ASPECTOS TÉCNICOS E LEGAIS DA COLETA E ANONIMIZAÇÃO DE TRÁFEGO DE REDES IP

MARCO AURÉLIO VILAÇA DE MELO

ASPECTOS TÉCNICOS E LEGAIS DA COLETA E ANONIMIZAÇÃO DE TRÁFEGO DE REDES IP

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: DORGIVAL OLAVO GUEDES NETO

Belo Horizonte

Setembro de 2009

Melo, Marco Aurélio Vilaça de
M528a Aspectos Técnicos e Legais da Coleta e Anonimização de Tráfego de Redes IP / Marco Aurélio Vilaça de Melo. — Belo Horizonte, 2009
xx, 84 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Dorgival Olavo Guedes Neto

1. Redes de Computação - Protocolos - Tese. 2. Redes de Computação - Direito à Privacidade - Tese. 3. Redes de Computação - Medidas de Segurança - Tese. 4. Anonimização - Tese. 5. Logs - Tese. I. Título.

CDU 519.6*22(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Aspectos técnicos e legais da coleta e anonimização de tráfego de redes IP

MARCO AURÉLIO VILAÇA DE MELO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. DORGIVAL OLAVO GUEDES NETO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. JOSÉ MARCOS SILVA NOGUEIRA
Departamento de Ciência da Computação - UFMG

DR. WILTON SPEZIALI CALDAS
Desenvolvimento e Pesquisa - IVision

Belo Horizonte, 29 de setembro de 2009.

À Deus, responsável por mais essa vitória.

Aos meus filhos, Matheus e Mariana pelo amor incondicional.

À Greisiele, pela dedicação e amor.

À minha mãe, pelo apoio irrestrito.

Ao Prof. Dorgival pela ajuda e paciência.

Ao meu pai, irmãos, familiares e amigos pelas palavras de apoio e amizade.

Resumo

Pesquisadores e administradores de rede encontram-se frente a um dilema ao trabalhar com arquivos de dados de tráfego coletado: como extrair informações úteis para seu trabalho, mas ainda garantir a privacidade dos usuários, cujas informações trafegam pela rede, e evitar o vazamento de informações sensíveis sobre a segurança da mesma?

Este trabalho faz um estudo sobre aspectos de privacidade e segurança no uso e compartilhamento de arquivos de registro de tráfego de rede (*logs* e propõe uma metodologia para análise do processo de anonimização de arquivos.

Inicialmente é explicada a necessidade crescente de se utilizar arquivos de *log* para as pesquisas sobre melhorias na Internet ou auditorias, mostrando em seguida os riscos que o uso e o compartilhamento desses arquivos pode acarretar para a privacidade dos usuários e a segurança da rede. Em seguida, analisamos as leis existentes em alguns países sobre a privacidade de dados e das comunicações eletrônicas, dando uma idéia da sua evolução histórica. No Brasil, são analisadas as leis existentes e alguns projetos e tramitação no congresso nacional, sendo apontadas as implicações legais que o uso desses arquivos pode ocasionar para usuários e administradores de redes.

Finalmente é feita uma análise dos principais protolocos da arquitetura TCP/IP com vistas à anonimização, indentificando quais campos daqueles protocolos podem revelar informações que afetem segurança da rede ou a privacidade dos usuários. Com base nessa informação, é apresentado um estudo das principais técnicas e ferramentas de anonização de dados e, por fim, é feita a especificação de uma metodologia para análise dos arquivos anonimizados que é complementada com a descrição do protótipo da ferramenta baseada nesta metodologia.

Abstract

Researchers and network administrators face a difficult dilemma when they work with traffic data files collected from the network: how to extract useful information for their work and yet to guarantee the privacy of users, whose information travel through the network, and prevent the leakage of sensitive information that may compromize network security?

This work presents a study of aspects of privacy and safety in the use and sharing of network traffic log files, and proposes a methodology for the analysis of the file anonimization process.

First we explain the reasons for the increasing need for the use of log files in network research and audits, showing the risks that the use and sharing of such files may carry for the privacy of users and the safety of the network. Next we discuss the existing laws in some major countries that deal with the privacy of data and electronic communications, showing their evolution over time. In Brazil, we discuss the current laws and some proposed projects being considered in Congress and their implication to users and network providers.

Finally, we analyze the major protocols of the TCP/IP architecture in relation to anonimization, identifying which protocol fields may reveal information sensitive to network safety or user privacy. Based on that analysis we present a discussion of the major tools and techniques for data anonimization and propose a methodology for the analysis of the quality of anonimization, which we complete with the description of a prototype based on that methodology.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Nos últimos anos o mundo presenciou um grande crescimento no uso da Internet, no Brasil, a cada dia aumenta o número de usuários conectados à rede mundial de computadores [1]. Além disso, houve também uma grande diversificação nas aplicações disponíveis através dessa rede. Todo esse crescimento se traduz em tráfego de rede, mensagens que circulam pelos canais da rede. Esse tráfego, além de seu interesse indireto para os usuários, que desejam obter informações da rede, é de grande interesse para duas comunidades ligadas à área de redes de computadores: pesquisadores e administradores de sistema.

Pesquisadores buscam entender o comportamento dos usuários e o impacto das diferentes aplicações sobre a infra-estrutura de rede, a fim de propôr novas soluções que garantam a contínua evolução dos serviços e a escalabilidade dos recursos da rede. Através da análise do padrão de acesso a páginas web, por exemplo, pesquisadores foram capazes de identificar a ocorrência frequente de acessos a páginas populares e propuseram soluções para reduzir a carga na rede usando mecanismos de *caches* [Rabinovich & Spatscheck, 2002]. Pela análise de tráfego, pesquisadores são também capazes de melhor entender o comportamento de novas aplicações como aquelas de compartilhamento de arquivos em redes *peer-to-peer* [Arthur & Panigrahy, 2006] e de identificar o comportamento de disseminadores de mensagens de *spam* por correio eletrônico [Steding-Jessen et al., 2008], podendo assim sugerir técnicas para seu controle.

Assim sendo, as informações obtidas através da monitoração de funções da rede são importantes para a evolução da pesquisa na área de rede [Bianchi et al., 2008b; Burkhart et al., 2008b; Pang & Paxson, 2003]. Nos últimos anos esses dados estão ganhando ainda maior importância; algumas conferências, por exemplo, para aceitação

[1] http://g1.globo.com/Noticias/Tecnologia/0,, MUL1274233-6174,00.html

de artigos, estão exigindo que os dados utilizados na pesquisa sejam disponibilizados para a comunidade científica. Outro elemento que desperta grande interesse nessas informações é a necessidade de se ter grandes massas de dados para testes de novas tecnologias, dando maior credibilidade à pesquisa. Para esses fins pesquisadores estão, cada vez mais, compartilhando os dados coletados por eles entre si.

Administradores de sistemas em rede precisam coletar e armazenar certas informações contidas no tráfego para fins de registro histórico das atividades da rede, para a identificação de comportamentos maliciosos na rede que possam indicar abusos ou ataques à infra-estrutura e serviços sob sua responsabilidade e para fins de auditoria[Bishop et al., 2006]. Certas organizações exigem que se mantenha registros dos momentos de conexão e desconexão de cada usuário do sistema, incluindo-se dados sobre suas atividades enquanto conectados. Para fins de planejamento estratégico, muitas vezes administradores se valem da coleta de tráfego para entender a evolução do uso da sua rede e desenvolver seus planos de expansão. Em outros momentos, a coleta e inspeção de tráfego é uma ferramenta essencial no combate a invasores que tentam acessar máquinas dentro da rede de uma organização, seja para obter informações confidenciais, seja para utilizá-las como intermediárias no lançamento de outros ataques à rede.

Exigências de coleta por parte dos administradores de redes vêm sendo, inclusive, objeto de algumas propostas de legislação em vários países e inclusive no Congresso Nacional Brasileiro [Senado Federal, 2008]. Cada vez mais, elementos de auditoria interna de empresas e até mesmo perícia criminal dependem de dados coletados em máquinas ou no tráfego de redes, de forma semelhante ao que ocorre com relação ao registro de ligações telefônicas.

Pelos motivos apresentados, a análise das informações obtidas através da monitoração do tráfego se torna cada vez mais importante. Esses dados de tráfego são obtidos através da monitoração direta dos canais e interfaces físicas da rede, de onde se pode obter uma cópia de cada pacote de dados que passam por eles em qualquer direção, bem como do registro detalhado da operação de alguns servidores da rede (por exemplo, as requisições feitas a um servidor web ou as mensagens recebidas por um servidor de correio eletrônico).

Apesar de sua importância, a coleta de tráfego tem implicações complexas, por poder incluir inclusive os dados dos usuários que trafegam durante sua interação com servidores e outros usuários da rede. Ao coletar o tráfego de uma rede, pode-se ter acesso ao conteúdo de mensagens de correio enviadas por cada usuário, identificar as páginas da Web visitadas por eles, acompanhar suas atividades em um *site* de comércio eletrônico ou suas interações com outros usuários em um *site* social como o Orkut.

Com exigências como as mencionadas anteriormente para a publicação e troca dos dados de medição de redes se tornando práticas constantes, algumas questões se tornam cada vez mais frequentes: Como disponibilizar essas informações sem prejudicar a segurança da rede? O que fazer para garantir a privacidade dos usuários da rede, quando dados precisam ser distribuídos para fins de pesquisa ou de uma auditoria, por exemplo? É legal/ético o uso desses dados sem nenhuma forma de tratamento para se garantir a privacidade dos usuários? Como devemos proceder para viabilizar a utilização desses dados na pesquisa ou para fins administrativos (ou mesmo legais) sem afetar a segurança e/ou a privacidade dos inocentes envolvidos?

Discussões como essas se tornam cada vez mais frequentes e necessárias, pois as pessoas têm se tornado mais conscientes desses problemas e, por consequência, se tornam mais preocupadas com as suas informações que transitam na rede. Além disso, a área jurídica começa a se preocupar com os impactos que o "mundo virtual" causa nas relações jurídicas. Um dos focos dessa preocupação é o quanto a privacidade é garantida quando se usa esses dados em pesquisas.

Diante desse quadro, várias técnicas e ferramentas para tornar anônimos os dados de rede têm sido propostas tentando garantir um determinado nível de privacidade aos dados distribuídos e, ao mesmo tempo, preservando as principais informações necessárias para a pesquisa e a segurança de rede. Essas ferramentas fazem a chamada anonimização, que é a técnica de excluir as informações considerados sensíveis à privacidade de determinado tipo de dado, proporcionando assim, uma maior tranquilidade e liberdade aos pesquisadores, auditores e investigadores, na utilização e compartilhamento dos dados utilizados por eles.

Tendo isso em mente, torna-se necessário um estudo para entender as implicações legais e analisar as várias técnicas e ferramentas de anonimização de dados existentes, para confirmar se elas satisfazem as exigências de privacidade enquanto mantêm as informações úteis para cada fim. Pang et al. [Pang et al., 2006] enfatizam a necessidade de uma ferramenta que analise os dados anonimizados para verificar se os mesmos estão realmente de acordo com determinada política de anonimização, dando uma maior confiabilidade e segurança ao se disponibilizar dados de redes. É nesse contexto que se insere o trabalho aqui apresentado.

1.1 Motivação

Garantir que o dado anonimizado realmente possui o nível esperado de anonimização é um problema de difícil solução, pois existem diversas questões sobre a anonimização

que despertam opiniões conflitantes, tanto na área jurídica, quanto na área técnica. Por exemplo, a divulgação do tipo e versão do sistema operacional de uma determinada máquina é considerada um risco por alguns administradores, enquanto não o é por outros. Do ponto de vista jurídico, em certos casos a divulgação das páginas acessadas a partir de determinada máquina não causa nenhum constrangimento, enquanto em outros fere gravemente a privacidade.

Considere-se por exemplo, um administrador que é abordado por um pesquisador que deseja uma amostra de tráfego da rede a fim de avaliar uma hipótese de pesquisa. Ou ainda, imaginem um diretor de uma universidade que procura o administrador da rede para discutir sobre a possibilidade da universidade firmar um convênio com um grupo de universidades, para passar a disponibilizar os dados de conexão de rede da universidade para toda a essa comunidade científica e, em troca, receber todos os dados dessa comunidade. Nesse caso, para garantir a confidencialidade e segurança da rede, o administrador deverá não só usar uma determinada ferramenta de anonimização de dados, mas deverá também usar uma política de anonimização pré-definida para que os dados tivessem o mesmo padrão e nível de qualidade dos dados disponibilizados pelas outras instituições.

O administrador pode até ter interesse no tipo de resultado da pesquisa, ou na possibilidade de ter acesso aos dados de outras universidades conveniadas para as pesquisas que elas desenvolvem, mas não deveria fornecer os dados se não tivesse garantias de que a privacidade dos seus usuários não seria violada em relação ao que exige a lei. Para esse fim, é importante que o administrador saiba quais são as informações sensíveis do ponto de vista da privacidade/segurança e as exigências e restrições legais envolvidas, bem como entenda o que oferecem as diversas ferramentas e técnicas de anonimização existentes.

Diante de situações similares a essa, torna-se necessária uma metodologia que valide determinada anonimização segundo um certo critério, por exemplo, garantir que não haja alguma forma de inferir que os endereços anonimizados das máquinas sejam mapeados para determinados endereços IP reais. Ou ainda, confirmar que determinado dado anonimizado manteve as mesmas características (por exemplo, distribuição estatística) que constam nos dados originais. Ou também, se é possível determinar o sistema operacional de um servidor específico a partir dos dados anonimizados.

1.2 Objetivos

Com base no exposto até aqui, o objetivo principal desta dissertação é oferecer elementos que auxiliem os administradores de sistemas em rede a decidir sobre a liberação de informações sobre tráfego de rede considerando aspectos de privacidade e anonimato de seus usuários. De forma mais detalhada, este trabalho tem os seguintes objetivos específicos:

- analisar a legislação existente sobre privacidade de dados em alguns países. No Brasil, serão analisadas as leis existentes bem como os principais projetos de lei que tramitam no nosso legislativo.
- identificar os principais elementos de informação contidos no tráfego de rede da Internet e discutir o impacto desses sobre a privacidade dos usuários e segurança de uma rede;
- identificar as principais ferramentas e técnicas de anonimização de dados de conexão de rede, confrontando-as com os vários tipos de ataques a estas técnicas.
- propor uma metodologia que avalie o grau de anonimato de uma técnica de anonimização que precise ser avaliada por um administrador de rede.

1.3 Contribuição

As principais contribuições deste trabalho endereçam diretamente os objetivos específicos mencionados.

