra forma "prejudicial", também pode haver divergência significativa entre países e comunidades sobre como definir tal conteúdo proibido.

O que muitas vezes é identificado como um debate sobre moderação de conteúdo indesejado em serviços com criptografia de ponta a ponta é na verdade uma discussão sobre a detecção de (qualquer) conteúdo sob criptografia de ponta a ponta.

Abaixo, examinamos cinco tipos de técnicas usadas tanto em sistemas *E2EE* e sistemas de texto simples ou sem criptografia, que geralmente tentam detectar um conteúdo já carregado ou adicionado ao sistema e/ou tentam impedir que um conteúdo indesejado seja adicionado ao sistema. Elas são: denúncia por usuário, rastreabilidade, análise de metadados, verificação[13] perceptiva e modelos preditivos.

Denúncia por usuário

Em um ambiente de texto simples, já existem várias opções para detectar conteúdo indesejado que já foi carregado ou postado no serviço de hospedagem. Uma das abordagens mais comuns é permitir alguma forma de denúncia iniciada por usuário. Um provedor de serviços pode disponibilizar ferramentas (por exemplo, botões de denúncia, formulários de reclamação, informações de contato) que permitem aos usuários alertar moderadores, outros intermediários ou outros usuários sobre conteúdos indesejados. Os moderadores podem visualizar o conteúdo diretamente e tomar medidas ou encaminhá-lo para uma análise posterior.

12 Consulte o Apêndice para obter mais detalhes técnicos sobre algumas propostas.

13 NdT. 8: Optou-se pelo termo "verificação" em correspondência ao termo original "*hashing*", mantendo-se no entanto o uso da palavra "*hash*".

As autoridades ou terceiros podem estabelecer canais de denúncia dedicados com os provedores de serviços, a fim de notificá-los sobre conteúdo potencialmente ilegal.[14] E os provedores podem tomar conhecimento de conteúdo indesejado em seus serviços por meio de e-mails, reportagens noticiosas ou outras comunicações que ocorram fora das ferramentas ou procedimentos de sinalização específicos que tenham desenvolvido.[15]

Usando essas opções como ponto de partida, pesquisadores e outros apresentaram várias propostas para permitir denúncias por usuários em serviços com *E2EE*. Atualmente, há uma variedade de esquemas criptográficos propostos para permitir a geração de denúncias (por usuários, para provedores de serviços) de mensagens com criptografia de ponta a ponta. Essas soluções são projetadas para que as mensagens só possam ser descriptografadas e verificadas pelo provedor de serviços e ninguém mais além dos originais remetente e destinatários. Esta categoria de esquemas é chamada de "franqueamento de mensagem". Dada uma conversa privada entre os usuários A e B, o franqueamento de mensagens garante que:

1. B pode provar ao provedor de serviços que recebeu uma determinada mensagem de A; e

2. B não pode reclamar ao provedor de serviços que recebeu uma mensagem de A que nunca recebeu.

A franquia de mensagens já está em uso em sistemas com E2EE. Por exemplo, o Facebook emprega franqueamento de mensagens em seu sistema de mensagens com criptografia de ponta a ponta, Secret Conversations (FACEBOOK, 2016). Depois que o Facebook introduziu o franqueamento de mensagens, o esforço que se seguiu melhorou o esquema original da empresa, tornando-o mais eficiente para arquivos anexos (DODIS et al., 2018; GRUBBS et al., 2017), permitindo apenas a abertura parcial de mensagens (apenas partes específicas de um mensagem são reveladas) (CHEN & TANG, 2018; LEONTIADIS & VAUDENAY, 2018) e estendendo o franqueamento de mensagem para provedores de serviços privados de metadados, ou seja, provedores de serviços que não revelam quem são o remetente e os destinatários das mensagens (TYAGI, GRUBBS, et al., 2019).

Embora o franqueamento de mensagens permita que o provedor de serviços visualize as mensagens, isso não viola as propriedades que esperamos de uma conversa criptografada,

14 Observe que isso é controverso.

15 Observe que esta é uma questão distinta de se um serviço tem conhecimento real de que uma postagem específica seja ilegal.

uma vez que um dos participantes originais da conversa (aqui, o receptor) escolhe explicitamente revelar a mensagem ao provedor de serviços. De uma perspectiva mais técnica, as chaves usadas para criptografar e descriptografar ainda são mantidas apenas pelo remetente e pelo receptor e não há portas clandestinas que permitam ao provedor de hospedagem ou qualquer outro terceiro acessar a conversa sem o conhecimento e aprovação de pelo menos de um dos participantes. Tudo o que o franqueamento de mensagens permite é que uma das partes em uma comunicação a revele ao provedor de serviços de forma que ele possa ter certeza da autenticidade da mensagem.

O franqueamento de mensagens vincula explicitamente os remetentes às suas mensagens para que possam ser responsabilizados se enviarem conteúdo indesejado ou ilegal, como mensagens de ódio e assédio, material de abuso sexual de crianças, propaganda terrorista ou spam.

Com o franqueamento de mensagens, uma conversa privada não pode mais ser repudiada por alguém que fez parte dessa conversa (ou seja, o moderador designado pode verificar uma mensagem denunciada). Isso é útil quando for importante o não-repúdio, ou seja, a necessidade de garantir que o autor de uma mensagem não possa negar sua autoria. Na verdade, o franqueamento de mensagens vincula explicitamente os remetentes às suas mensagens para que possam ser responsabilizados se enviarem conteúdo indesejado ou ilegal, como mensagens de ódio e assédio, material de abuso sexual de crianças, propaganda terrorista ou *spam*. Essa técnica também pode permitir que a pessoa que está denunciando a mensagem se proteja de qualquer responsabilidade caso receba conteúdo abusivo ou ilegal; por exemplo, o franqueamento de mensagem permite que a pessoa que está denunciando a mensagem comprove se ela é a destinatária, e não a origem, de uma mensagem contendo *CSAM*.

Embora as técnicas de franqueamento de mensagem descritas acima não violem a garantia de ponta a ponta de uma conversa privada, as variantes mais recentes poderiam e, portanto, é importante que as implementações práticas do franqueamento de mensagem sejam transparentes sobre as propriedades exatas que elas garantem.

Rastreabilidade

Uma preocupação relacionada é como identificar quais usuários compartilharam conteúdo que foi sinalizado como problemático. Por exemplo, o provedor de serviços ou uma agência governamental pode querer rastrear como um determinado conteúdo foi distribuído online. Como a criptografia de ponta a ponta padrão torna mais difíceis a detecção e o rastreamento, os pesquisadores estão estudando até que ponto o rastreamento pode ser feito em plataformas de mensagens com criptografia de ponta a ponta.

Uma proposta de Tyagi, Miers, *et al.*, (2019), estende as técnicas de franqueamento de mensagens para rastrear todos os usuários que encaminharam ou receberam determinado

conteúdo. Antes de alguém sinalizar o conteúdo, as mensagens são mantidas em sigilo e apenas o remetente e o(s) destinatário(s) de uma mensagem podem descriptografá-la. Depois que o conteúdo é denunciado, no entanto, o provedor de serviços pode aprender o conteúdo de uma conversa e rastreá-la para localizar todas as mensagens com o mesmo conteúdo que não foram denunciadas diretamente na cadeia de encaminhamento. Embora isso permita que um provedor de serviços rastreie a disseminação de conteúdo malicioso viral, também oferece uma oportunidade para que os usuários denunciem conteúdo confidencial e exponham a privacidade de todos os remetentes e destinatários na cadeia.

Essas técnicas de rastreamento são construídas sobre o franqueamento de mensagem e, embora o franqueamento não viole as propriedades que esperamos de conversas criptografadas, o rastreamento sim. Lembre-se que, no uso básico do franqueamento de mensagens, a mensagem só pode ser revelada ao provedor de serviço se um dos participantes da conversa decidir denunciá-la, preservando, assim, a garantia de privacidade da criptografia de ponta a ponta de que somente o remetente e destinatários autorizados têm acesso os dados.

Com o rastreamento, no entanto, o provedor de serviços pode obter informações que não foram explicitamente reveladas a ele pelo remetente ou destinatário. Por exemplo, se o usuário A enviar uma mensagem para o usuário B, que, então, enviará a mesma mensagem para o usuário C, o usuário C pode relatar a mensagem e o esquema de rastreamento revelará não apenas que B enviou a mensagem para C, mas que ela foi enviada por A para B sem A ou B explicitamente revelarem isso ao provedor de serviços. Em uma cadeia de encaminhamento mais expandida, um indivíduo que pode estar a 1000 conexões de distância do remetente original pode denunciar uma mensagem que, então, revelaria que ela foi recebida pelos 999 destinatários anteriores.

A demanda por rastreabilidade entre os governos é frequentemente baseada em propostas politicamente motivadas com pouca ou nenhuma orientação técnica em termos de viabilidade. No ano passado, o governo brasileiro introduziu a chamada “Lei das *Fake News*” exigindo uma forma de rastreabilidade e identificação do usuário que não apenas quebraria a criptografia e prejudicaria a privacidade e a liberdade de expressão, mas também representaria um fardo significativo para os prestadores de serviços reterem grandes quantidades de dados de usuário (MAHESHWARI, 2020).

Em outro exemplo, o governo indiano promulgou recentemente “As Regras de Tecnologia da Informação (Diretrizes para Intermediários e Código de Ética para Mídia Digital) de 2021” que enfraquecem a liberdade de expressão (MAHESHWARI

& LLANSÓ, 2021), ao exigirem que grandes serviços de rede sociais, aqueles com mais de 5 milhões de usuários registrados, revelem o remetente original ou a origem de uma determinada mensagem.

A rastreabilidade como um conceito não é consistente com as garantias de privacidade para sistemas com *E2EE* e corrigir problemas de design nesses exemplos com falhas não resolverá essa tensão inerente.

O governo defendeu duas abordagens possíveis para fazer isso: exigir que os provedores incluam as informações de identidade criptografadas da origem nos metadados de cada mensagem; ou exigir que os provedores mantenham um banco de dados *hash* de todo o conteúdo veiculado em sua plataforma. Quando um conteúdo problemático for identificado, seu *hash* pode ser comparado aos *hashes* no banco de dados para identificar a origem. Essas propostas, bem como as próprias regras, são falhas de várias maneiras porque o conceito de uma origem é ambíguo, as propostas não são viáveis na prática e as regras são juridicamente questionáveis (MAHESHWARI & NOJEIM, 2021). E ainda mais importante, embora a proposta possa evitar a identificação de todos em uma cadeia de mensagens (ao contrário da proposta de Tyagi, Miers, *et al.*, (2019)), ela revela a identidade da origem a um terceiro, minando a garantia de privacidade da criptografia de ponta a ponta.

Em suma, dependendo da abordagem, a rastreabilidade pode fornecer a um terceiro acesso às informações sobre a origem de uma mensagem e todos os seus outros destinatários anteriores em uma cadeia de encaminhamento expandida, sem seu consentimento. De qualquer forma, isso significa que a rastreabilidade como um conceito não é consistente com as garantias de privacidade para sistemas com criptografia de ponta a ponta; e que corrigir problemas de design nesses exemplos falhos não resolverá essa tensão inerente.

Análise de Metadados

Análise de metadados – que são "dados sobre dados" (no caso, dados sobre uma mensagem criptografada) – pode incluir uma quantidade surpreendentemente robusta de detalhes, incluindo tamanho do arquivo, tipo, data/hora, remetente/destinatário, etc. A análise baseada em metadados é relevante, por exemplo, na detecção de *spam* em comunicações de texto simples. Com a criptografia de ponta a ponta, um provedor de serviços também pode detectar *spam* (ou conteúdo problemático semelhante) examinando o volume ou tamanho das mensagens enviadas por uma conta e tomar medidas se esse volume se desviar de sua classificação de atividade normal de mensagens.

As técnicas de aprendizado de máquina (*ML*, iniciais de "*machine learning*") podem ser aplicadas aos metadados para prever até que ponto um determinado usuário pode compartilhar conteúdo problemático em um serviço com *E2EE*. Aprendizado de máquina é um processo pelo qual um sistema analisa dados para extrair características e relacionamentos dentro dos

dados. Por exemplo, a análise de perfil não criptografado ou informações de bate-papo em grupo (por exemplo, fotos de perfil) usando classificadores de *ML* podem contribuir para a detecção de contas envolvidas na distribuição de material de abuso sexual de crianças em serviços com criptografia de ponta a ponta. De acordo com o WhatsApp, eles banem mais de 300.000 contas por mês por suspeita de compartilhamento de *CSAM* usando esses tipos de abordagens (WHATSAPP, 2021b).

As técnicas de aprendizado de máquina (*ML*, iniciais de "*machine learning*") podem ser aplicadas aos metadados para prever até que ponto um determinado usuário pode compartilhar conteúdo problemático em um serviço com *E2EE*. Aprendizado de máquina é um processo pelo qual um sistema analisa dados para extrair características e relacionamentos dentro dos dados. Por exemplo, a análise de perfil não criptografado ou informações de bate-papo em grupo (por exemplo, fotos de perfil) usando classificadores de *ML* podem contribuir para a detecção de contas envolvidas na distribuição de material de abuso sexual de crianças em serviços com criptografia de ponta a ponta. De acordo com o WhatsApp, eles banem mais de 300.000 contas por mês por suspeita de compartilhamento de *CSAM* usando esses tipos de abordagens (WHATSAPP, 2021b).

Em geral, desde que a análise de metadados ocorra exclusivamente no dispositivo de um usuário e não armazene, use ou envie mensagens descriptografadas, a privacidade do usuário é preservada e as garantias da criptografia de ponta a ponta não são violadas.

Essas técnicas também podem se concentrar no comportamento do usuário. Por exemplo, os modelos de aprendizado de máquina podem ser treinados sobre o comportamento de usuários que foram banidos de um serviço com *E2EE* (por exemplo, suas práticas de criação de conta, frequência de envio de mensagens ou denúncias por outros usuários sobre conteúdo problemático). Isso pode, então, ser usado para analisar o comportamento de novos usuários que desejam ingressar no serviço ou de usuários existentes. (WHATSAPP, 2019).

Obviamente, nem todas as análises de metadados podem identificar com segurança o conteúdo problemático. O aplicativo WhatsApp no dispositivo de um usuário coloca um rótulo no conteúdo que é encaminhado (compartilhado) com os usuários várias vezes (WHATSAPP, 2021a). No entanto, sua utilidade como técnica de detecção de conteúdo é limitada porque as mensagens encaminhadas com frequência não são necessariamente problemáticas – elas podem simplesmente ser interessantes.