- O capítulo 3 apresenta uma discussão dos aspectos legais segundo a legislação brasileira específica, ainda incipiente na área de comunicação de dados, e com base no material legal já desenvolvido para outros meios de comunicação que pode ser co-relacinado com a área de dados de rede. Esse capítulo também discute elementos da legislação dos EUA e da União Européia, mais desenvolvida nessa área.
- O capítulo 4 apresenta uma análise detalhada dos principais protocolos da arquitetura da Internet (a arquitetura TCP/IP), discutindo a informação normalmente disponível em cada campo desses protocolos e sua implicação para a obtenção de dados que possam afetar o anonimato/privacidade dos usuários e a segurança da rede, bem como uma descrição das principais técnicas e ferramentas de anonimização disponíveis, suas qualidades e limitações.

- O capítulo 5 apresenta a metodologia e um protótipo da ferramenta de verificação proposta, que permitirá aos administradores de sistemas em rede analisar o efeito da aplicação de uma certa técnica ou ferramenta de anonimização externa sobre um arquivo de tráfego coletado. Com base nessa análise seria possível avaliar se as restrições de anonimato que devem ser observadas pelo administrador estão sendo atendidas no processo de geração do arquivo de dados a ser disponibilizado.

1.4 Organização do restante do texto

Nos capítulos seguintes, apresentamos primeiramente, no capítulo 2, os principais conceitos relacionados ao anonimato, formas de reverter a anonimização de dados, aspectos legais relacionados a coleta e análise de dados de rede e conceitos técnicos sobre a arquitetura TCP/IP, coleta e anonimização de tráfego. Em seguida, os capítulos 3, 4 e 5 apresentam as contribuições já mencionadas. Finalmente, o capítulo 6 apresenta as conclusões da dissertação e sugestões para trabalhos futuros.

Capítulo 2

Conceitos e Trabalhos Relacionados

A fim de compreendermos melhor os diversos aspectos relacionados à anonimização de arquivos de tráfego de rede e desenvolvermos as contribuições deste trabalho é importante discutirmos os conceitos gerais de privacidade, anonimização, aspectos de segurança de rede, as características da arquitetura TCP/IP, utilizada na Internet atual e que determina qual informação acompanha cada pacote de dados na rede, algumas das principais ferramentas de anonimização de tráfego existentes e outros trabalhos relacionados que mereçam destaque. Para esse fim, as seções seguintes discutem cada um desses tópicos em mais detalhes.

2.1 Privacidade

A privacidade é um termo subjetivo e por isso de difícil definição, pois o seu conceito e amplitude variam de pessoa para pessoa; por exemplo, ter seu nome impresso em uma lista telefônica pode representar uma invasão de privacidade para um cantor famoso, que não gostaria de ter seu nome e telefone divulgados a todos. Por outro lado, um prestador de serviços autônomo provavelmente irá considerar essa divulgação benéfica para os seus negócios. Por causa desses diferentes sentimentos quanto à privacidade, os autores divergem entre conceitos amplos e restritos.

Warren & Brandeis [1890] diz que "a privacidade é um direito de estar só", conceituando o tópico de forma simples e restrita. Já a dimensão desse conceito é ampliado por José Afonso da Silva [Silva, 1997], ao dizer que a nossa Constituição assegura "direito à indenização por dano material ou moral decorrente da violação da intimidade, da vida privada, da honra e da imagem das pessoas, em suma, do direito à privacidade". Nos países democráticos o direito à privacidade é considerado um direito fundamental e é protegido por lei.

Com a evolução das tecnologia esse direito passa a ficar mais fragilizado, pois a cada dia cresce o número de câmeras de segurança, de empresas com cadastros informatizados de clientes, etc. A partir disso, o conceito de privacidade começa a englobar também os dados, surgindo em seguida as legislações para proteção desses dados.

Para melhor entender a legislação sobre privacidade dos dados é interessante fazer uma classificação desses dados em dados cadastrais, dados necessários para estabelecer uma conexão e dados de conteúdo de tráfego. Apesar de cada um desses tipos ter utilidade díspares todos, *a priori*, contêm informações privadas. Dados casdastrais são os dados encontrados nos vários bancos de dados existem nas empresas. Os dados necessários para estabelecer uma conexão são as informações usadas para controlar a conexão de um cliente a uma página `web` de um banco, por exemplo. Por último, os dados de contéudo de tráfego contêm a informação de interesse do usuário durante cada interação do mesmo com os sistemas em rede.

2.2 Anonimização

No dicionário Aurélio [Ferreira, 2008] a definição de anonimato é "sem o nome ou assinatura do autor; sem nome ou nomeada; obscuro". Portanto, podemos dizer que no contexto da informatização dos dados a informação anônima é aquela que não seja possível identificar a quem ela se refere.

A nossa Constituição, no inciso IV do artigo 5^o, diz que "é livre a manifestação do pensamento, sendo vedado o anonimato" [Congresso Nacional, 1988], entendendo então que é vedada a não identificação do autor. Apesar de a privacidade ser protegida em nosso texto constitucional, o anonimato não o é [Pinheiro, 2008], permitindo que os dados de cadastros e de conexões possam ser levantados através dos meios legais.

Diante disso, a anonimização de dados de tráfego de rede é o processo de retirar as informações que possam levar à identificação dos usuários da conexão. Mais abrangentemente, essa anonimização engloba também o conteúdo da informação trocada e também as informações que interferem na segurança da rede de origem e destino dos dados.

2.3 Ataques Utilizando *Logs*

O problema de coleta de dados se resume nos limites da lei. Ou seja, a lei define se um determinado tipo de dado pode ser coletado ou não. Além disso, mesmo com a

permissão legal para coleta, a legislação deverá especificar se o dado coletado pode ser compartilhado e a forma para isso ocorrer.

No caso das pesquisas, caso haja um consentimento legal para utilização desses dados, isso deve ocorrer somente com o uso de anonimização, ou seja, a tendência legal é admitir o uso apenas com informações que não levem à identificação do usuário e seus dados privados.

Levando em consideração que os dados compartilhados para a pesquisa sejam anonimizados, surge a necessidade de garantir que esses dados de rede divulgados não serão passíveis de quebra do anonimato. Além disso, é preciso considerar os possíveis ataques que afetam não só o anonimato, mas também a segurança da rede/sistema, pois a falta de segurança de uma rede pode implicar na violação da privacidade dos seus usuários.

Um ataque comum com relação à segurança de uma rede é o que tenta identificar o sistema operacional que gerou um certo tipo de dado; para isso foram desenvolvidas as ferramentas baseadas na técnica de *passive OS Fingerprint* [Nmap, 2009; Spangler, 2003]. A principal atuação dessas ferramentas é verificar em determinados campos dos cabeçalhos da pilha TCP/IP, o tipo de informação que eles contêm. Isso se justifica porque nem sempre os desenvolvedores dos sistemas operacionais seguem as definições e padronização completamente. Ou seja, o padrão indica que determinado campo deve conter um valor padrão, mas muitos sistemas colocam valores diferentes. Dessa forma, as ferramentas de identificação do sistema operacional comparam o valor do campo com o valor padrão de cada sistema operacional. Caso os valores sejam iguais, deduz-se qual o sistema operacional que originou aquele pacote. Existem diversos campos em vários níveis da arquitetura TCP/IP que podem ser utilizados por esse tipo de ferramenta, como veremos com maiores detalhes esses campos na seção 4.1.

Outro tipo de ataque, analisado por Kohno et al. [2005], é a técnica de identificar determinada máquina através de um padrão de tempo de envio de pacotes, onde segundo os autores cada equipamento possui um padrão de intervalo entre o envio dos pacotes, esse padrão se torna uma "assinatura" ou impressão digital das máquinas.

Os ataques anteriores levam em consideração arquivos de *logs* anonimizados disponibilizados pelas empresas. Um outro tipo de ataque é o chamado ataque de injeção de *logs* [Gattani & Daniels, 2008; King et al., 2009; Ribeiro et al., 2008], onde o adversário sabe que determinada empresa disponibiliza periodicamente, para a comunidade, seus arquivos de *logs* anonimizados. Então, o adversário tenta inserir, nos arquivos que futuramente serão disponibilizados, informações que o ajude a identificar que determidados dados, mesmo após a sua anonimização, foram gerados por ele.

Essa inserção de dados pode ser feita através de uma sequência de requisições

ICMP ou através da inclusão de uma informação em campos não utilizados pela arquitetura TCP/IP, por exemplo, o campo RESERVADO do protocolo TCP. No futuro, quando os arquivos forem disponibilizados para a comunidade, o adversário localiza nesses dados o tráfego gerado por ele. Dessa forma, ele saberá o conteúdo do dado original e o padrão de sua anonimização, facilitando o trabalho de identificar os valores originais do restante dos dados anonimizados.

Existe ainda um outro tipo de ataque a arquivos de *logs* descrito por Coull et al. [2007], que é possível inferir a topologia da rede e até identificar determinados computadores/usuários através da análise de comportamento do tráfego, por exemplo, tipo de tráfego específico ou horário de conexão determinado, segundo os autores a anonimização não é eficaz contra este tipo de ataque.

2.4 Arquitetura TCP/IP

Atualmente a Internet é formada por milhares de pequenas e grandes redes de computadores interligadas uma às outras. Essas redes são formadas pelos mais variados tipos de *hardware*, sistemas operacionais, aplicativos e tecnologias de roteamento, configurando umas das principais características da Internet, que é a sua heterogeneidade.

Para que máquinas conectadas em diferentes pontos dessa variedade de tecnologias conseguisse se comunicar, foi necessário criar padrões de comunicação que permitissem a troca de informação entre as diferentes redes.

A arquitetura TCP/IP, desenvolvida a partir do projeto da ARPANET no início da década de 1970, se tornou o padrão *de facto* que permitiu essa troca de informações. Também chamada de "pilha TCP/IP", essa arquitetura é baseada na comutação de pacotes e é formada por um conjunto de quatro camadas de protocolos (aplicação, transporte, rede (ou inter-rede) e tecnologia de rede local), onde cada uma das camadas possui vários protocolos que têm a função de resolver determinados problemas envolvidos na comunicação, por exemplo, a identificação do serviço que está sendo utilizado, a identificação do destinatário, etc.

Ao se comunicarem usando a arquitetura TCP/IP, as máquinas dividem a informação em vários pacotes de dados que devem ser transferidos pela rede até seu destino. Nesse processo de divisão da informação para o envio, cada pacote de dados da camada de aplicação é repassado para cada uma das camadas inferiores da pilha. Essas camadas adicionam ao pacote informações de controle para que as camadas equivalentes da arquitetura no destinatário entendam como a informação deve ser traduzida. Essas informações constituem os cabeçalhos de cada camada da arquitetura e determinam

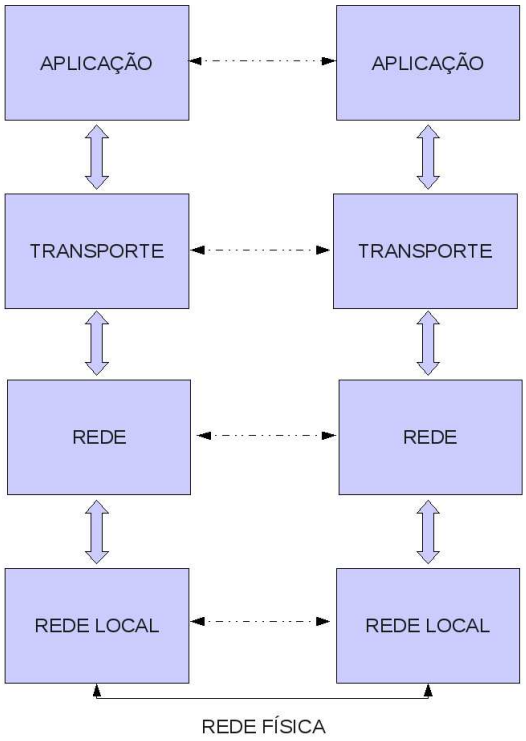


Figura 2.1. Pilha de Protocolos TCP/IP

como cada pacote deve ser processado ao longo do caminho.

No destinatário, cada pacote recebido é processado a partir da camada de tecnologia de rede local. Cada camada retira o respectivo cabeçalho do início do pacote e utiliza a informação ali contida para decidir como processar os dados do pacote. Normalmente isso implica na entrega do pacote a um protocolo da camada superior, que por sua vez retira seu cabeçalho e repete o processo, até que os dados sejam entregues à aplicação. Essa técnica de inclusão/retirada de cabeçalhos na mensagem pelos protocolos de cada camada é chamada de encapsulamento, onde o pacote que sai de uma camada, incluindo seu cabeçalho, é entendido pela camada abaixo como sendo a mensagem de dados. A figura 2.2 mostra como funciona o encapsulamento de dados pelas camadas da arquitetura TCP/IP.

A seguir detalhamos a função de cada uma das camadas da arquitetura TCP/IP.

- **Aplicação**

  A camada aplicação é a camada onde se localizam os programas dos usuários, os quais implementam diferentes serviços. Essa camada recebe as solicitações daqueles usuários e as transformam em mensagens para outras aplicações em ou-

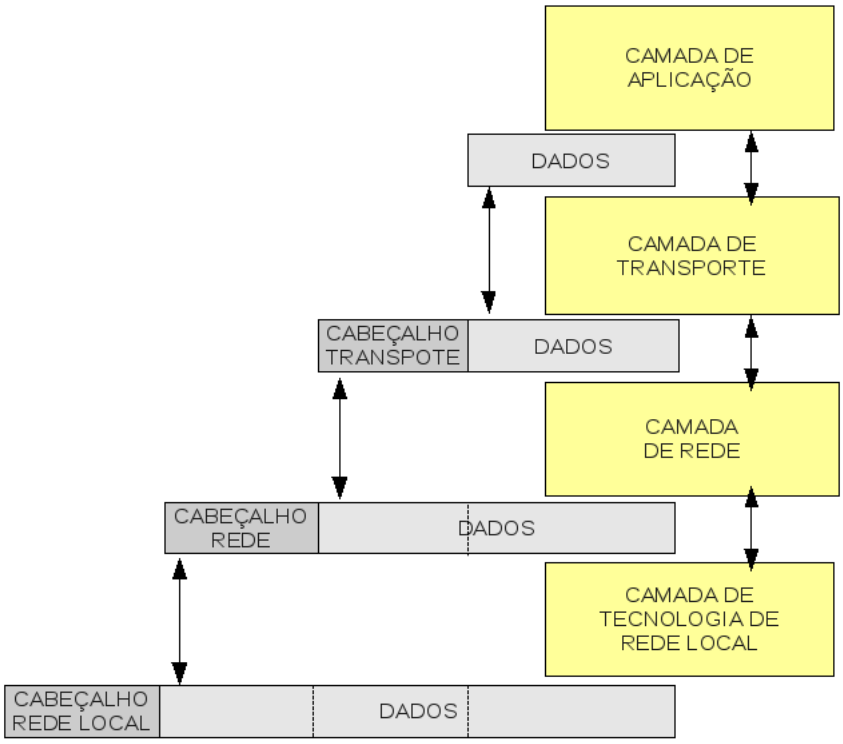


Figura 2.2. Encapsulamento de dados na arquitetura TCP/IP

tros pontos da rede. Essas mensagens precisam ser repassadas para a camada de transporte, para que sejam entregues aos programas de destino. Exemplos de protocolos dessa camada são HTTP, SMTP, SSH e DNS, que tratam, respectivamente, de requisições das aplicações de web, do envio de mensagens de correio eletrônico, do acesso a computadores remotos e da resolução de nomes na Internet.