Além disso, também é importante reconhecer os riscos à privacidade inerentes ao acesso de metadados por provedores de serviços, uma vez que podem ser usados para revelar dados delicados, como a identidade do remetente ou do destinatário (GRESCHBACH *et al.*, 2012). Esses riscos estão presentes no uso da análise de metadados em comunicações de texto simples e, potencialmente, também em sistemas com criptografia de

ponta a ponta. Limitar a análise de metadados ao dispositivo do usuário (e a um aplicativo nesse dispositivo) pode ser uma abordagem para reduzir esses riscos. Em geral, desde que a análise de metadados ocorra exclusivamente no dispositivo de um usuário e não armazene, use ou envie mensagens descriptografadas, a privacidade do usuário é preservada e as garantias de criptografia de ponta a ponta não são violadas.

Verificação perceptiva em criptografia de ponta a ponta

Existem duas categorias amplas de ferramentas de aprendizado de máquina usadas para detecção e análise de conteúdo: modelos de correspondência e modelos de predição (SHENKMAN et al., 2021), ambos propostos para sistemas com *E2EE*. Os modelos de correspondência de *ML* têm como objetivo reconhecer o conteúdo como idêntico ou suficientemente semelhante ao conteúdo visto anteriormente. Uma técnica importante é a verificação de *hash*, uma abordagem para criar uma impressão digital ou representação de uma parte do conteúdo para fins de comparação de uma forma que seja mais eficiente e flexível do que depender do conteúdo original para comparação. Essas impressões digitais podem ser comparadas entre si para identificar correspondências. Para ser útil na detecção de conteúdo indesejado em grande escala, os provedores de hospedagem de conteúdo contam com bancos de dados de *hashes* de conteúdo problemático previamente identificado. O provedor de hospedagem de conteúdo executa o algoritmo de *hash* em cada arquivo fornecido pelo usuário no *upload* e compara esse *hash* com os *hashes* no banco de dados.

Existem dois tipos principais de modelos de correspondência: verificação criptográfica e verificação perceptiva. O *hash* criptográfico usa uma função criptográfica para gerar uma impressão digital *hash* aleatória, que é altamente sensível a alterações. Esta abordagem pode ser eficaz em identificar conteúdo conhecido sem alterações. A verificação perceptiva, por outro lado, permite que o provedor de serviços determine o grau em que duas partes do conteúdo devem ser semelhantes para serem consideradas correspondentes. Isso pode ser importante quando pequenas alterações são feitas em uma parte do conteúdo para ignorar a detecção.

A verificação perceptiva é usada em um contexto de texto simples para identificar automaticamente o conteúdo que o host determinou anteriormente que não deseja em seu sistema; veja, por exemplo, a ferramenta de verificação material de abuso sexual de crianças da empresa Cloudflare, (PAINE & GRAHAM-CUMMING, 2019). É também o tipo de verificação que é objeto de propostas de pesquisa para detecção de conteúdo sob criptografia de ponta a ponta. A verificação perceptiva pode

Existem alguns bancos de dados hash que são compartilhados por toda a indústria, incluindo o banco de dados da ONG dos Estados Unidos NCMEC – National Center for Missing & Exploited Children (inglês para "Centro Nacional para Crianças Desaparecidas e Exploradas") de supostas imagens de material de abuso sexual de crianças e o banco de dados sobre potencial conteúdo terrorista e extremista violento compartilhado pela indústria do Global Internet Forum to Counter Terrorism ("Fórum Global da Internet para Conter o Terrorismo"). Além desses recursos compartilhados, alguns serviços podem criar seus próprios conjuntos de hash de conteúdo que desejam bloquear, por exemplo, o programa do Facebook para verificação de imagens íntimas enviadas pelo usuário que estão sendo compartilhadas de forma não consensual no serviço. https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/.

ser usada por um provedor de serviços ao receber conteúdo; isso é chamado de varredura pelo lado do servidor. Ou pode ser usada pelo aplicativo de mensagens (ou outro aplicativo) no dispositivo do usuário antes que o conteúdo seja enviado; isso é chamado de varredura pelo lado do cliente. Em ambos os casos, se uma correspondência for encontrada, a mensagem pode ser impedida de chegar ao destinatário e o usuário pode enfrentar consequências adicionais.

No caso de conversas criptografadas, pode-se usar a varredura do lado do servidor ou do cliente para detectar conteúdo indesejado ou abusivo, mas essas abordagens violam as garantias de privacidade que se espera de uma conversa criptografada, introduzem novas vulnerabilidades de segurança, ou ambos.

No caso de varredura do lado do servidor, o conteúdo seria convertido em *hashes* antes de ser criptografado e os *hashes* seriam enviados a um servidor da plataforma para verificação. Revelar o *hash* para o servidor, no entanto, é uma violação de privacidade, porque os *hashes* podem revelar informações sobre o conteúdo. Por exemplo, alguém com acesso ao servidor pode criar *hashes* de imagens específicas ou outro conteúdo de interesse e, quando houver uma correspondência, eles podem determinar quem enviou esse conteúdo ao servidor.

Embora essas soluções tentem minimizar a quantidade de informações reveladas ao servidor, saber que uma mensagem criptografada corresponde ou não a um conteúdo indesejado é uma violação da privacidade que se espera no armazenamento e nas conversas criptografados, já que um terceiro agora tem acesso a algumas informações sobre as comunicações.

Para lidar com essas limitações, uma proposta sugere uma abordagem de verificação perceptiva de preservação de privacidade que permitiria a um servidor da plataforma verificar se algum conteúdo criptografado corresponde a um material indesejado conhecido sem aprender mais nada sobre o próprio material (KULSHRESTHA & MAYER, 2021). Além disso, uma vez que esta solução ocorre pelo lado do servidor, o banco de dados de *hashes* indesejados seria mantido apenas pela plataforma e não distribuído aos dispositivos dos usuários. Embora essas soluções tentem minimizar a quantidade de informações reveladas ao servidor, saber que uma mensagem criptografada corresponde ou não a um conteúdo indesejado é uma violação da privacidade que esperamos do armazenamento criptografado e das conversas, já que um terceiro agora tem acesso a algumas informações sobre as comunicações. Essas soluções também podem produzir correspondências falsas, o que, dependendo da finalidade do banco de dados de *hash*, pode significar que o provedor de serviços interpretará o conteúdo analisado como abusivo; portanto, é crucial que eles incluam maneiras de corrigir falsas correspondências.[16]

16 As soluções tentam manter a taxa de falsa correspondência de verificação perceptiva em dados criptografados de ponta a ponta quase a mesma que a verificação perceptiva em dados de texto simples.

Por outro lado, com a varredura do lado do cliente, o conjunto de *hashes* indesejados é armazenado no dispositivo do usuário para que as comparações de *has*h possam ser feitas no dispositivo. Se os resultados da comparação de *hash* forem fornecidos apenas ao usuário, isso pode não violar as garantias de privacidade que esperamos de uma conversa criptografada; entretanto, se os resultados da comparação de *hash* forem compartilhados com o servidor, as garantias de privacidade da criptografia de ponta a ponta são violadas. Existem várias preocupações adicionais sobre a verificação perceptiva em texto simples que levantam questões sobre sua eficácia. Por exemplo, só é eficaz em conteúdo compartilhado mais de uma vez. Um estudo descobriu, no entanto, que 84% das imagens de material de abuso sexual de crianças relatadas (por provedores de serviços dos EUA usando ferramentas de detecção automatizadas ou pelo público dos EUA) foram relatadas apenas uma vez, demonstrando que grande parte do material abusivo é novo e, portanto, não poderia ser bloqueado por uma ferramenta de correspondência (BURSZTEIN *et al.*, 2019).

Além disso, a filtragem de *hash*, particularmente onde o algoritmo é público, também é vulnerável à adição deliberada de hashes ao banco de dados para gerar falsos positivos, ou seja, um ataque de envenenamento (DOLHANSKY & FERRER, 2020). Ataques de envenenamento podem ser usados para censurar a fala, adicionando *hashes* de material politicamente sensível em um banco de dados de *hash*. Os pesquisadores já encontraram algumas evidências disso na China, onde os atores podem estar usando a varredura pelo lado do cliente para essa finalidade (KNOCKEL *et al.*, 2020). E mesmo que a varredura do lado do cliente e do servidor seja usada apenas para detectar qualquer coisa que possa ser definida como conteúdo abusivo ou prejudicial em uma jurisdição, construir a tecnologia em plataformas e dispositivos de usuário pode permitir esse tipo de exploração por governos autoritários (ou não autoritários) em outro lugar (PFEFFERKORN, 2020).

Em uma implementação de varredura do lado do cliente, o conjunto completo de hashes a serem bloqueados é armazenado no dispositivo de cada usuário, tornando esse conjunto de hash potencialmente detectável por um usuário mal intencionado.

A implementação prática da varredura pelo lado do cliente em criptografia de ponta a ponta introduz essas vulnerabilidades no sistema. Em particular, distribuir o banco de dados de *hashes* indesejados para dispositivos de usuários pode permitir que atores mal intencionados subvertam o processo de detecção, manipulando o banco de dados de *hash*.

Outra proposta de Reis *et al.*, (2020) explora a ideia de usar a verificação perceptiva para detectar desinformação[17] no WhatsApp e, por extensão, outros serviços com criptografia de ponta a ponta, usando um modelo de varredura do lado do cliente. O trabalho deles envolveu principalmente a compreensão dos padrões de compartilhamento de informações no WhatsApp,

17 NdT. 9: O texto original usa o termo "*misinformation*" nesse trecho.

mas no processo propôs uma arquitetura que poderia ser introduzida no WhatsApp para detectar e sinalizar informações incorretas nos dispositivos dos usuários. Em sua proposta, o Facebook manteria um conjunto de *hashes* perceptivas para imagens que checadores de fatos consideraram desinformação (por exemplo, imagens compartilhadas fora do contexto ou manipuladas usando técnicas simples para criar as chamadas "falsificações baratas" (PARIS & DONOVAN, 2019)). Esses *hashes* são então armazenados diretamente no dispositivo do usuário e atualizados periodicamente. Ao enviar uma imagem, seu *hash* seria comparado aos já armazenados no dispositivo do remetente e outros avisos ou notificações poderiam ser exibidos ao remetente se o conteúdo for identificado como informação incorreta. A mesma verificação pode ser feita no dispositivo do destinatário e com avisos e notificações semelhantes exibidos.

Nessa proposta, os usuários que enviam essas informações não seriam sinalizados pela plataforma. As ações de notificação são realizadas apenas nos dispositivos do remetente e dos destinatários. Os destinatários podem, então, optar por denunciar a mensagem, mas não há meios automatizados de responsabilidade para enviar ou receber informações incorretas de acordo com esta proposta. No entanto, embora esta abordagem de varredura do lado do cliente possa proteger a privacidade do usuário (a detecção é feita apenas no dispositivo do usuário e nenhum terceiro está envolvido), ela também pode ser contornada, o que pode limitar sua utilidade geral. Em uma implementação de varredura do lado do cliente, o conjunto completo de *hashes* a serem bloqueados é armazenado no dispositivo de cada usuário, tornando esse conjunto de *hashes* potencialmente detectável por um usuário mal intencionado. Pode ser possível que um adversário malicioso use este conjunto de *hashes* para identificar quais imagens são refletidas no banco de dados (por exemplo, fazendo *hash* das imagens que eles desejam compartilhar e determinando se há um *hash* correspondente no banco de dados). Isso pode permitir que o adversário malicioso desenvolva métodos de aplicação de transformações ao conteúdo que são refletidas no banco de dados, a fim de evitar a detecção.

Essa fraqueza no modelo de varredura do lado do cliente se aplica à maioria das formas desta varredura. É particularmente problemático em casos de uso que envolvem o compartilhamento de mídia ilegal, como material de abuso sexual de crianças, nos quais agentes mal intencionados podem estar altamente motivados para desenvolver métodos de fraude. Na verdade, isso explica parcialmente porque algoritmos para métodos bem conhecidos, como PhotoDNA[18], não são públicos ou executados localmente em dispositivos, pois podem ser

18 Uma ferramenta desenvolvida pela Microsoft e Dartmouth College para detectar *CSAM* usando verificação preditiva (MICROSOFT, n.d.).

vulneráveis a ataques.

Além dessas limitações, pode haver outras considerações de implementação, como a capacidade de processamento do dispositivo, armazenamento, conectividade com a Internet e uso da bateria. Isso pode ter implicações importantes para a equidade, dependendo do contexto em que essa abordagem é usada. Por exemplo, entre populações, países ou regiões de baixa renda, *feature phones* ou *smartphones* baratos são frequentemente usados com o WhatsApp em vez de *smartphones* mais poderosos como iPhones e Samsung que são populares em países de renda alta (JAMES, 2020).

Em suma, as técnicas de correspondência de *hash*, como varredura do lado do servidor ou do lado do cliente, fornecem a terceiros o acesso à mensagem, introduzem vulnerabilidades de segurança significativas no sistema ou ambos. Mesmo propostas para varredura do lado do cliente que não envolvem acesso de terceiros (ou seja, apenas o remetente ou destinatário são notificados sobre a detecção de conteúdo indesejado) apresentam o potencial de manipulação do banco de dados de *hash* por malfeitores. Essas abordagens, portanto, não são consistentes com as garantias de privacidade e segurança da criptografia de ponta a ponta.

Modelos Preditivos para detecção de conteúdo em criptografia de ponta a ponta

A segunda categoria de técnicas de aprendizado de máquina consiste em modelos de predição que visam reconhecer as características do conteúdo com base no aprendizado prévio da máquina. Essa abordagem é frequentemente usada para conteúdo novo ou anteriormente desconhecido. Requer (geralmente grandes quantidades de) dados para treinar o modelo para prever se uma parte do conteúdo tem determinados atributos. Isso inclui modelos de visão computacional, que cobrem a análise de formas, texturas, cores, etc., e modelos de audição por computador, que se concentram em conteúdo de áudio. Um exemplo básico de uma dessas técnicas, um classificador de imagens, pode procurar prever se uma imagem enviada por um usuário é um cachorro ou gato. Em uma configuração de texto simples, as ferramentas de aprendizado de máquina são usadas para detecção e identificação de conteúdo baseado em texto (ver DUARTE *et al.*, 2017) e conteúdo multimídia (SHENKMAN *et al.*, 2021).

Com base nessas técnicas, Mayer (2019) oferece diretrizes para pesquisadores desenvolverem modelos para prever a existência de conteúdo problemático em um contexto de criptografia de ponta a ponta. Uma maneira de fazer isso poderia ser

usar algoritmos de aprendizado de máquina para detectar mensagens problemáticas de texto simples (como *spam*) usando classificadores pré-treinados instalados, por exemplo, em um aplicativo de mensagens ou outro aplicativo no dispositivo de um usuário. Depois que um usuário descriptografar uma mensagem, o classificador pode sinalizar a mensagem como possível *spam*. Novamente, existem limitações práticas a serem consideradas, como a vida útil da bateria do telefone e os recursos de processamento. Tal como acontece com a análise de metadados, se esse processo ocorrer exclusivamente no dispositivo de um usuário e nenhuma informação sobre a mensagem for divulgada a terceiros, as garantias de criptografia de ponta a ponta podem não ser violadas. No entanto, mais pesquisas são necessárias para desenvolver técnicas viáveis usando essa abordagem.