- **Transporte**

  A camada seguinte é denominada transporte. Ela é responsável por receber os dados da camada de aplicação e garantir que eles sejam entregues à máquina destino. Nessa camada os protocolos existentes são o *User Datagram Protocol* (UDP) e o *Transmission Control Protocol* (TCP), que oferecem serviços de entrega diferentes: o primeiro oferece um serviço baseado em mensagens independentes, sem garantias de entrega, enquanto o segundo oferece um canal de comunicação de *bytes*, que são entregues garantidamente em ordem e sem perdas (desde que não haja uma interrupção da rede subjacente).

- **Rede**

A camada de rede tem como principal protocolo o *Internet Protocol* (IP), que acrescenta aos pacotes da camada de transporte informações como endereços de origem e destino e garante que esses pacotes sejam roteados através de uma rede local a outra, até que eles atinjam seu destino. Esse processo, entretanto, é feito no modelo denominado "melhor esforço" (*best effort*), onde nenhuma garantia é feita sobre a entrega final dos dados (daí a importância do TCP, que deve corrigir quaisquer falhas ocorridas na comunicação por IP).

Nesse sentido, as principais atribuições dessa camada são prover um padrão de identificação de máquinas na rede que seja válido para toda a Internet e fornecer uma forma de garantir o encaminhamento correto dos pacotes entre a máquina de origem e a máquina destino (roteamento). Esse padrão de identificação é o que se denomina endereço IP, ele possui quatro *bytes* e tem como função identificar unicamente uma máquina na Internet; ele também tem como função identificar a rede em que se encontra determinada máquina.

Para garantir essas funcionalidades encontramos também nessa camada, além do IP, o protocolo **Internet Control Message Protocol** (ICMP) e os protocolos de roteamento, o primeiro tem como função principal permitir que os elementos da rede se comuniquem para troca de mensagens de erro ou de controle que porventura sejam necessárias durante a comunicação. Já os protocolos de roteamento, são responsáveis por permitir que os caminhos entre as diversas origens e destinos possíveis sejam conhecidos ao longo da rede. Exemplos de protocolos de roteamento são RIP, OSPF e BGP.

Normalmente se inclui nesta camada o protocolo ARP (*Address Resolution Protocol*), usado pelas máquinas para transformar os endereços IP em endereços reconhecidos pela tecnologia de rede local existente em cada caso.

- **Tecnologia de Rede Local**

  A camada inferior, na concepção original da arquitetura TCP/IP, é denominada de tecnologia de rede local e tem como responsabilidade receber os pacotes da camada de rede e os converter em quadros que em seguida são transformados em sinais elétricos e transmitidos pela rede física até uma outra máquina da rede local, esta máquina poderá ser o destinatário final da conexão, ou pode ser um roteador (*gateway*) que através do protocolo IP identificará o próximo canal/rede por onde aquele pacote deverá ser roteado em seguida. Cada tecnologia de rede local pode ter sua forma interna de identificar cada máquina a ela conectada, daí a imporância do protocolo ARP, mencionado anteriormente, já que cada rede

local não necessariamente tem ciência dos endereços definidos pela camada de rede (IP).

Exemplos de tecnologias de rede local são as diveras variedades de redes Ethernet, as redes sem fio conhecidas como WiFi e WiMax, e tecnologias para canais ponto-a-ponto (usualmente linhas discadas) como o protocolo PPP.

2.5 Processo de Coleta e Análise de Dados

A fim de se obter informações sobre o comportamento dos usuários e as demandas sobre a infra-estrutura de rede, diversos tipos de dados podem ser de interesse durante o trabalho de monitoração e análise de *logs*.

2.5.1 Tipos de coleta

Para cada tipo de dado, um tipo de coleta específico pode ser necessário, ao se focar em um tipo de aplicação ou serviço específico, é como que administradores e pesquisadores se valham de registros de atividade (*logs*) gerados pelos programas servidores que implementam determinados serviços. Esse é o caso, por exemplo, quando se estuda a carga de um servidor Web através do *log* das requisições atendidas por ele. No extremo oposto do espectro de coleta de dados encontra-se a coleta de tráfego bruto que circula pela rede, onde todo o conteúdo de qualquer comunicação que atravessa um canal pode ser monitorado e coletado. A coleta de análise de *logs* de aplicações é preferida quando o objetivo é analisar um serviço específico. Nesse caso, os registros já são por natureza mais processados, pois pode-se resumir a informação a ser coletada com base no entendimento da semântica da aplicação. Entretanto, para fazê-lo, é normalmente necessário realizar a coleta nas extremidades da rede, seja na máquina do usuário ou no servidor da aplicação, já que são os únicos pontos que possuem conhecimento suficiente para interpretar as requisições do usuário e as respostas do servidor. Esse tipo de análise permite se obter um conhecimento aprofundado sobre um certo serviço, mas não permite uma visão abrangente sobre a rede como um todo ou sobre a interação entre diferentes serviços.

Do ponto de vista de privacidade e anonimato o fato da informação ser derivada com base na semântica de cada serviço torna o problema de se verificar o anonimato em qualquer *log* desse tipo um problema diferente para cada tipo se serviço ou formato de *log*. As questões de anonimato que surgem em um serviço de correio são de natureza diferente daquelas de um servidor `Web`, por exemplo. Dessa forma, trabalhos nesse nível devem focar em serviços específicos.

Já a coleta de tráfego bruto de rede permite que se obtenha uma visão global de toda comunicação que utiliza um certo elemento da rede (um canal, roteador ou chave/*switch*). Esse tipo de coleta exige que o interessado tenha acesso direto ao elemento da rede onde se pretende observar o tráfego, o que normalmente implica na participação do administrador da rede em questão. O problema desse tipo de coleta é o grande volume de dados que pode ser gerado, pois em última instância pode-se optar por coletar cada *byte* trafegado. Esse volume também implica em um maior trabalho na análise dos dados coletados. Por ser uma coleta bruta, é em princípio possível derivar quase toda informação sobre cada aplicação, pelo menos até o ponto em que essa informação tenha relação com os *bytes* trafegados. Isso se deve ao fato de que todos os dados de cada aplicação podem, em princípio, ser incluídos na coleta. Além disso, os cabeçalhos dos diversos protocolos trazem diversas informações que podem servir para se identificar a máquina de origem/destino da comunicação e até mesmo o usuário envolvido.

Uma solução intermediária em relação ao tipo de dado coletado, que é utilizada para análises onde o objetivo não vai além do entendimento dos padrões de tráfego (volumes, origens e destinos), sem preocupação com a semântica dos serviços, é a coleta de dados sobre fluxos (*flows*). Esse tipo de informação é comumente disponível em roteadores através do protocolo NetFlow [Netflow, 2009] e informa apenas o volume de dados trafegados entre cada par origem/destino observado através de um canal ou roteador por unidade de tempo. Esse tipo de dado possui basicamente apenas o endereço IP de origem e destino como informação que pode afetar a privacidade do usuário e/ou a segurança da rede. Dessa forma, questões de anonimato nesse caso se limitam a esses endereços; dessa forma, anonimização do tráfego `netflow` é apenas um sub-conjunto das questões associadas ao tráfego bruto.

Este trabalho tem como foco o estudo do problema de anonimização de registros de tráfego bruto, por ser um problema abrangente e independente de aplicações específicas. Além disso, muito do que se discute aqui sobre anonimização de endereços de rede que se aplica diretamente ao problema de anonimização de coletas de fluxos, como explicado anteriormente.

Tráfego bruto de rede compreende todo o conteúdo de cada pacote que trafega pela rede. Esse tipo de dado pode ser obtido nos elementos de conexão e roteamento, como roteadores ou *switches*. Dele podem ser obtidas informações sobre origem e destino dos dados, tipo de serviço que está sendo usado, horário da conexão e até mesmo o conteúdo da comunicação, como por exemplo, identificação de usuário, senha e número de cartão de crédito em uma interação com um servidor de comércio eletrônico e também todo o conteúdo de uma mensagem de correio eletrônico.

Os dados referentes à conexão geralmente são coletados pelos administradores para observar o uso da rede, identificar possíveis ataques, identificar a origem de cada tipo de tráfego, ou qualquer tipo de informação que ajude na manutenção e bom funcionamento da rede. Já os pesquisadores podem usar esses dados para caracterizar tráfego e analisar o comportamento da rede após a disponibilização de um novo serviço, por exemplo.

2.5.2 O processo de coleta

O processo de coleta pode ser dividido em três partes principais: a obtenção de um acesso direto ao tráfego a ser coletado, a coleta propriamente dita e sua análise posterior.

2.5.2.1 Acesso aos dados do tráfego

O primeiro passo para se realizar a coleta de dados brutos é encontrar uma forma de se ter acesso ao conteúdo de todos os pacotes que passam por um canal de interesse. Se esse canal de interesse é apenas um canal que leva a uma máquina específica, como um servidor, basta se ter acesso àquela máquina para se realizar a coleta. Por outro lado, quando se deseja coletar/analisar todo o tráfego de entrada e saída de uma rede, é necessário ter acesso ao canal que conecta essa rede ao restante da Internet. Nesse caso, é comum que haja apenas roteadores ou chaves Ethernet (*switches*) nas extremidades do canal, onde normalmente não é possível se realizar diretamente uma coleta (já que normalmente precisa-se de um equipamento especialmente configurado para esse fim).

Nesse caso, há normalmente duas formas de se resolver esse problema, dependendo dos recursos de *hardware* disponíveis: interceptação ou espelhamento do tráfego. A figura 2.3 ilustra as duas opções.

No caso da interceptação, um computador com duas interfaces de rede deve ser colocado no meio do fluxo de dados, usando-se cada uma das suas interfaces para se conectar a um dos dois extremos do canal original que se deseja monitorar. O sistema operacional daquele computador deve ser configurado para copiar todos os pacotes que cheguem em uma interface para a outra, garantindo que o fluxo de pacotes no canal seja mantido inalterado. Paralelamente, o sistema deve copiar cada pacote recebido para um arquivo de armazenamento local, que constituirá o arquivo de registro de tráfego.

Já no caso do espelhamento, é essencial que se tenha um elemento de rede (roteador ou chave Ethernet) com essa funcionalidade. Nesse caso, o elemento de rede pode ser programado para realizar uma cópia de cada pacote recebido ou enviado através de uma certa interface de rede (a interface de terminação do canal de interesse). Essa

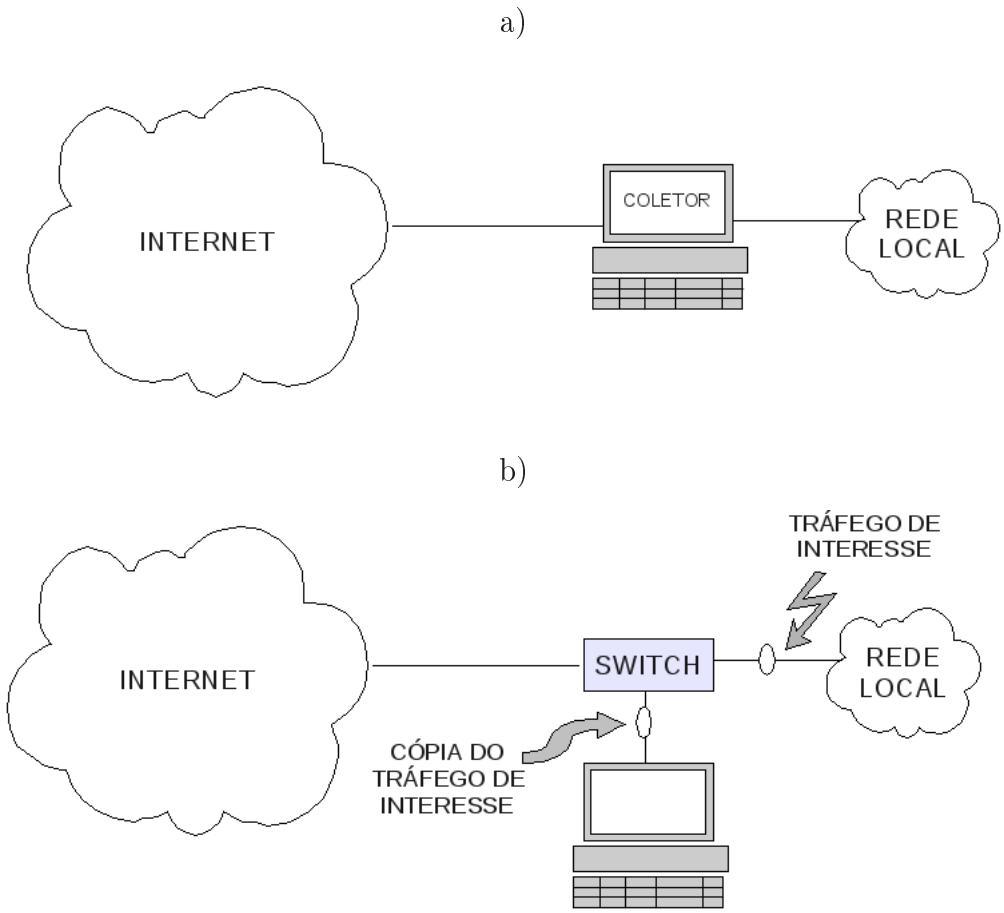


Figura 2.3. Formas de coleta de tráfego: a) interceptação; b) espelhamento

cópia é então transmitida por uma outra interface do mesmo elemento de rede, à qual pode-se então conectar o computador de coleta. Esse computador precisa apenas armazenar cada pacote que recebe através daquela interface, sem nenhum outro tratamento extra.