Moderação de conteúdo em ambientes com criptografia de ponta a ponta - Próximas etapas para pesquisa

Nós definimos ambientes com criptografia de ponta a ponta como um serviço ou aplicativo em que as chaves usadas para criptografar e descriptografar dados são conhecidas apenas pelos remetentes e destinatários desses dados. Uma parte crucial disso é o princípio de ponta a ponta. Os terceiros que roteiam, armazenam, fazem *backup* e processam os dados criptografados não têm acesso às chaves e, portanto, não podem obter nenhuma informação sobre os dados.

Usando essa definição, avaliamos as propostas técnicas atuais que visam fornecer alguma forma de detecção de conteúdo em serviços com criptografia de ponta a ponta. Nossa avaliação identificou duas propostas que preservam as garantias de segurança e privacidade da *E2EE* sem introduzir nenhuma nova vulnerabilidade de segurança no sistema. A primeira é a denúncia por usuário, que inclui o franqueamento de mensagens, um meio para o provedor de serviços autenticar que o remetente realmente enviou o conteúdo que foi denunciado como problemático pelo destinatário. O franqueamento de mensagens permite que o usuário denuncie conteúdo problemático, tal como conteúdo abusivo, informação errada e desinformação ,ou material de abuso sexual de crianças, incluindo configurações de conversas criptografadas de um para um e em grupo. A segunda abordagem para detecção de conteúdo que é consistente com as promessas da criptografia de ponta a ponta é o uso de análise de metadados, que poderia ser usada, por exemplo, para detectar conteúdo problemático como *spam* e material de abuso sexual de crianças.

Existem várias outras propostas para habilitar a detecção de conteúdo em sistemas com criptografia de ponta a

ponta, mas por diferentes razões, elas introduzem novas vulnerabilidades no sistema; são incapazes de fornecer as garantias de privacidade e segurança que o usuário espera; ou ainda não são viáveis. Embora algumas possam parecer promissoras no início (por exemplo, REIS *et al.*, 2020), elas apresentam vulnerabilidades como as decorrentes de se ter o banco de dados de *hash* no dispositivo do usuário. Outros violam efetivamente as garantias de privacidade da *E2EE* (por exemplo, a rastreabilidade). E ainda outras abordagens potencialmente promissoras (MAYER, 2019), como classificadores de aprendizado por máquina que operam apenas no dispositivo do usuário e não divulgam informações a terceiros, ainda não estão disponíveis e apontam áreas para investigação, especialmente em termos de modelos de predição para detecção de conteúdo abusivo no dispositivo.

• Explorar a aplicabilidade das abordagens de moderação de conteúdo para além da detecção de conteúdo em serviços com criptografia de ponta a ponta. Cada serviço com *E2EE* é diferente, mas algumas ferramentas e abordagens úteis no contexto de texto simples podem ser relevantes em ambientes de criptografia de ponta a ponta. A educação do usuário sobre as políticas de um serviço ou regras de um fórum específico, opções para os usuários atuarem como moderadores em discussões com várias pessoas e a avaliação precisa do conteúdo relatado são todos pontos de potencial intervenção em um sistema de moderação de conteúdo. As ferramentas automatizadas que estão atualmente configuradas para detectar, avaliar e aplicar medidas contra o conteúdo em um único processo podem ser reconfiguradas e incorporadas aos sistemas de moderação que dependem mais de denúncias por usuários. Esses tipos de intervenções podem ser especialmente úteis para pensar sobre como projetar serviços com *E2EE* para reduzir a probabilidade de conteúdo e atividades abusivas, como assédio e discurso de ódio.

• As soluções de detecção de conteúdo devem enfatizar a agência do usuário, que é o caso das denúncias por usuário, incluindo o franqueamento de mensagens. A análise de metadados combinada com a denúncia por usuário pode permitir que o usuário determine as ações apropriadas onde o conteúdo problemático for detectado. Isso também pode incluir pesquisas adicionais sobre como permitir que os usuários escolham ou criem seus próprios filtros para bloquear conteúdo indesejado no aplicativo com criptografia de ponta a ponta em seus dispositivos, desde que esses filtros não exponham informações sobre as mensagens a terceiros.

• Se os serviços com criprografia de ponta a ponta precisarão depender substancialmente de denúncias por usuários para detectar conteúdo indesejado ou potencialmente ilegal, então, significativamente mais pesquisas são necessárias

para determinar as técnicas mais eficazes para encorajar a denúncia de conteúdo por usuários, como design de interface do usuário, alinhamento dos usuários com os valores do serviço e promoção do desenvolvimento de comunidades saudáveis.

• Pesquisas adicionais são necessárias para evitar abusos por parte de infratores reincidentes, incluindo usuários que foram proibidos de criar novas contas por um *E2EE* ou serviço semelhante.

• As propostas devem ser explícitas sobre as propriedades exatas que elas garantem, e que qualquer alteração de um sistema precisa de notificação, consentimento e opção de sair ("*opt-out*") do usuário. É essencial basear a pesquisa no princípio da ponta a ponta que possibilita manter as garantias de segurança e privacidade esperadas pelo usuário.

Denúncias por usuários e análise de metadados fornecem ferramentas eficazes na detecção de quantidades significativas e diferentes tipos de conteúdo problemático em serviços com E2EE, incluindo mensagens abusivas e de assédio, spam, informações erradas e desinformação, bem como material de abuso sexual de crianças.

Em suma, observamos que denúncias por usuários e análise de metadados fornecem ferramentas eficazes na detecção de quantidades significativas e diferentes tipos de conteúdo problemático em serviços com criptografia de ponta a ponta, incluindo mensagens abusivas e de assédio, *spam*, informações erradas e desinformação, bem como material de abuso sexual de crianças. No entanto, ainda é necessário mais trabalho. Incentivamos a realização de pesquisas adicionais sobre moderação de conteúdo em serviços com *E2EE* com base no próprio uso de metadados e métodos que atuam dentro dos limites de um aplicativo de mensagens no dispositivo de um usuário para capacitar o usuário a sinalizar, ocultar ou denunciar conteúdo indesejado para o prestador de serviços. Esses métodos não devem modificar de forma alguma os esquemas de criptografia subjacentes, nem interferir nas garantias de privacidade e segurança da criptografia de ponta a ponta. Em última análise, devemos reconhecer que as soluções tecnológicas para detectar apenas conteúdo problemático, seja em um sistema de texto simples ou com criptografia de ponta a ponta, não endereçarão os problemas maiores de, digamos, distribuição de desinformação ou material de abuso sexual de crianças. Em vez disso, como sociedade, também precisamos considerar as causas sociais e políticas por trás desses fenômenos e endereçá-los em sua essência.

Apêndice: Resumos ampliados de algumas propostas para detectar conteúdo em plataformas com criptografia de ponta a ponta

Grubbs, P., Lu, J., & Ristenpart, T. (2017)

Os autores formalizam uma definição de segurança para franqueamento de mensagens e avaliam o esquema do Facebook. No processo, eles também formalizam as propriedades de vinculação do remetente, o que garante que, uma vez que uma mensagem tenha sido atribuída, uma atribuição será posteriormente verificada corretamente, dada a mensagem original. Isso impede que um remetente possa enviar uma mensagem ao destinatário que descriptografa corretamente, mas não verifica no momento da denúncia; e a vinculação de recebimento garante que, uma vez que um destinatário receba uma atribuição para uma mensagem, ele não possa abrir essa atribuição para qualquer outra mensagem que não a enviada originalmente. Isso impede que o destinatário possa atribua ao remetente um conteúdo que ele não enviou.

Isso se baseia no trabalho anterior através da introdução de um esquema mais eficiente que permite a franquia de anexos de arquivos. Embora o esquema anterior possa ser eficiente para mensagens de texto, anexos de arquivos grandes contendo imagens ou vídeos exigem um mecanismo de franquia mais rápido. Os autores são motivados diretamente pelo esquema de franquia de anexo de arquivo do Facebook, que, embora seja eficiente, não cumpre a propriedade de vinculação do remetente discutida anteriormente.

Dodis, Y., Grubbs, P., Ristenpart, T., & Woodage, J. (2018)

Os autores primeiro demonstram um ataque contra o esquema de franquia de anexos do Facebook. Eles fazem isso explorando o fato de o Facebook usar o AES-GCM, um esquema AEAD seguro, em uma configuração não padrão. Um remetente de mensagem maliciosa sob o esquema de franquia do Facebook é capaz de construir textos cifrados de uma maneira que impede um destinatário de denunciar uma mensagem maliciosa, violando, assim, a segurança de vinculação do remetente. Esse ataque é realizado por um remetente malicioso criando primeiro duas mensagens diferentes, m1,m2, onde m1 é inócuo e m2 contém conteúdo malicioso. O remetente malicioso pode então encontrar duas chaves, k1,k2, de modo que a criptografia de ambas as mensagens, sob chaves diferentes, resulte no mesmo texto cifrado (Enc(k1,m1) = Enc(k2,m2)). O remetente envia primeiro a mensagem inócua e depois a mensagem maliciosa.

Quando criptografadas, essas duas mensagens produzem texto cifrado idêntico, portanto, o Facebook pode assumir internamente que a segunda mensagem é uma duplicata. Se o Facebook entregar a segunda mensagem, mas não a sinalizar, uma denúncia de abuso posterior do destinatário falhará.

Os autores apontam que isso demonstra a necessidade de um meio mais eficiente de atribuir anexos de arquivos sem violar a segurança de vinculação do remetente, o que os motivou a criar uma primitiva segura de uso único chamada criptoatribuição[19]. Eles, então, constroem um esquema chamado encadeamento de função *hash* que permite a franquia de um anexo de arquivo com um único cálculo SHA-256 ou SHA-3. A segurança subjacente vem de uma propriedade dos tipos de funções de *hash* usadas, chamada resistência à colisão. Sob resistência à colisão, um adversário mal intencionado computacionalmente limitado não pode encontrar duas entradas com *hash* para a mesma saída. Aproveitando essa propriedade, os autores encadeiam várias chamadas de função de *hash*, com o *xor* de uma chave e o anexo como entrada. Isso produz uma atribuição de vinculação para um determinado anexo de arquivo.

Usando esse esquema, eles são capazes de alcançar a não-falseabilidade, segurança de vinculação ao remetente, segurança de vinculação ao destinatário e refutabilidade. Eles também aproveitam os resultados intermediários da computação de uma atribuição como material chave para criptografar o próprio anexo.

Chen, L., & Tang, Q. (2018)

Em trabalhos anteriores, o franqueamento exige que um destinatário de conteúdo malicioso revele o conteúdo completo da mensagem ao denunciar uma mensagem. Isso pode levar os destinatários a não denunciar uma mensagem com conteúdo malicioso por medo de revelar outras informações confidenciais. Um remetente de mensagens mal intencionadas também pode tirar proveito disso e enviar mensagens mal intencionadas com informações confidenciais sobre o destinatário para impedir que ele denuncie essas mensagens posteriormente. Para resolver esses problemas, os autores criam um esquema que permite que partes de mensagens sejam reveladas ao denunciar.

Tyagi, N., Grubbs, P., Len, J., Miers, I.,

19 Ndt.10: Optou-se pela palavra “criptoatribuição”, pela aglutinação das palavras “criptografia” e “atribuição”, para corresponder ao neologismo “*encryptment*”, no original em inglês composto pelos termos “*encryption*” e “*commitment*”, respectivamente, “encriptação” e “atribuição”, em tradução literal.

& Ristenpart, T. (2019)

O franqueamento assimétrico de mensagens estende o trabalho anterior nessa área para abordar plataformas de mensagens que empregam a privacidade de metadados. No contexto anterior, se o usuário A enviar uma mensagem com criptografia de ponta a ponta para o usuário B, a plataforma não poderá descobrir o conteúdo da mensagem, a menos que um usuário denuncie a mensagem. A plataforma, no entanto, vê que uma mensagem foi enviada do usuário A para B. No contexto da privacidade de metadados, quando o usuário A enviou uma mensagem para B, a plataforma vê apenas que uma mensagem foi enviada, e não as identidades do remetente ou destinatário. As propostas anteriores usam a identidade do usuário como parte da atribuição de uma mensagem e, portanto, não funcionariam nessa configuração.

Esse trabalho apresenta um esquema de franqueamento assimétrico de mensagens que é compatível com plataformas de mensagens com criptografia de ponta a ponta privadas de metadados que cumprem com o requisito de não-falseabilidade, a vinculação do receptor e do remetente e a negabilidade. Em plataformas privadas com criptografia de ponta a ponta de metadados, cada usuário tem duas chaves, uma chave privada e uma chave pública. Se o usuário A enviar uma mensagem ao usuário B, A primeiro deve assinar digitalmente o seguinte e criar uma atribuição usando a chave privada de A: a mensagem, a chave pública de B, a chave pública do moderador. Um moderador pode verificar posteriormente essa atribuição usando sua chave privada, as chaves públicas de A e B e a mensagem denunciada.

Kulshrestha, A., & Mayer, J. (2021)

Este trabalho mostra que a verificação perceptiva ainda pode ocorrer em cima de mensagens com criptografia de ponta a ponta de uma maneira em que um servidor apenas aprende se o conteúdo de uma mensagem criptografada corresponde a um conteúdo nocivo conhecido sem saber qual é a mensagem original do usuário ou o *hash*. O usuário também não aprende nada sobre o conteúdo do banco de dados. Os autores fazem uso de uma técnica criptográfica chamada recuperação de informações privadas, que permite que um elemento seja recuperado do banco de dados sem que ele saiba qual era o elemento.

Embora isso reduza o risco de revelar *hashes* ao servidor, ainda existem os mesmos riscos de permitir a vigilância. Mesmo um protocolo que não revela nenhuma informação sobre a mensagem original pode ser abusado pelas plataformas para realizar vigilância em cima da criptografia de ponta a ponta. Em vez de encontrar correspondências com conteúdo prejudicial,

uma plataforma também pode substituir o *hash* pelo relativo a outro conteúdo que pode ser desfavorável à plataforma ou talvez a um governo específico. Os autores também observam que a implementação desses protocolos no contexto de um governo democrático pode permitir que governos mais autoritários usem as mesmas ferramentas para suprimir o discurso e conduzir a vigilância.