A interceptação exige normalmente um computador com mais recursos (o equipameno precisa ter duas interfaces de rede, ser configurado para copiar o tráfego recebido em cada interface para outra, agindo como uma *bridge*, e com desempenho suficiente para fazer a cópia e o armazenamento do tráfego sem perdas). Entretanto pode ser mais facilmente colocada em prática, pois não impõe maiores exigências sobre a rede a ser monitorada. Desde que o canal monitorado seja da mesma tecnologia das duas interfaces de rede do computador de monitoração ela pode ser implantada. Já o espelhamento reduz a demanda sobre o computador de coleta, que precisa ser capaz apenas de copiar os dados recebido para um arquivo, mas depende da existência de um elemento de rede no ponto da coleta que possua recursos de espelhamento de tráfego.

2.5.2.2 Coleta de tráfego

Independente da técnica adotada para se ter acesso ao tráfego, o próximo elemento necessário é o programa de coleta propriamente dito. Nesse caso, a ferramenta mais comum para obter esses dados, dentre outros aplicativos existentes, é o programa `tcpdump` [Tcpdump & libpcap, 2009] através da `libpcap` que é uma biblioteca para processamento dos *logs*, estes emprestam seus nomes para os arquivos gerados por eles.

A coleta de tráfego não é um procedimento automático dos elementos de rede: ela deve ser configurada pelo administrador do sistema e a partir daí, como mencionado anteriormente, pode-se obter e armazenar todo o conteúdo dos dados que trafegam pela rede nesse tipo de monitoração. Entretanto, geralmente o que é coletado e analisado são apenas os primeiros *bytes* de cada pacote trafegado, já que neles encontram-se os cabeçalhos dos protocolos, de onde se pode obter a maior parte da informação de interesse para análise.

O `tcpdump` é uma ferramenta que é executada através da linha de comando e consegue ler tanto os dados diretamente da interface de rede, quanto de um arquivo de coleta gerado anteriormente. Ele pode gerar um arquivo de saída no formato texto ou no formato do próprio programa. Ele usa a biblioteca `pcap` (*packet capture*), que proporciona um ambiente de alto nível para captura e processamento de pacotes de rede.

O `tcpdump` tem como padrão, na maioria dos sistemas operacionais, ler apenas os primeiros 68 *bytes* dos pacotes que trafegam na rede. Esses 68 *bytes* normalmente são suficientes para se obter toda a estrutura de cabeçalhos ICMP, IP, TCP e UDP. Entretanto, como o tamanho dos pacotes desses protocolos pode variar, é possível que dentro dos 68 *bytes* salvo exista uma quantidade de *bytes* do chamado *payload* (que são os dados da aplicação propriamente dita). Além disso, o programa pode ser configurado para reter todo o pacote, aumentando consideravelmente a quantidade de dados armazenados e consequentemente afetando o desempenho do sistema de leitura e gravação dos pacotes. O armazenamento de *bytes* de *payload* é sempre uma questão delicada, devido à variedade de aplicações é praticamente impossível se criar uma forma de anonimizar esses dados.

Uma característica muito importante do `tcpdump` é que ele permite especificar filtros para ele coletar apenas determinado tipo de informação. Por exemplo, ele pode ser configurado para coletar apenas o tráfego de determinado endereço IP de origem, ou todos os pacotes que forem do protocolo TCP, ou até mesmo excluir os pacotes que sejam endereçados para a porta 80 (geralmente tráfego web).

2.5.2.3 Mecanismos de análise

Uma vez de posse de uma cópia do tráfego em um canal, diversas ferramentas podem ser utilizadas para se analisar esse tráfego. Muitas delas, como por exemplo, o `tcpstat` [1] e o `tcpflow`[2] apenas geram informações estatísticas agregadas, não se constituindo, a princípio, em ameaça à privacidade dos usuários. Entretanto, outros programas permitem que se obtenha um grande volume de informações sobre os usuários e suas comunicações.

A primeira ferramenta nessa linha é sem dúvida o próprio `tcpdump`, que pode ser usado para gerar relatórios textuais com informações extraídas de cada pacote. Além dele, entretanto, diversos outros programas podem ser usados. Duas ferramentas que merecem destaque nesse caso são o `ntop` e o `wireshark`.

O `ntop` [Ntop, 2009] é um analisador desenvolvido para atuar em tempo real (normalmente no mesmo ponto onde se usaria o `tcpdump` para coleta) que gera diversos relatórios sobre o tráfego observado. Ele pode, entretanto, ser utilizado também para obter relatórios sobre tráfego previamente coletado. Os relatórios do `ntop` são de forma geral estatísticos; entretanto, os dados podem ser divididas por endereços de origem/destino, tipo de máquina e outros elementos que podem afetar a privacidade dos usuários. A figura 2.4 apresenta exemplos do tipo de informação disponível através da sua interface.

Já o `wireshark` [Wireshark, 2009] (previamente chamado `ethereal`) é um programa de inspeção de pacotes com interface gráfica. Com ele é possível se inspecionar cada bit em um pacote, sendo que a interpretação dos campos dos cabeçalhos da maioria dos protocolos existentes já é feita automaticamente pela aplicação. Com essa ferramenta é possível também, por exemplo, reconstruir toda uma comunicação entre duas partes na rede a partir dos pacotes individuais, o que pode ter sérios impactos em questões de privacidade. A figura 2.5 mostra três janelas de análise. Na primeira são mostrados todos os pacotes do arquivo `tcpdump`, na qual o pacote número 34 foi selecionado. Na segunda janela é possível visualizar detalhadamente os campos das camadas do TCP/IP, no caso, são mostrados os valores dos campos do protocolo IP. Finalmente, na terceira janela é mostrado o conteúdo do pacote em hexadecimal.

Além desses programas, diversos outros existem com funcionalidades semelhantes ou complementares. Além disso, diversas bibliotecas existem para linguagens como `C`, `Java`, `Python`, `Perl` e outras, que simplificam o desenvolvimento de programas que interpretam o tráfego de rede em busca de informações específicas. Esse recurso será

[1] http://www.frenchfries.net/paul/tcpstat/
[2] http://www.circlemud.org/ jelson/software/tcpflow/

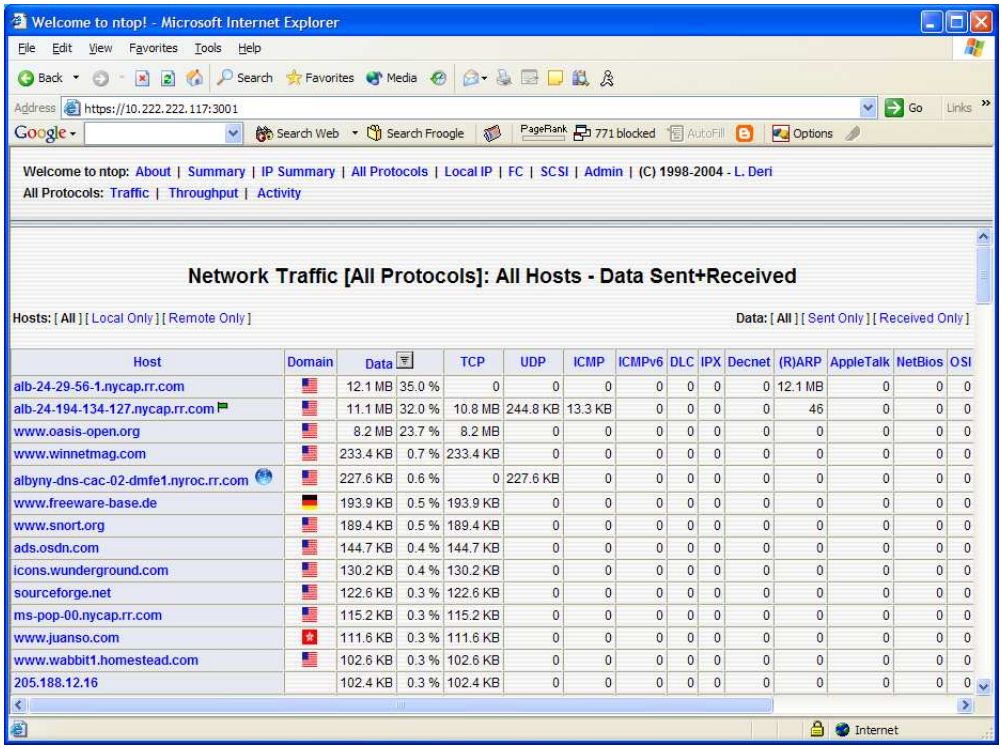


Figura 2.4. Exemplo de informação mostrada pelo `ntop`

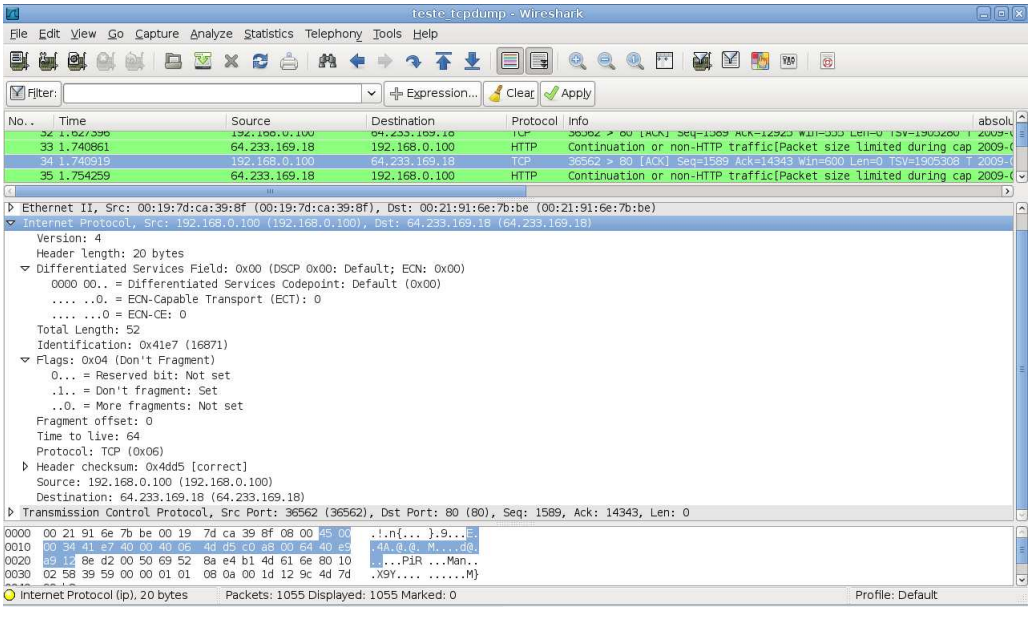


Figura 2.5. Exemplo de informação mostrada pelo `wireshark`

inclusive explorado no capítulo **??**, no desenvolvimento da ferramenta de análise de anonimização

2.5.3 Discussão

Uma vez que se consiga um computador com acesso ao tráfego utilizando uma das técnicas anteriores, o administrador/pesquisador precisa explicitamente executar o programa tcpdump para coletar os dados, ou seja, até que isso ocorra não há qualquer problema de privacidade envolvido. As dúvidas surgem a partir do momento em que uma cópia desses dados começa a ser armazenada. Por exemplo, apenas coletar estes dados sem o aviso prévio do usuário da rede já caracteriza uma invasão de privacidade? Afinal, os pacotes armazenados por padrão do `tcpdump`, como visto, podem conter informações pessoais. Por outro lado, o administrador não manipulou ou compartilhou esses dados. E se os dados coletados contiverem apenas com os cabeçalhos do pacote, isso configuraria também uma quebra de privacidade? O endereço IP é uma informação pessoal?

Avisar previamente o usuário sobre coleta é suficiente para o uso de seus dados não configurar invasão de privacidade? O pesquisador poderá compartilhar esses dados ou guardá-los por tempo indeterminado? Poderá fazer qualquer tipo de análise nesses dados? Essas são questões que precisam ser consideradas.

Como vimos, o tcpdump permite ainda filtrar as informações a serem coletadas, ou seja, é possível fazer uma monitoração direcionada para identificar quais computadores acessam determinados *sites* ou qual o perfil de acesso de determinados usuários. Este tipo de monitoração é legal? Existe algum tipo de procedimento especial que torne esta coleta legal ou ilegal? Ao filtrar apenas um tipo de tráfego aumenta-se a garantia de privacidade do que é coletado? Podendo filtrar um tipo de tráfego, a coleta pode ser direcionada e tornar a privacidade mais ameaçada?

A anonimização de dados torna a coleta legal? Pois a princípio nem o *payload* e nem as informações de indificação da máquina foram mantidas. Ou apenas é legal a coleta executada com permissão judicial?

Achar respostas concretas para essas questões é um tarefa complexa, esperamos ao final desse trabalho discutir as principais dificuldades que envolvem esse assunto.

2.6 Ferramentas de anonimização

Diante da necessidade de se manter a privacidade dos dados e a segurança das redes, surgiram as chamadas ferramentas de anonimização de dados, que definem um conjunto de políticas e técnicas para tentar garantir a privacidade dos usuários de redes e outros serviços, sem, no entanto, afetar a qualidade das informações necessárias para o desenvolvimento de pesquisa, auditorias e análises gerenciais.

Existem diversos tipos de ferramentas e métodos de anonimização, cada uma delas usando abordagens diferentes. Algumas fazem anonimização em um nível específico da pilha de protocolos, outras em informações restritas como por exemplo, as URLs e nomes dos arquivos [Kuenning & Miller, 2003], mas a maioria tenta anonimizar campos em todos os níveis da arquitetura TCP/IP.

A seguir discutimos brevemente algumas ferramentas de anonimização existentes e suas características principais. Posteriormente, na seção 4.4 detalharemos as principais ferramentas existentes.

O `tcpdpriv` [Minshall, 1996] é uma das mais conhecidas ferramentas de anonimização, desenvolvida para anonimizar dados coletados diretamente da interface de rede utilizando o `tcpdump`. Ela se preocupa apenas com os cabeçalhos dos pacotes IP, UDP e TCP, sendo capaz de gerar diversos níveis de anonimização, pois permite a escolha de vários campos do cabeçalho para serem anonimizados.

O `ipsumdump` [Ipsumdump, 2009] é uma ferramenta de anonimização que sumariza os dados obtidos do `tcpdump` utilizando o `tcpdpriv` e transforma esses dados para formato ASCII.

Outra ferramenta disponível é a `tcpurify` [Blanton, 2009], utilizada para obtenção de dados na interface de rede. Similar ao `tcpdump`, mas com o enfoque em privacidade, ela anonimiza diretamente o dados antes de serem armazenados e despreza o restante do pacote IP ou Ethernet, logo após reconhecer o último cabeçalho que se deseja coletar.

A ferramenta `APPI` [Koukis et al., 2006], é uma API baseada em linguagem `C`, tem como principal objetivo de projeto ser extensível, aplicando essa característica em três aspectos diferentes: permite a adição de novas funções de anonimização, possui suporte a novos protocolos e aceita entrada para vários tipos de coletores de tráfego.

O `tcpmkpub` [Pang et al., 2006] é uma ferramenta de análise de dados do tcpdump que não prevê anonimização de dados online. Ela procura ser o mais genérica possível, para permitir uma implementação fácil de uma política de anonimização através dos níveis de protocolo, ou seja, ela fornece um *framework* geral para anonimizar dados de rede que pode alojar uma gama de políticas de protocolos e de decisões.

Na mesma linha do `tcpmkpub` existe também o `FLAIM` [Slagell et al., 2006], que tem uma linguagem de especificação dos campos a serem anonimizados, tornando a configuração muito flexível. Além disso, disponibiliza várias técnicas de anonimização para cada um dos campos dos protocolos da arquitetura TCP/IP.

2.7 Outros Trabalhos Relacionados

A discussão de disponibilizar *logs* sem prejudicar a privacidade e segurança, preservando ainda a qualidade dos dados para a pesquisa, tem ganhado mais destaque a cada dia. Atualmente, temos trabalhos discutindo as mais variadas técnicas de anonimização com por exemplo, anonimização usando criptografia [Xu et al., 2002; Ramaswamy & Wolf, 2007], outros trabalhos seguem a linha de criar ambientes seguros para a coleta e análise de logs [Shanmugasundaram, 2003; Bianchi et al., 2008a; Hussain et al., 2006].

De forma geral as pesquisas se concentram em novas técnicas e ferramentas de anonimização [Luo et al., 2006], e técnicas para recuperar informações anonimizadas, ou seja, ataques contra as anonimizações [King et al., 2009; Kohno et al., 2005; Ribeiro et al., 2008]. Há também artigos que se concentram apenas em anonimização de um determinado campo, por exemplo, o endereço IP [Keardsri et al., 2009].

Também são encontrados alguns artigos que analisam a ética e os problemas jurídicos que o compartilhamento de dados pode gerar [Allman & Paxson, 2007; Ohm et al., 2007] e artigos que apresenta técnicas de avaliação da qualidade da forma de anonimização [Coull et al., 2008; Kelly et al., 2008].

Capítulo 3

Aspectos Legais

Como vimos anteriormente, o crescimento no uso da Internet para atividades do dia-a-dia ocorrem em um ritmo cada vez maior, aumentando a necessidade de melhorias da infraestrutura da Internet e de seus protocolos.

Consequentemente, houve um grande aumento na quantidade de informação privada trafegada na Internet além, é claro, de um maior número de pessoas cadastradas nas bases de dados das empresas. Isso gerou em muitos países uma preocupação em regular a proteção, manutenção e circulação dessas informações na Internet, pois as empresas trocam, entre si, informações de suas bases, assim como os pesquisadores utilizam arquivos de *logs* de rede para sua pesquisa.

Outra consequência desse crescimento da Internet foi o aumento significativo do número de crimes e fraudes pela rede. Esses fatos levaram a uma preocupação crescente da área jurídica em tornar a Internet um ambiente mais controlado. Para isso, vários países começaram a regulamentar políticas de combate aos chamados cybercrimes, dentre elas a tipificação de novos crimes relacionados ao ambiente computacional e regras para maior controle do acesso a Internet, principalmente com relação ao armazenamento de arquivos de *logs* de conexão para facilitar a identificação de criminosos.

Diante dessa regulamentação crescente entre os países e, consequentemente, uma maior preocupação com o controle dos dados que circulam na Internet, aumenta a necessidade de um rigoroso processo de coleta, armazenamento e utilização desses dados por parte de empresas, impactando diretamente a utilização e compartilhamento de dados de conexão de rede pelos pesquisadores da área.

Atualmente, a maioria dos pesquisadores coletam, manipulam e, muitas vezes, compartilham os arquivos de *logs* de conexão sem se preocupar com as normas de confidencialidade/privacidade que protegem esse tipo de dado [Ohm et al., 2007]. Entretanto, mesmo quando se preocupam com o conteúdo desses arquivos eles não sabem

como devem proceder para coletar e manipular esses arquivos de forma a garantir a segurança da rede, a privacidade dos usuários, além de se preservarem contra um processo de indenização ou até mesmo criminal.

Por outro lado, no meio dessa corrida para regulamentar o uso da Internet, organismos de proteção das liberdades individuais tentam combater excessos que tais regulamentações possam causar, por exemplo, na privacidade dos indivíduos, que é um direito básico previsto no artigo 12 da Declaração Universal dos Direitos Humanos [ONU, 1950].

Para tentar elucidar as questões jurídicas que envolvem a coleta e análise de dados de conexão de rede, neste capítulo discutiremos o que alguns países estão fazendo para tentar controlar a circulação de dados e reduzir a criminalidade na Internet. Inicialmente, a legislação da União Européia será discutida, por ser um grupo dos principais países do mundo e por já estar muito avançada nesse assunto. Em seguida, discutiremos as leis sobre o controle dos dados pessoais nos Estados Unidos e alguns países da América do Sul.

Por fim, veremos o que existe no Brasil em termos de legislação em vigor e ainda apresentaremos as características relevantes dos principais projetos de lei que tramitam no Senado Federal. Finalmente, tentaremos delinear qual a melhor forma de lidar com a coleta de dados, para garantir o alto grau de conformidade com as leis.

3.1 União Européia

Com o aumento do armazenamento de informações por meios eletrônico a União Européia já se preocupava em regulamentar o uso dessas informações por parte de empresas privadas e órgãos governamentais desde a década de 80. Em 1981 foi aprovada a convenção 108, que tem seus objetivos descritos em seu artigo 1º: "A presente Convenção destina-se a garantir, no território de cada Parte, a todas as pessoas singulares, seja qual for a sua nacionalidade ou residência, o respeito pelos seus direitos e liberdades fundamentais, e especialmente seu direito à vida privada, face ao tratamento automatizado dos dados de caráter pessoal que lhes digam respeito (protecção dos dados)" [Parlamento Europeu, 1981].

Essa convenção define dados de caráter pessoal como sendo qualquer informação sobre uma pessoa identificada ou que possa vir a ser identificada (titularond dos dados). Ela também define padrões mínimos de segurança nos quais os dados pessoais de arquivos automatizados devem estar resguardados, sob uma política apropriada, contra a destruição, acidental ou não, a perda acidental e a manipulação e divulgação

não autorizadas.

Ela também especifica a forma como esses dados podem ser obtidos e como devem ser mantidos, dando prerrogativa ao titular dos dados para acessá-los, retificá-los ou eliminá-los. Ela ainda permite que as empresas utilizem esses dados para fins de estatística ou de pesquisa científica, desde que não causem risco à privacidade dos titulares.

Observamos que no decorrer dos anos regular a proteção à privacidade sem prejudicar a livre circulação dos dados entre os países membros foi sempre uma preocupação no Parlamento Europeu, tanto que foram lançadas diversas diretivas e regulamentos (que são uma espécie de tratado entre os países membros que se comprometem a adequar sua legislação interna às diretrizes desses documentos) definindo regras sobre o assunto.

Em 1995 foi adotada a diretiva 46 [Parlamento Europeu, 1995] que busca regulamentar, novamente, a livre circulação dos dados pessoais entre os países membros, resguardando sobretudo os direitos fundamentais, dentre eles o direito à vida privada. Nessa diretiva são criadas outras duas exceções a essa regra de privacidade: a primeira é o consentimento expresso do titular para o uso dos dados; a segunda é quando os dados são anonimizados, antes de serem manipulados, garantindo que as pessoas não sejam identificadas.

Na diretiva 46/95 foi mantida a definição de dado pessoal, especificando algumas das formas de identificação indireta do titular dos dados, por exemplo, através de um número identificador, ou características físicas culturais, etc. Ela ainda define o que é o tratamento automatizado de dados pessoais, exemplificando os tipos de manipulação que esses dados podem sofrer, determinando sigilo e segurança adequados. Determina que os estados-membros devem garantir recursos judicais para quem se sentir prejudicado e sanções para os responsáveis.

Além disso, ela regulariza a transferência desses dados para países não-comunitários e, preocupada com o dinamismo na evolução da informática, cria no seu artigo 29 o "grupo de proteção das pessoas no que diz respeito ao tratamento de dados pessoais", que é formado por representantes de cada país membro, especificando entre suas atribuições, a de dar parecer sobre nível de proteção dos países membros e não-membros e dar recomendações sobre proteção das pessoas relativas ao tratamento dos dados pessoais na Comunidade Europeia.

Em 1997 foi adotada a diretiva 66 [Parlamento Europeu, 1997] que regulamentava o setor de telecomunicações e foi revogada em 2002 pela diretiva 58 de 2002, que aumentou a sua abrangência para o setor de comunicações eletrônicas. Mesmo revogada, veremos alguns detalhes de alguns artigos da diretiva 66 de 1997, para contextualizar-

mos a evolução histórica da legislação europeia sobre o tratamento de dados pessoais e proteção a privacidade. Ela regulamentava o uso dos dados pessoais dos assinantes do setor de telecomunicações. Até então, as diretivas anteriores eram direcionadas aos bancos de dados existentes nas empresas, mas nessa diretiva foi dado um destaque aos dados relativos à conexão.

Como destaque da diretiva 66/97 pode-se citar o artigo 5^o, que determinava que os estados-membros deveriam garantir a confidencialidade das comunicações na rede pública de telecomunicações e seus serviços. Isso incluía coibir a escuta, o armazenamento ou outros meios de interceptação de comunicações por terceiros sem o consentimento dos usuários, excetuando-se quando legalmente autorizados, nos casos de segurança do estado, investigação criminal, etc.

Também o artigo 6^o da diretiva 66/97 merece destaque neste trabalho, pois ele reconhecia a importância dos dados relativos à conexão, principalmente para as telecomunicações, devido à utilidade para o faturamento dos assinantes e apoio para área comercial e estatística. Além disso, determinava que o acesso aos dados de tráfego deveria ser restrito às pessoas que utilizam para aquele fim. Entretanto, o mais importante é que essa diretiva ainda estabelecia que esses dados deveriam ser apagados ou tornados anônimos após a conclusão da conexão ou do seu uso para as tarefas acima (faturamento, assistência ao cliente, detecção de fraudes, etc).

Em 2001 foi adotado o regulamento 45 [Parlamento Europeu, 2001] sobre a proteção no tratamento de dados pessoais e sua livre circulação pelas instituições e pelos órgãos comunitários. Na mesma linha de regulação do tratamento de dados da diretiva 46 de 1995 que regulava empresas privadas, este regulamento é específico para os órgãos públicos acima citados.

Para adaptar as regras à evolução de novas tecnologias a diretiva 58 de 2002 [Parlamento Europeu, 2002], como dito anteriormente, revogou a diretiva 66 de 1997. Essa nova diretiva visa uma maior abrangência em relação a anterior, pois ela não só regula o setor de telecomunicações como regula todo o setor de comunicações eletrônicas. Ela traz uma inovação em seu artigo 2^o, diferenciando dados de tráfego e a comunicação propriamente dita, onde os primeiros são as informações trocadas para estabelecer a conexão, já a comunicação propriamente dita é qualquer informação trocada entre as partes através de um serviço de comunicação eletrônica.

O artigo 5^o determina que os Estados-membros garantirão a confidencialidade das comunicações e dos dados de tráfego impedindo escutas, armazenamento, etc, exceto quando legalmente autorizados. Ainda no artigo 6^o, foi mantida a regra de descartar os dados de tráfego ou anonimizá-los assim que passam a ser desnecessários para a comunicação, excetuando também os dados necessários para a tarifação.

Com o número crescente de crimes realizados através da Internet, em 2006 foi adotada a diretiva 24 [Parlamento Europeu, 2006] que tem como objetivo regular a conservação de dados gerados no contexto dos serviços de comunicação eletrônica pública para efeitos de investigação e de repressão a crimes graves, alterando a diretiva 58/2002.

A diretiva 24/2006 mantém a distinção entre dados de tráfego e a informação privada que é trocada durante comunicação e deixa claro no artigo 1^o, número 1, que ela determina a manutenção apenas dos dados de tráfego, ou seja, os dados que são usados para estabelecer a conexão, ficando excluídos, portanto, os dados relativos ao conteúdo das comunicações eletrônicas. No item 23 das considerações iniciais, ela esclarece que só são obrigados a conservar os dados os fornecedores que geram ou tratam os mesmos, dando a entender que ela desobriga os provedores dessa determinação.

No seu artigo 1^o são derrogados os artigos 5^o, 6^o e 9^o da diretiva 58/2002. O artigo 5^o cria seis categorias de dados que devem ser conservados. A primeira, são dados necessários para encontrar e identificar a fonte de uma comunicação, depois são os dados necessários para encontrar e identificar o destino de uma comunicação, em seguida são os dados necessários para identificar a data, hora e duração de uma comunicação. Outra categoria são os dados necessários para identificar o tipo de comunicação; e também os dados para identificar o equipamento de telecomunicações dos utilizadores e por fim, os dados para identificar a localização do equipamento de comunicação móvel.

O artigo 6^o determina, aos estados-membros, que o tempo mínimo de conservação dos dados que as legislações internas devem estipular é de seis meses e não devem ultrapassar dois anos.

Aspectos Relevantes para a Coleta e Anonimização

Como vimos, inicialmente foram adotados diversos atos no sentido de regular a troca de informações entre os países-membros, priorizando as garantias individuais dos usuários, nesses casos, o direito à vida privada. Entretanto, como dito anteriormente, com a proliferação do uso da Internet também houve um aumento no número de crimes relacinados a esse meio.

Em contrapartida, os países começaram a regular de forma autonôma a retenção dos dados de comunicação eletrônica. Então, a União Européia se viu na obrigação de estabelecer regras que unificassem as legislações dos países-membros, determinando a retenção de dados de conexão pela operadora e estabelecendo critérios para essa

retenção, dentro dos princípios da privacidade e do Estado de Direito.

Para o nosso estudo fica claro, como vimos na diretiva 24 de 2006 que a coleta do conteúdo da comunicação só pode ser feita através de ordem judicial. Com relação aos dados da conexão, se esses forem anonimizados ou se tiverem o consetimento do usuário, eles poderão ser usados tanto pelos pesquisadores quanto pelos administradores de redes. Mas os dados de conexão não anonimizados só podem ser coletados para garantir o bom funcionamento da rede. Sendo assim, os administradores não poderão repassar esses dados para a pesquisa.

3.2 América

O Canadá possui dois decretos sobre a privacidade dos dados: o primeiro é de 1982 e regulamenta a coleta, o uso e a divulgação de dados pelos órgão governamentais e o segundo é de 2001, que estabelece princípios que as organizações em geral devem seguir na coleta, armazenamento e uso dos dados pessoais.

Nos Estados Unidos vigora a *common law*, que é o sistema de formação de leis através dos costumes e de decisões judiciais. Devido a esse sistema, os EUA possuem uma diversidade grande de decisões judiciais sobre a privacidade de dados, leis estaduais e leis federais. Devido a essa descentralização, o congresso americano começou a criar diversos *acts*, que são as leis federais, regulamentando a privacidade de determinados tipos de dados, por exemplo, o *Health Information and Portability Accountability Act* (HIPAA), que trata sobre a manutenção e tratamento dos dados relativos à saúde, o *Children's Online Privacy Protection Act* (COPPA), que proíbe aos *sites* a coleta de dados de crianças sem a autorização dos pais e o *Driver's Privacy Protection Act*, que proíbe o estado a revelar dados pessoais dos cidadãos, como o endereço, número do seguro social, etc.

Em 1986 entrou em vigor o *Electronic Communications Privacy Act* (ECPA), que regula a interceptação da comunicação de dados, proibindo que se intercepte, acesse e divulge informações de uma comunicação eletrônica, prevendo algumas exceções a essa regra; por exemplo, a invasão não autorizada de sistemas por *hackers* é considerada ilegal, mesmo que esta invasão não cause dano. Após os ataques terroristas, em 2001, entrou em vigor USA *Patriot Act* que entre outras coisas, permite a interceptação de comunicação de voz em computadores suspeitos.

Na América do Sul alguns países já possuem lei específica de proteção de dados. O Chile, por exemplo, aprovou um lei de proteção de dados em 1999, dando direito às pessoas de acesso e correção de suas informações. A Argentina, em 2000, sancionou a

Lei 25.326 sobre a proteção dos dados pessoais, seguindo a tendência das leis internacionais, que prevê a proteção dos dados pessoais, estabelecendo regras de informação sobre o tratamento dos dados.

Além disso, ela criou um órgão de regulamentação e aplicação da lei proteção aos dados pessoias. Dessa forma, a Argentina em 2003 obteve um parecer de adequação de proteção da União Européia, se tornando o primeiro país da América do Sul com autorização de transferência de dados de/para a Europa.

3.3 Brasil

No Brasil, apesar de não estarmos tão avançado com relação às normas de troca, preservação e privacidade dos dados dos meios de comunicação eletrônicos, não se pode afirmar que não exista nenhuma regra sobre o assunto. Nessa seção, discutiremos as leis que falam de privacidade e interceptação de dados e que atualmente vigoram no país. Além disso, veremos o principal projeto de lei que está em tramitação no Congresso Nacional, que de alguma forma ajudará no compreensão de como esse assunto deve evoluir.

3.3.1 Legislação em Vigor

3.3.1.1 Código Penal Brasileiro

O Decreto-lei n^o 2.848 de dezembro de 1940, o nosso Código Penal [Congresso Nacional, 1940], já descrevia em seu artigo 151, o crime de violação de correspondência, que prevê uma pena de um a seis meses ou multa; e diz no inciso II do parágrafo 1^o que incorre na mesma pena quem praticar o tipo penal VIOLAÇÃO DE COMUNICAÇÃO TELEGRÁFICA, RADIOELÉTRICA OU TELEFÔNICA que é descrito da seguinte forma:

> II- quem indevidamente divulga, transmite a outrem ou utiliza abusivamente comunicação telegráfica ou radioelétrica dirigida a terceiro, ou conversação telefônica entre pessoas.

Sendo assim, vemos que o nosso Código Penal tornava crime apenas quem divulga ou transmite a outrem conversação telefônica entre outras pessoas. Isso significa que simples ato de interceptar e/ou gravar uma comunicação telefônica não era considerado crime, pois o crime era consumado somente no momento da divulgação ou transmissão da informação a outrem [Jesus, 1997] . Esse inciso se resume também a apenas comunicações telefônica e radioelétrica, não incluindo nosso assunto que é tráfego de redes,

mas como não temos leis específicas sobre a comunicação eletrônica e o procedimento na comunicação são similares, faremos sempre um paralelo entre esses dois tipos de comunicação, ajudando a ilustrar a evolução do tratamento dado pela lei em nosso país para a interceptação de dados.

3.3.1.2 Constituição da República de 1988

Nossa Constituição de 1988 [Congresso Nacional, 1988], prevê em seu artigo 5º, inciso X, a inviolabilidade da intimidade e da vida privada das pessoas e, no inciso XII do mesmo artigo, prevê a inviolabidade da correspondência e das comunicações, como mostrado a seguir:

> "Art. 5º: Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade, nos termos seguintes:
>
> X - são invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação;
>
> XII - é inviolável o sigilo da correspondência e das comunicações telegráficas, de dados e das comunicações telefônicas, salvo, no último caso, por ordem judicial, nas hipóteses e na forma que a lei estabelecer para fins de investigação criminal ou instrução processual penal".

A princípio, lendo rapidamente o inciso XII, parece que os legisladores deixaram claro o seu interesse em tornar inviolável o sigilo da correspondência, das comunicações telegráficas e de dados, abrindo exceção às comunicações telefônicas quando houver ordem judicial.

O que aparentemente já está definido é, na verdade, uma grande polêmica entre os juristas do país, pois a expressão, "salvo, no último caso", não deixa claro a que se refere [Delmanto et al., 1998], criando pelo menos duas correntes de interpretação desse inciso. A primeira corrente defende que o inciso possui quatro itens (correspondência, comunicações telegráficas, comunicações de dados e comunicações telefônicas) sendo, assim, que a exceção prevista diante de autorização judicial é relativa apenas às comunicações telefônicas, tornando o sigilo da correspondência, da comunicações telegráficas e de dados absoluto [Greco Filho, 1996]. Ao defender a exceção somente às comunicações telefônicas, Delmanto et al. [1998], citando Themistocles Cavalcanti [1],

[1] Themistocles Cavalcanti, Do Controle da Constitucionalidade, 1986, p. 164, apud Alberto silva Franco, Crimes Hediodos, 1994, p.90

diz que as garantias individuais devem ser interpretadas de forma extensiva, ou seja, diante de uma regra com texto duvidoso deve-se ampliar a garantia de liberdade e não restringi-la.

Por outro lado, existem autores que defendem a idéia de que esse inciso é dividido em apenas duas partes, sendo a primeira o direito ao sigilo da correspondência e das comunicações telegráficas e a segunda, o direito ao sigilo comunicações de dados e das comunicações telefônicas. Dessa forma, a exceção prevista de quebra do sigilo se destina tanto às comunicações de dados, quanto às comunicações telefônicas [Gomes & Cervini, 1997]. Em seu voto no julgamento do pedido 577 [Mello, 1992] de quebra de sigilo bancário, o Ministro do Supremo Tribunal Federal (STF) Marco Aurélio Mello, declara esse entendimento sobre esse preceito.

Além de falar da exceção vista acima, o inciso XII determina que a legislação infra-constitucional a regulamente na sua forma e hipóteses, para fins de investigação criminal ou instrução penal. Nesse ponto não há discussão, ou seja, as duas correntes concordam que para haver a quebra do sigilo é preciso uma ordem judicial e isso somente para fins de investigação criminal ou instrução processual penal.

3.3.1.3 Lei 9296

Conforme previsto em nossa Constituição, a lei 9296 [Congresso Nacional, 1996] foi promulgada em 1996 para regulamentar o seu inciso XII do artigo 5^o. No parágrafo único do artigo 1^o, os legisladores deixam claro sua interpretação do referido inciso da Constituição, "O disposto nesta Lei aplica-se à interceptação do fluxo de comunicações em sistemas de informática e telemática". Infelizmente este artigo não acabou com a polêmica, pois os defensores de que a Constituição autoriza apenas a interceptação telefônica afirmam que a lei 9296 é inconstitucional, pois ela estende o alcance da norma constitucional, restringindo o direito à privacidade e uma norma infra-constitucional não pode contrariar o texto da Constituição.

Atualmente está no STF uma Ação Direta de Inconstitucionalidade (ADI) da lei 9296 pedindo a inconstitucionalidade de cinco dispositivos dessa lei; entre eles, o parágrafo único do artigo 1^o. A decisão a ser tomada pelo STF deverá resolver a questão em definitivo.

A lei 9296 no artigo 2^o define três hipóteses onde a interceptação telefônica não será admitida: quando não houver indícios de autoria ou participação, quando a prova puder ser feita por outros meios e quando o fato investigado for uma infração penal punida, no máximo, com detenção. Nesse artigo, a lei 9296 também é criticada por alguns autores, pois ela contraria a boa prática da legislação, onde deveriam ser descritas as

hipóteses em que a interceptação é admitida [Greco Filho, 1996]. Apesar da polêmica, esse artigo mostra a importância dada pelo legislador ao direito à privacidade pois, de acordo com essas exceções, esse bem só poderá ser maculado diante de sérias razões.

Outros dois pontos que se destacam nessa lei é que a interceptação telefônica se dará em autos apartados, apensados aos autos do inquérito policial ou do processo penal e que no artigo 9º ela define as formas de destruição das gravações que não interessar em uma investigação ou processo. Dessa forma, o legislador, mesmo permitindo exceções ao direito de sigilo nas comunicações, demonstra uma preocupação em presevar ao máximo a privacidade do investigado.

Finalmente, a lei 9296 revoga parcialmente o artigo 151 inciso II do Código Penal Brasileiro [Congresso Nacional, 1940], visto anteriormente. tornando crime não só a transmissão ou divulgação indevida de conteúdo da comunicação telefôncia, mas também o ato de interceptação de comunicações telefônicas, de informática ou telemática sem autorização judicial, conforme o artigo 10 a seguir:

> "Constitui crime realizar interceptação de comunicações telefônicas, de informática ou telemática, ou quebrar segredo da Justiça, sem autorização judicial ou com objetivos não autorizados em lei"

. Com esta nova redação, a simples interceptação constitui crime, o que é fundamental para a análise deste trabalho. Além disso, esse artigo prevê uma pena de reclusão de dois a quatro anos e multa, tornando a punição para quem incorre nesse crime muito mais severa do que a lei anterior.

Aspectos Relevantes para a Coleta e Anonimização

Diante do exposto, confrontaremos nessa seção o nosso entendimento da legislação em vigor com a coleta de dados de tráfego de rede. Apesar dos bons argumentos contra a permissão constitucional da intercepção com autorização judicial de comunicações de dados, tornando esse direito absoluto; ao analisar os argumentos da corrente contrária, junto com algumas decisões de nossos tribunais, me parece que o entendimento de que a Constituição permite a interceptação judicial tanto das comunicações telefônicas quanto nas comunicações de dados, deve ser a interpretação válida, entendendo portanto, que o parágrafo único da artigo 1º da lei 9296/96 é constitucional.

Devemos primeiramente separar os dois tipos de dados que a comunicação de dados possui: o primeiro é o conteúdo da comunicação, o outro são os dados de registros,

os quais são necessários para a realização e controle da comunicação.

Fazendo um paralelo com as conexões telefônicas, as operadoras têm necessidade de manter os dados de registros para fazer a tarifação de serviços. Esses dados possuem informações técnicas como, por exemplo, hora da chamada, duração, etc, e informações pessoais como número do telefone, registro de chamada, etc. Consequentemente, podemos concluir que o armazenamento dos dados de conexão não é proibido, mas a sua divulgação fere o direito de privacidade instituído na Constituição de 1988 em seu artigo 5^o inciso X e só pode ser autorizado através de uma ordem judicial.

Portanto, se os dados de conexão de rede forem equiparados aos da telefonia, os administradores de rede podem coletar e armazenar esses dados e até mesmo manipulá-los, desde que sejam utilizados para o bom funcionamento da atividade. Entretanto, quanto ao repasse desses dados a outrem, entendo que estaria contrariando a determinação legal. Entendo também que há a possibilidade de compartilhamento desses dados caso os mesmos passem por um processo de anonimização que torne inviável a identificação do dono da informação e consequentemente garanta a privacidade dos dados.

Com relação ao conteúdo da comunicação, ou seja, o *payload* do pacote TCP ou UDP, entendo que a lei 9296 é clara: só pode ser coletado/interceptado por ordem judicial e esta só poderá ser concedida quando houver indícios de autoria de crime punível com reclusão e não houver possibilidade de prova por outros meios. Esses dados devem ainda ficar armazenados em local seguro ao qual apenas o responsável ou a autoridade policial tenham acesso.

Continuando o paralelo entre a interceptação de comunicações de dados e de telefone, acredito que tecnicamente seja muito mais complicado executar a coleta da comunicação de uma determinada pessoa ou do seu computador pessoal. Isto ocorre porque no caso da interceptação telefônica, o pedido deve ser feito para um número de telefone determinado, mas na interceptação de dados esse pedido na maioria das vezes não terá condições de especificar qual endereço deve ser interceptado, pois o usuário poderá a cada conexão à Internet obter um endereço IP diferente.

Dessa forma, a monitoração deverá ser feita na operadora que o usuário suspeito é cliente, pois só ela teria condições de saber qual o endereço o computador do suspeito estaria utilizando. Na telefonia o número do telefone não se altera, mas a quebra de sigilo é solicitada junto à operadora do suspeito.

3.3.2 Projetos de Lei em Tramitação

Devido à grande repercusão que os crimes de informática vêm tendo, temos no Congresso Nacional diversos projetos que regulamentam questões relativas a informática. Dentre eles, destacamos dois projetos que achamos de maior relevância, o primeiro é o projeto 494/2000 [Senado Federal, 2008] que tem como principal característica seguir em linhas gerais o conteúdo das diretivas europeias. Já o segundo é o projeto que tem tido muito destaque na mídia devido à polêmica criada sobre o controle do uso da Internet, mais conhecido como o Projeto do Senador Eduardo Azeredo [Senado Federal, 2009], ele é um substitutivo de outros três projetos que estavam em tramitação no Senado Federal, como veremos em seguida.

3.3.2.1 Projeto de Lei 494 de 2008

Esse projeto de lei foi proposto pela Comissão Parlamentar de Inquérito (CPI) sobre a pedofilia e tem em seu artigo 1^o a descrição dos seu objetivo.

> "Essa lei disciplina a forma, os prazos e os meios de preservão e transferência de dados informáticos mantidos por fornecedores de serviço e autoridades públicas, para fins de investigação de crimes praticados contra crianças e adolescentes".

Apesar de direcionada para investigação de crimes de pedofilia, esse projeto tem muita similiaridade com as diretivas 58/2002 e 24/2006 da União Européia, conforme mencionado no início desse capítulo. No artigo 2^o o projeto define três tipos de fornecedores de serviço: de telecomunicações, de acesso e de conteúdo ou interativo. Neste mesmo artigo são definidas três categorias de dados: de conexão, cadastrais do usuário e relativos ao conteúdo da comunicação. Na primeira categoria estão os dados necessários para realizar uma conexão; a segunda engloba apenas os cadastros dos usuários/clientes; finalmente a categoria dos dados trafegados propriamente ditos, onde se encontra o conteúdo da comunicação.

No artigo 3^o, é determinado um prazo de 3 anos para a manutenção dos dados e de conexões para os fornecedores de serviço de telecomunicações e de acesso e de 6 meses para os fornecedores de conteúdo ou interativo. Além disso, ela determina em seu artigo 7^o que em qualquer fase da investigação criminal envolvendo delitos contra crianças e adolescentes, esses dados devem ser transferidos para a autoridade policial ou Ministério Público sem ordem judicial prévia e os dados de conteúdo apenas com autorização policial.

Outro artigo que gerará polêmica é o artigo 8^o, pois ele determina que a autoridade policial poderá, sem autorização judicial, solicitar a preservação imediata dos dados de conteúdo, para fins de investigação de crimes envolvendo crianças e adolescentes. A transferência desses dados para a autoridade solicitante deverá ser feita apenas com autorização judicial.

Finalmente, no artigo 14, esse projeto prevê que o Poder Executivo estabelecerá padrões e formatos para solicitações e as respostas a pedidos, por dados. Isso ajudará os administradores a fazer a coleta com mais segurança.

3.3.2.2 Substitutivo dos Projetos de Lei 89/2003, 137/2000 e 76/2000

Este projeto é conhecido popularmente com a Lei do Azeredo, pois ele foi o relator dos projetos de lei citados no título e propos um projeto que os substituiu. Inicialmente, esse projeto substitutivo causou muita polêmica, pois ele criava vários crimes que no entendimento de muitos eram mal definidos e por consequência criava situações absurdas.

Depois de várias discussões este projeto sofreu diversas modificações na sua redação inicial que amenizaram as polêmicas sobre o assunto. Este substitutivo foi aprovado no Senado em julho de 2009 e foi encaminhado para a Câmara dos Deputados, e mesmo depois de modificado, ele ainda apresenta alguns artigos que afetam o nosso tema e, portanto, discutiremos a seguir.

Em seu artigo 3^o, ele acrescenta um tipo penal ao artigo 154 do Código Penal, criminalizando, a divulgação ou utilização indevida de informações e dados pessoais.

> "Divulgar, utilizar, comercializar ou disponibilizar dados e informações pessoais contidas em sistema informatizado com finalidade distinta da que motivou o seu registro, salvo nos casos previstos em lei ou mediante expressa anuência da pessoa a que se referem, ou de seu representante legal. Pena: detenção de um a dois anos, e multa".

No artigo 16, este projeto traz várias definições de termos relacionados a informática, tais como, dispositivo de comunicação, sistema informatizado, rede de computadores, código malicioso, dados informáticos e dados de tráfego. As duas últimas definições são de nosso interesse relatar.

> "Dados informáticos: qualquer representação de fatos, de informações ou de conceitos sob forma suscetível de processamento numa rede de computadores ou dispositivo de comunicação ou sistema informatizado;

> Dados de tráfego: todos os dados informáticos relacionados com sua comunicação efetuada por meio de uma rede de computadores, sistema informatizado ou dispositivo de comunicação, gerados por eles como elemento de uma cadeia de comunicação, indicando origem da comunicação, o destino, o trajeto, a data, o tamanho, a duração ou o tipo de serviço subjacente".

Para finalizar no seu artigo 22 esse projeto propõe o armazenamento dos dados relativos à conexão por três anos, delimitando em seu inciso I o prazo e o tipo de dado que deve ser armazenado, mas limitando o seu fornecimento à autoridade investigatória mediante autorização judicial. E em seu inciso II ele determina que outros dados, que se presume que são os dados da comunicação, devem ser preservados após requisição judicial, "respondendo civil e penalmente pela sua confidencialidade e inviolabilidade".

Aspectos Relevantes para a Coleta e Anonimização

Após analisar a legislação existente e compará-la com a regras estabelecidas em outros países, acho que alguns artigos do projeto 494/2008 são inconstitucionais.

No artigo 7^o, que dispõe sobre o acesso e transfêrencia dos dados, em seu *caput* ele autoriza a transferência dos dados em qualquer fase da investigação ou processual. Esse texto é constitucional, mas acredito que ele altera o inciso III do artigo 2^o da lei 9296, que diz que o fato investigado deve ser uma infração penal punida com reclusão.

Já no inciso I desse mesmo artigo, considero que o mesmo contraria o direito à privacidade, pois ele determina a transferência dos dados de conexão para as autoridades competentes sem autorização judicial. Novamente fazendo um paralelo com as comunicações telefônicas, o sigilo das conexões só é quebrado com autorização judicial. Nessa mesma linha o inciso II desse artigo não contraria a nossa Constituição.

Outro artigo que nos parece inconstitucional é o artigo 8^o, pois ele determina ao fornecedor de conteúdo que preserve os dados relativos ao conteúdo da comunicação apenas com a solicitação da autoridade policial. Como determinado pela Constituição, a interceptação somente é permitida com a autorização judicial.

Como mencionado anteriormente o artigo 14 é interessante ao estabelecer que o Poder Executivo irá determinar os padrões e formatos de solicitações e respostas, facilitando a tarefa do administrador de redes pois, como vimos, existem diversas formas de coleta de dados.

No que diz respeito ao tema do nosso trabalho esse projeto não interfere muito no que já é estabelicido na legislação atual, pois para a pesquisa ele continuará não

podendo utilizar-se do conteúdo dos pacotes e em relação aos dados de conexão, esses só poderão ser utilizados se forem anonimizados antes do uso.

Já o projeto substitutivo do Senador Eduardo Azeredo, apesar de não distinguir claramente os dados de comunicação dos dados de conexão, enumera os dados que têm obrigatoriedade de ser mantidos. Outro ponto mal definido é o responsável por fazer a manutenção desses dados, o termo está muito genérico dentro a gama de entidades que fazem o provimento de acesso a rede de computadores mundial. Os pontos positivos são os itens relativos à retenção dos dados, este projeto está de acordo com o que foi visto na nossa legislação em vigor e portanto, parece estar de acordo com o que estabelece nossa Constituição.

Capítulo 4

Aspectos Técnicos

Como visto nos capítulos anteriores, os pacotes de tráfego contêm várias informações essenciais para a comunicação de rede. Além das informações contidas dentro dos dados transmitidos pela aplicação, diversos campos dos cabeçalhos da pilha TCP/IP podem conter informações que identificam algum usuário e/ou equipamento de rede e que afetam diretamente a sua privacidade e a segurança da rede.

Neste capítulo analisaremos os aspectos técnicos da anonimização de dados, começando com uma análise de cada um dos campos dos cabeçalhos dos principais protocolos da arquitetura TCP/IP, destacando quais campos podem ser usados para a violação da privacidade e a segurança dos sistemas. Em seguida vamos discutir as técnicas de anonimização existentes, com ênfase para a anonimização de endereços. Por fim, veremos mais detalhadamente algumas das principais ferramentas de anonimização de dados existentes.

4.1 Aspectos relacionados à anonimização na arquitetura TCP/IP

Como vimos anteriormente, a arquitetura TCP/IP é composta de quatro camadas, onde cada camada possui um ou mais protocolos, que fazem o encapsulamento das mensagens na transmissão de dados. Como discutido anteriormente, em cada camada um protocolo é responsável por incluir seu cabeçalho à frente dos dados da camada superior e repassar esse pacote (dados recebidos mais o cabeçalho incluído) para a camada inferior. No recebimento dos dados, o processo é invertido e cada um dos protocolos retira seu cabeçalho e repassa os dados para a camada acima, até que a informação seja entregue à aplicação.

As seções a seguir descrevem as principais funcionalidades de cada camada e os principais protocolos encontrados em cada uma. Primeiramente será feita um análise das informações contidas nos cabeçalhos de cada protocolo e, em seguida, serão identificados os campos que possuem implicações quanto a questões de privacidade dos usuários e segurança da rede. A figura 4.1, dá uma idéia de como os protocolos estão distribuídos nas camadas.

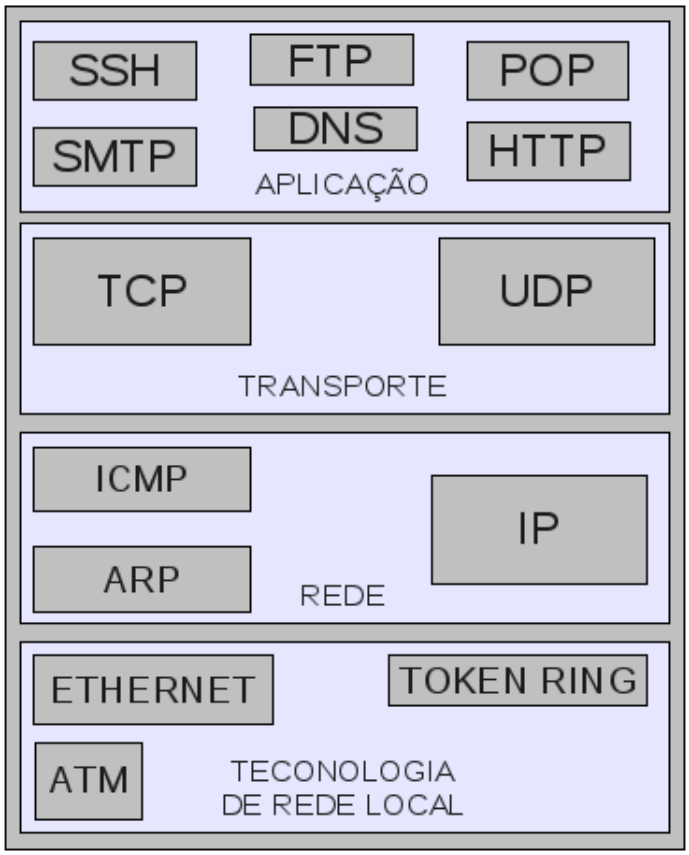


Figura 4.1. Camadas TCP/IP e alguns de seus protocolos

4.1.1 Aplicação

Nessa camada encontram-se os protocolos das aplicações existentes. Como discutido na seção 2.5.1, a análise de protocolos de aplicação exige conhecimento específico sobre cada protocolo e aplicação em particular, o que não é tarefa simples. Entretanto, um adversário que tenha interesse em extrair informações do conteúdo dessas mensagens teria condições de fazê-lo caso essa informação estivesse disponível. Como cada aplicação tem seus detalhes particulares, manter esses dados de forma anonimizada e segura

em um arquivo de registro de tráfego não é considerado uma tarefa factível na maior parte dos casos.

Devido a essa dificuldade, na maioria das vezes essa informação não é coletada: administradores comumente configuram coletores como o `tcpdump` para coletar apenas os primeiros bytes de cada pacote, em número suficiente para cobrir os cabeçalhos dos protocolos até a camada de transporte. Ferramentas de anonimização também normalmente removem essa informação.

Isto não significa que essa camada não seja importante e não seja necessária a sua verificação, mas que a sua análise no caso de tráfego de rede se resume a identificar a existência de dados ou não. A informação disponível nessa camada, mesmo não sendo todo o conteúdo do pacote, pode causar um grande prejuízo para a privacidade dos usuários. Nesse sentido, uma questão importante seria identificar se, no processo de coleta/anonimização de uma certa amostra de tráfego, os dados foram realmente removidos de cada pacote, ou se todos os pacotes de rede foram apenas truncados em um certo comprimento, o que pode permitir que, em determinados casos, alguns bytes de dados ainda estejam presentes e em certos casos, apenas alguns bytes podem revelar informações privadas.

4.1.2 Transporte

A camada de transporte na arquitetura TCP/IP, atualmente, tem dois protocolos de maior relevância: o *Transmission Control Protocol* (TCP) e o *User Datagram Protocol* (UDP).

4.1.2.1 UDP

O protocolo UDP é conhecido por não ser orientado a conexões e não oferecer garantias de entrega. Com isso, ele possui um número reduzido de campos em seu cabeçalho, pois não tem funções mais complexas. Os campos que ele possui são porta de origem e destino do pacote, comprimento do cabeçalho e soma de verificação.

Os campos PORTA DE ORIGEM e PORTA DE DESTINO, são os campos que identificam a terminação da conexão. Geralmente, aplicações padronizadas possuem uma porta padrão na qual o sistema operacional ficará aguardando uma conexão, por esse motivo, portas trazem a identificação da aplicação.

Em seguida, temos o campo COMPRIMENTO DO CABEÇALHO (HLEN); como o próprio nome diz, ele informa o tamanho do cabeçalho UDP, já que este pode ter tamanho variável. Por último, temos o campo SOMA DE VERIFICAÇÃO DE CABEÇALHO (*checksum*), sua função é verificar a integridade do pacote recebido.

Aspectos Relevantes para a Anonimização

Os campos PORTA DE ORIGEM e PORTA DE DESTINO, como vimos, servem para identificar o processo/aplicação de origem e destino do datagrama. Esses campos são muito importantes para a pesquisa em redes, porque muitas análises levam em consideração qual a porta que está sendo acessada, ou aplicação está sendo utilizada.

O problema é que a descoberta, por um adversário, que determinado servidor aceita conexão em uma certa porta é considerada uma falha de segurança, pois essa informação indica qual serviço é executado nesse servidor, com essa informação o adversário poderá explorar possíveis falhas de segurança existentes nesse serviço.

Com relação à privacidade, ela pode ser afetada na medida em que se descobre que a máquina de determinada pessoa acessou um serviço de um determinado servidor. Por exemplo, se uma pessoa conectou um servidor HTTP que disponibiliza apenas conteúdo ilegal. Entretanto, para isso é necessário que se descubra o endereço das máquinas envolvidas e quem usou esse endereço, ou seja, apenas a informação da porta não é suficiente para que a privacidade de alguém seja invadida.

O campo SOMA DE VERIFICAÇÃO (*checksum*), como visto, é formado pelo resultado da soma de complemento de um do cabeçalho e dados do pacote UDP. Com essa soma de complemento de um, campos menores que 32 bits são passíveis de serem descobertos se apenas um campo de até 16 bits for anonimizado, pelo *checksum* é possível inferir qual seria o valor original. Como os dados utilizados em alguns campos são dados considerados sensíveis ao anonimato e a segurança, esse campo deve ser anonimizado para evitar esses riscos. Esse campo é importante para a pesquisa devido à sua função de identificar se o pacote foi transmitido sem erros. Assim sendo, existem técnicas que anonimizam esse campo utilizando códigos para identificar se houve erros ou não [Blanton, 2009], mas isso dificulta a análise dos dados, pois o administrador de rede deverá saber qual o padrão de código utilizado pela ferramenta de anonimização.

Outra opção, diante da necessidade de anonimização, é que ele seja recalculado de acordo com o novo cabeçalho. Caso ele tenha identicado erro, ele deve ser regerado com erro, dessa forma atenderia à necessidade de anonimização sem afetar a pesquisa. O grande problema dessa solução é que os dados são utilizados para calcular esse campo. Ou seja, levando em consideração que os pacotes tenham sido coletados sem os dados, ou pelos menos, apenas uma pequena parte foi coletada devido à opção padrão do `tcpdump`, não é viável saber se o campo está correto ou não. Isso nos leva à conclusão que o campo pode ser anonimizado sem uma solução que atenda as expectativas dos pesquisadores.

4.1.2.2 TCP

O protocolo TCP é o protocolo orientado a conexões e como visto na seção 2.4, possui funcionalidade de ordenação e confirmação de recebimentos dos pacotes. Para implementar essas características possui diversos campos de controle no seu cabeçalho, conforme ilustra a figura 4.2.

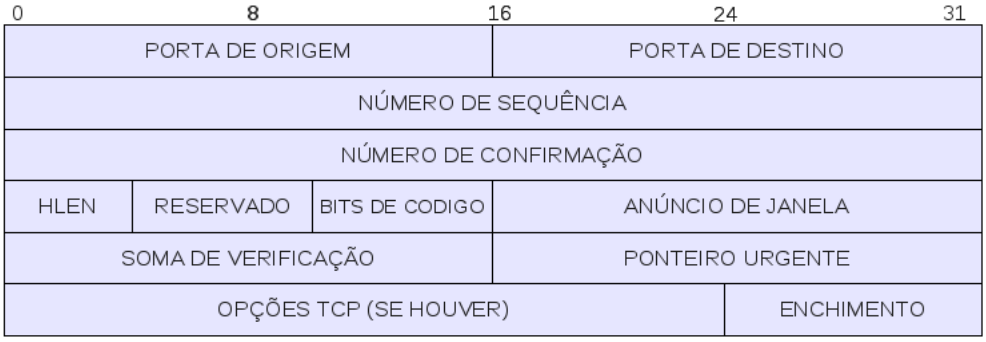


Figura 4.2. Cabeçalho do pacote TCP

Os primeiros dois campos são as PORTAS DE ORIGEM e PORTAS DE DESTINO, e têm a mesma função dos campos homônimos do protocolo UDP.

Outro campo desse protocolo é o NÚMERO DE SEQUÊNCIA, que tem como principal função enumerar e identificar cada pacote de uma conexão, para que as funções de confirmação e ordenação dos pacotes possam ser implementadas. O campo seguinte é o NÚMERO DE CONFIRMAÇÃO, que indica o número do último byte que o receptor recebeu com sucesso na sequência de dados. Em seguida, temos o campo COMPRIMENTO DO CABEÇALHO (HLEN), como no UDP informa o tamanho do cabeçalho TCPP, já que este pode ter tamanho variável. O campo RESERVADO foi criado para utilização futura, mas atualmente ele não é utilizado.

Além de transmitir o dados de uma aplicação, o cabeçalho TCP pode também ser utilizado para confirmar o recebimento dos dados, solicitar o estabelecimento ou encerramento da conexão. Para identificar essas informações em pacotes TCP, foi definido o campo FLAGS ou bits de código, que possui 6 bits, onde cada um indica uma funcionalidade do pacote. O primeiro bit, URG, indica a existência de dados urgentes no pacote; em seguida, o bit ACK indica que o valor do campo de reconhecimento é válido; o bit PSH indica que o receptor deve repassar os dados para a camada superior; o bit RST encerra uma conexão com erro; o bit SYN sincroniza o número de sequência no estabelecimento de uma conexão e, finalmente, o bit FIN indica o fim de uma conexão.

A seguir, temos o campo ANÚNCIO DE JANELA, que é muito importante na funcionalidade de controle de congestionamento do TCP, pois indica o número de bytes que o receptor pode aceitar ser enviado a partir do último byte confirmado [Peterson & Davie, 2003]. Temos ainda o campo SOMA DE VERIFICAÇÃO que é similar ao campo com mesmo nome visto no protocolo UDP.

Em seguida o campo OPÇÕES disponibiliza funcionalidades não obrigatórias, como a negociação do tamanho máximo dos pacotes (MSS), o uso de confirmação seletiva (SACK), marca de tempo dos segmentos (*timestamp*) e aumento do tamanho da janela de transmissão (WSCALE). E por fim o campo PREENCHIMENTO (*padding*) tem a função de garantir que o cabeçalho será múltiplo de 32 bits.

Aspectos Relevantes para a Anonimização

O campo de NÚMERO DE SEQUÊNCIA identifica a posição dos bytes dos dados no seguimento a ser transmitido. Esse campo é utilizado por algumas ferramentas para identificação do sistema operacional da máquina que enviou o pacote, pois cada sistema operacional pode ter um padrão diferente na inicialização desse número [Nmap, 2009]. A descoberta dessa informação pode afetar diretamente a segurança da rede, pois se o adversário souber o endereço da máquina, ele pode tentar explorar alguma falha de segurança específica daquele sistema operacional. Ao mesmo tempo, essa informação pode facilitar a identificação de determinada máquina, caso o sistema operacional seja muito específico. Sendo assim, ele também é um campo importante para sofrer alterações no processo de anonimização. Como nos campos PORTA DE ORIGEM E PORTA DE DESTINO ele é importante para os pesquisadores, pois é através desse campo que se consegue identificar que aquele pacote pertence a determinado fluxo de informação.

Da mesma forma que o campo NÚMERO DE SEQÜÊNCIA, através dos campos de ANÚNCIO DE JANELA e OPÇÕES do TCP pode ser possível identificar o sistema operacional que originou o pacote, portanto eles podem ocasionar as mesmas falhas de segurança e privacidade vistas no campo anterior. Por outro lado, a anonimização desses campos não afeta muito a pesquisa, pois geralmente eles não são usados para avaliação de nenhum evento. Outros campos relevantes para a anonimização, como visto no protocolo UDP, são as PORTAS DE ORIGEM e PORTAS DE DESTINO e a SOMA DE VERIFICAÇÃO. Todos devem seguir as recomendações apresentadas na seção anterior.

4.1.3 Camada de Rede

A camada de rede é principalmente definida, na arquitetura TCP/IP, pelo protocolo IP (*Internet Protocol*). Além dele, outros protocolos importantes que merecem menção neste trabalho são ARP, ICMP e protocolos de roteamento como RIP, OSPF e BGP.

No contexto atual é possível encontrar na rede pacotes de duas versões diferentes do protocolo IP: IPv4 e IPv6. Apesar da primeira ainda ser dominante, é possível que no futuro a Internet migre para usar a versão 6 de forma predominante. Sendo assim, neste trabalho discutimos as duas versões, apesar de mantermos um foco maior na versão 4, ainda a mais comum.

4.1.3.1 IPv4

Como ilustrado na figura 4.3, o cabeçalho IPv4 possui vários campos de controle para conseguir enviar um pacote para outra máquina conectada à Internet. O primeiro campo é a VERSÃO DO PROTOCOLO e, como o próprio nome indica, identifica a versão do protocolo IP utilizada para criar o pacote. Dependendo do valor desse campo, o restante do pacote é interpretado de acordo com a definição do cabeçalho daquela versão do protocolo IP (4 ou 6).

Em seguida, o campo COMPRIMENTO DO CABEÇALHO (HLEN) especifica o comprimento do cabeçalho. O TIPO DE SERVIÇO é um campo de 8 bits que determina como o pacote deve ser tratado pelos roteadores, por exemplo, com prioridade, alta confiabilidade, etc. Originalmente ele era ignorado, mas mais recentemente algumas aplicações de multimídia passaram a usar esse campo e alguns roteadores passaram a interpretá-los para tentar melhorar a qualidade de transmissões desse tipo de serviço. O campo seguinte é o COMPRIMENTO TOTAL do pacote, incluindo os dados.

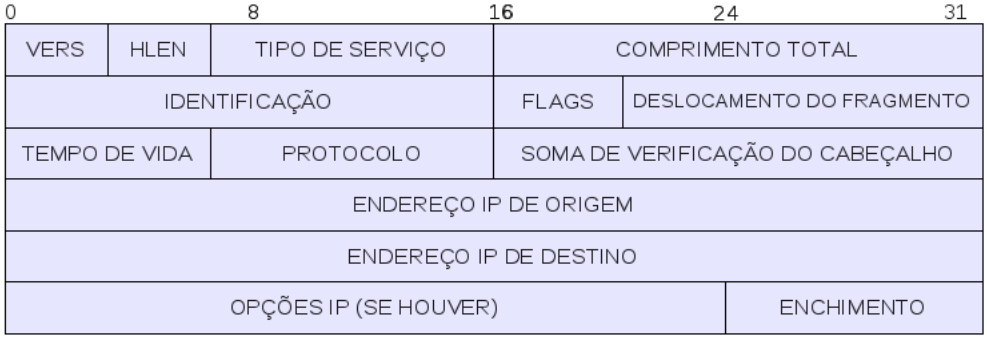


Figura 4.3. Cabeçalho do pacote IPv4P

Na segunda linha é visto o campo IDENTIFICAÇÃO que é um número único que identifica cada pacote enviado por uma máquina. Em seguida vem o campo FLAGS, usado para controlar a fragmentação do pacote IP. Também com a função de ajudar a controlar a fragmentação do pacote, o campo DESLOCAMENTO DE FRAGMENTO (*offset*) identifica o deslocamento do fragmento do pacote em relação ao tamanho total do pacote original.

Na linha seguinte temos o campo TEMPO DE VIDA (TTL), que controla o tempo máximo que o pacote pode permanecer na rede. Ao gerar um pacote, o transmissor coloca um valor inicial nesse campo. A cada roteador ou máquina por onde esse pacote passa, esse valor é decrementado. Caso o valor do campo chegue a zero, o roteador deverá descartar o pacote. Em seguida, o campo PROTOCOLO especifica qual o protocolo da camada de transporte criou a mensagem que está sendo transportada no pacote.

O próximo campo é VERIFICAÇÃO DA SOMA DE CABEÇALHO (*checksum*) que é utilizado para verificar se os valores do cabeçalho do pacote recebido estão íntegros. Diferentemente do campo de verificação dos protocolos TCP e UDP, o cálculo é feito apenas sobre o cabeçalho, não sendo utilizado o *payload*. Esse campo é atualizado por todos os roteadores que manipulam o pacote, pois, como visto, eles fazem alterações no cabeçalho IP, ao decrementar o TTL.

Os campos seguintes são os ENDEREÇOS IP de origem e destino. Esses campos são formados por quatro bytes cada e identificam unicamente o emissor e o destinatário do pacote IP. O ENDEREÇO IP contém duas informações: parte dele endereça a rede, ou seja, identifica a qual rede o equipamento pertence e a segunda parte identifica unicamente o equipamento na rede.

Por último temos os campos OPÇÕES e ENCHIMENTO. O primeiro é utilizado basicamente para especificar funcionalidades opcionais. E o segundo serve apenas para garantir que cabeçalho IP seja múltiplo de 32 bits, já que as opções podem ter comprimento variável.

Aspectos Relevantes para a Anonimização

Dos campos apresentados acima, alguns podem ser usados para a quebra de privacidade e/ou segurança da rede. Por exemplo, o campo TIPO DE SERVIÇO é utilizado por algumas ferramentas na identificação de sistema operacional (*OS FingerPrint*). A forma como esse campo é usado tem certa relação com o tipo de sistema operacional.

Outros campos do cabeçalho IP que são importantes para este tipo de ataque são os campos de comprimento total, identificação, o bit de não fragmentação do campo flags e o tempo de vida (TTL).

Como visto, o campo de verificação da soma do cabeçalho (*checksum*) do protocolo IP se diferencia do TCP, por não estar presente no seu cálculo os dados do pacote. Entretanto, como ele também é calculado com dados do cabeçalho, é possível usá-lo para extrair alguma informação do cabeçalho e isto pode afetar a segurança da rede ou o anonimato do usuário.

Como no protocolo TCP, esse campo também é importante para os pesquisadores, pois é possível extrair a informação de número de pacotes que são recebidos com erros. A diferença é que no caso do protocolo IP, os dados usados para o seu cálculo geralmente estão disponíveis e, portanto, é possível verificar se ele foi gerado com erro ou não. Dessa forma, a ferramenta de anonimização poderá analisar se o campo está correto e regerá-lo a partir do cabeçalho anonimizado, mantendo tanto a informação sobre se o pacote foi entregue corretamente, quanto a segurança da rede e anonimato do usuário.

Por terem a função de identificar unicamente a origem e destino do pacote, os campos de endereço IP de origem e de destino, contêm claramente informações sensíveis à privacidade e à segurança. Como visto, esses endereços, além de identificar unicamente um equipamento, podem fornecer informações sobre a frequência de acesso de um determinado equipamento ou de uma determinada rede, portanto são muito importantes para os pesquisadores. Devido à sua importância existem vários estudos para se encontrar um meio-termo para a questão de anonimização de endereços IP, onde se consiga preservar as informações interessantes para a pesquisa, sem que essas informações afetem a privacidade e a segurança da rede. Pela importância desse ponto, a seção 4.3 trata desse assunto com maiores detalhes.

4.1.3.2 IPv6

Devido à limitação de números IP na versão IPv4, foi proposta uma nova versão para o protocolo IP que é o IPv6, onde foram feitas várias mudanças no cabeçalho. A principal é o aumento do tamanho do endereço IP, que passa de 32 para 128 bits. Apesar dessa mudança, será necessário o mesmo tipo de análise da versão anterior, e os métodos de anonimização deverão ser os mesmos, considerando apenas o aumento do tamanho do campo.

Outra modificação foi no campo opções, mas como no endereço IP esse campo precisará ainda de anonimização, pois a nova versão prevê opções com valores padrão, que poderão causar o mesmo problema na identificação do sistema operacional, caso