Tyagi, N., Miers, I., & Ristenpart, T. (2019)

Os autores introduzem um esquema que permite o rastreamento retroativo de conteúdo malicioso, mas, como isso não captura todos os usuários que podem ter recebido o conteúdo, eles propõem um segundo esquema. Um sistema de moderação pode então rastrear pra trás e pra frente todos os indivíduos que podem ter recebido o conteúdo malicioso e notificar os usuários, bem como identificar a fonte do conteúdo. Antes de uma denúncia por usuário, a confidencialidade é preservada e apenas o remetente e o destinatário de uma mensagem podem descriptografá-la. Após uma denúncia, a plataforma, por meio do rastreamento, aprende o conteúdo das mensagens que não foram denunciadas diretamente na cadeia de encaminhamento. Embora isso permita que uma plataforma rastreie a disseminação de conteúdo malicioso viral, também oferece uma oportunidade para usuários maliciosos denunciarem conteúdo sensível e expor a privacidade de todos os remetentes e destinatários na cadeia.

Os autores conseguem rastrear conteúdo malicioso pra trás e pra frente introduzindo dois esquemas separados para (1) rastreamento pra trás e (2) rastreamento pra trás e pra frente. No primeiro esquema, os autores criam uma cadeia do que chamam de indicadores criptografados, no qual as mensagens encaminhadas apontam para o remetente anterior da mensagem. Quando uma mensagem é enviada, o remetente faz uma amostra de uma chave de rastreamento e a usa para criar uma atribuição da mensagem. Se um usuário A enviar uma mensagem (m) para o usuário B, que então encaminha a mensagem (m) para o usuário C, o usuário A primeiro analisa aleatoriamente uma amostra de um chave de rastreamento kA e a usa para criar uma confirmação da ligação com (m). Essa atribuição, $comm$A,juntamente com a kA, é enviada para B. Quando B encaminha (m) para C, B analisa uma amostra de sua própria chave, kB, e, então, criptografa a kA usando a kB. Isso cria o que é chamado de indicador criptografado, que aponta de volta para o remetente. Ao iniciar um rastreamento a partir de C, a kB é usada para recuperar a kA, que pode ser usada para verificar a atribuição de A com a mensagem (m). Esse esquema funciona para rastreamento de encaminhamento usando o mesmo conceito e cria indicadores criptografados que apontam para destinatários e remetentes.

Os autores são capazes de obter a vinculação do destinatário e a vinculação do remetente de maneira semelhante ao trabalho anterior, contando com a propriedade de resistência à colisão subjacente ao seu esquema de confirmação. Eles também mantêm a possibilidade de repúdio, apenas permitindo que a plataforma realize o rastreamento.

Referências

Abelson, H., Anderson, R., Bellovin, S. M., Benaloh, J., Blaze, M., Diffie, W., Gilmore, J., Green, M., Landau, S., Neumann, P. G., Rivest, R. L., Schiller, J. I., Schneier, B., Specter, M. A., & Weitzner, D. J. (2015). Keys under doormats: Mandating insecurity by requiring government access to all data and communications. *Journal of Cybersecurity*, 1(1), 69–79. https://doi.org/10.1093/cybsec/tyv009

Bai, W., Pearson, M., Kelley, P. G., & Mazurek, M. L. (2020). Improving Non-Experts' Understanding of End to-End Encryption: An Exploratory Study. *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, 210–219. https://doi.org/10.1109/EuroSPW51379.2020.00036

Bloch-Wehba, H. (2020). Automation in Moderation. *Cornell International Law Journal*, 53, 41–96.

Bursztein, E., Clarke, E., DeLaune, M., Elifff, D. M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K., & Bright, T. (2019). Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. *The World Wide Web Conference*, 2601–2607. https://doi.org/10.1145/3308558.3313482

Caplan, R. (2018). *Content or Context Moderation?* Data & Society. https://datasociety.net/library/content-or context-moderation/

Chen, L., & Tang, Q. (2018). People Who Live in Glass Houses Should not Throw Stones: Targeted Opening Message Franking Schemes. *IACR Cryptol. EPrint Arch.*, 2018, 994.

Dodis, Y., Grubbs, P., Ristenpart, T., & Woodage, J. (2018). Fast Message Franking: From Invisible Salamanders to Encryptment. In H. Shacham & A. Boldyreva (Eds.), *Advances in Cryptology – CRYPTO 2018* (pp. 155–186). Springer International Publishing. https://doi.org/10.1007/978-3-319-96884-1_6

Dolhansky, B., & Ferrer, C. C. (2020). Adversarial collision attacks on image hashing functions. *ArXiv:2011.09473 [Cs]*. http://arxiv.org/abs/2011.09473

Duarte, N., Llansó, E., & Loup, A. C. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy & Technology. https://cdt.org/insights/mixed-messages-the-limits-of automated-social-media-content-analysis/

Facebook. (2016). *Facebook: Messenger Secret Conversations—Technical Whitepaper*. Facebook. https://about. fb.com/wp-content/uploads/2016/07/messenger-secret-conversations-technical-whitepaper.pdf

Federal Bureau of Investigation. (2014, October 16). *Going Dark: Are Technology, Privacy, and Public Safety on a Collision Course?* [Speech]. Federal Bureau of Investigation. https://www.fbi.gov/news/speeches/going dark-are-technology-privacy-and-public-safety-on-a-collision-course

Fiesler, C., Jiang, J., McCann, J., Frye, K., & Brubaker, J. (2018). Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), Article 1. https://ojs.aaai.org/index.php/ICWSM/article/view/15033

François, C. (2020). *Actors, Behaviors, Content: A Disinformation ABC* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.). Transatlantic Working Group. https://cdn. annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/ABC_Framework_TWG_Francois_ Sept_2019.pdf

Gillespie, T. (2012). The dirty job of keeping Facebook clean. *Culture Digitally*, 22. https://culturedigitally. org/2012/02/the-dirty-job-of-keeping-facebook-clean/

Global Encryption Coalition. (2020). *Breaking Encryption Myths – Global Encryption Coalition*. Global Encryption Coalition. https://www.globalencryption.org/2020/11/breaking-encryption-myths/

Goldman, E. (2021). Content Moderation Remedies. *Michigan Technology Law Review*, Forthcoming. https:// papers.ssrn.com/sol3/papers.cfm?abstract_id=3810580

Green, M. (2019, December 8). *Can end-to-end encrypted systems detect child sexual abuse imagery? A Few Thoughts on Cryptographic Engineering*. https://blog.cryptographyengineering.com/2019/12/08/on-client-side media-scanning/

Greschbach, B., Kreitz, G., & Buchegger, S. (2012). The devil is in the metadata—New privacy challenges in Decentralised Online Social Networks. *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, 333–339. https://doi.org/10.1109/PerComW.2012.6197506

Grimmelmann, J. (2015). The virtues of moderation. *Yale JL & Tech.*, 17, 42.

Grubbs, P., Lu, J., & Ristenpart, T. (2017). Message Franking via Committing Authenticated Encryption. In J. Katz & H. Shacham (Eds.), *Advances in Cryptology – CRYPTO 2017* (pp. 66–97). Springer International Publishing. https://doi.org/10.1007/978-3-319-63697-9_3

International Centre for Missing & Exploited Children (ICMEC). (2018). *Child Sexual Abuse Material: Model Legislation & Global Review* (9th Edition). International Centre for Missing & Exploited

Children (ICMEC). https://cdn.icmec.org/wp-content/uploads/2018/12/CSAM-Model-Law-9th-Ed FINAL-12-3-18-1.pdf

James, J. (2020). The smart feature phone revolution in developing countries: Bringing the internet to the bottom of the pyramid. *The Information Society*, 36(4), 226–235. https://doi.org/10.1080/01972243.2020.17614 97

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer*

Interaction, 3(CSCW), 150:1-150:27. https://doi.org/10.1145/3359252

Kaye, D. (2015). *Report on encryption, anonymity, and the human rights framework* (A/HRC/29/32). OHCHR. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/CallForSubmission.aspx

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev., 131*, 1598.

Klonick, K. (2020). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *Yale Law Journal, 129*, 2418–2434.

Knockel, J., Parsons, C., Ruan, L., Xiong, R., Crandall, J., & Deibert, R. (2020). *We Chat, They Watch: How International Users Unwittingly Build up WeChat's Chinese Censorship Apparatus*. https://citizenlab. ca/2020/05/we-chat-they-watch/

Knodel, M., Baker, F., Kolkman, O., Celi, S., & Grover, G. (2021). *Definition of End-to-end Encryption (IETF Active Internet-Draft)*. IETF. https://datatracker.ietf.org/doc/draft-knodel-e2ee-definition/

Kulshrestha, A., & Mayer, J. (2021). *Identifying Harmful Media in End-to-End Encrypted Communication: Efficient Private Membership Computation*. 30th {USENIX} Security Symposium ({USENIX} Security 21). https://www.usenix.org/conference/usenixsecurity21/presentation/kulshrestha

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*,

543–550. https:// doi.org/10.1145/985692.985761

Leontiadis, I., & Vaudenay, S. (2018). Private Message Franking with After Opening Privacy. *IACR Cryptol. EPrint Arch.*, 2018, 938.

Maheshwari, N. (2020, June 23). Traceability Under Brazil's Proposed Fake News Law Would Undermine Users' Privacy and Freedom of Expression. *Center for Democracy and Technology*. https://cdt.org/insights/ traceability-under-brazils-proposed-fake-news-law-would-undermine-users-privacy-and-freedom-of expression/

Maheshwari, N., & Llansó, E. (2021). Part 1: New Intermediary Rules in India Imperil Free Expression, Privacy and Security. Center for Democracy and Technology. https://cdt.org/insights/ part-1-new-intermediary rules-in-india-imperil-free-expression-privacy-and-security/

Maheshwari, N., & Nojeim, G. (2021). Part 2: New Intermediary Rules in India Imperil Free Expression, Privacy and Security. *Center for Democracy and Technology*. https://cdt.org/insights/ part-2-new-intermediary-rules in-india-imperil-free-expression-privacy-and-security/

Marlinspike, M. (2013). Simplifying OTR deniability. *Signal Messenger*. https://signal.org/blog/simplifying-otr deniability/

Masnick, M. (2018, August 9). Platforms, Speech And Truth: Policy, Policing And Impossible Choices. *Techdirt*. https://www.techdirt.com/articles/20180808/17090940397/platforms-speech-truth-policy-policing impossible-choices.shtml

Mayer, J. (2019). Content Moderation for End-to-End Encrypted Messaging. *Princeton University*. https://www. cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf

Microsoft. (n.d.). *PhotoDNA* | Microsoft. Retrieved March 24, 2021, from https://www.microsoft.com/en-us/ photodna

Murdock, J. (2020, July 3). *Earn It Act gets "unanimous approval" but experts warn of privacy risks.* Newsweek. https://www.newsweek.com/senate-earn-it-act-encryption-privacy-aclu-human-rights-watch-eff-internet freedoms-1515326

National Academies of Sciences, Engineering, and Medicine. (2018). *Decrypting the Encryption Debate: A Framework for Decision Makers*. https://doi.org/10.17226/25010

Newman, L. H. (2020, March 5). *The EARN IT Act Is a Sneak Attack on Encryption*. Wired. https://www.wired. com/story/earn-it-act-sneak-attack-on-encryption/

Paine, J., & Graham-Cumming, J. (2019, December 18). Announcing the CSAM Scanning Tool, Free for All Cloudflare Customers. *The Cloudflare Blog*. https://blog.cloudflare.com/the-csam-scanning-tool/

Paris, B., & Donovan, J. (2019). *Deepfakes and Cheap Fakes*. Data & Society. https://datasociety.net/library/ deepfakes-and-cheap-fakes/

Pfefferkorn, R. (2020, May 11). Client-Side Scanning and Winnie-the-Pooh Redux (Plus Some Thoughts on Zoom). *Center for Internet and Society*. http://cyberlaw.stanford.edu/blog/2020/05/client-side-scanning and-winnie-pooh-redux-plus-some-thoughts-zoom

Reis, J. C. S., Melo, P., Garimella, K., & Benevenuto, F. (2020). Can WhatsApp benefit from debunked fact checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review*. https://doi. org/10.37016/mr-2020-035

Ruane, K. (2020, June 30). *The EARN IT Act is a Disaster for Online Speech and Privacy, Especially for the LGBTQ and Sex Worker Communities*. American Civil Liberties Union. https://www.aclu.org/news/free speech/the-earn-it-act-is-a-disaster-for-online-speech-and-privacy-especially-for-the-lgbtq-and-sex-worker communities/

Shenkman, C., Thakur, D., & Llansó, E. (2021). *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*. Center for Democracy & Technology. https://cdt.org/insights/do-you-see what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/

Swartz, A. (2006, September 7). Who Runs Wikipedia? (Aaron Swartz's Raw Thought). *Raw Thought*. http:// www.aaronsw.com/weblog/whorunswikipedia

Swire, P., & Ahmad, K. (2011). Encryption and Globalization. *Columbia Science and Technology Law Review, 13*, 416.

Thompson, A. W., & Park, C. (2020). *Privacy's Best Friend: How Encryption Protects Consumers, Companies, and Governments Worldwide (Exploring the Twenty-First Century Privacy Debate)*. New America. http:// newamerica.org/oti/reports/privacys-best-friend/

Twitch. (2021, March 2). *Twitch.tv—Transparency Report*. Twitch. Tv. https://www.twitch.tv/p/en/legal/ transparency-report/

Tyagi, N., Grubbs, P., Len, J., Miers, I., & Ristenpart, T. (2019). Asymmetric Message Franking: Content Moderation for Metadata-Private End-to-End Encryption. In A. Boldyreva

& D. Micciancio (Eds.), *Advances in Cryptology – CRYPTO 2019* (pp. 222–250). Springer International Publishing. https://doi.org/10.1007/978-3-030-26954-8_8

Tyagi, N., Miers, I., & Ristenpart, T. (2019). Traceback for End-to-End Encrypted Messaging. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 413–430. https://doi. org/10.1145/3319535.3354243

Uhlig, U., Knodel, M., Oever, N. ten, & Cath, C. (2021). *How the Internet Really Works*. No Starch Press. https:// nostarch.com/how-internet-really-works

U.S. Department of Justice. (2020, October 11). I*nternational Statement: End-To-End Encryption and Public Safety*. https://www.justice.gov/opa/pr/international-statement-end-end-encryption-and-public-safety

WhatsApp. (2019). *Stopping Abuse: How WhatsApp Fights Bulk Messaging and Automated Behavior*. Facebook. https://faq.whatsapp.com/general/security-and-privacy/unauthorized-use-of-automated-or-bulk messaging-on-whatsapp/?lang=en

WhatsApp. (2021a). *WhatsApp Help Center—About forwarding limits*. WhatsApp.Com. https://faq.whatsapp. com/general/chats/about-forwarding-limits/?lang=en

WhatsApp. (2021b). *WhatsApp Help Center—How WhatsApp Helps Fight Child Exploitation*. WhatsApp.Com. https://faq.whatsapp.com/general/how-whatsapp-helps-fight-child-exploitation/?lang=en


CENTER FOR
DEMOCRACY
& TECHNOLOGY


INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE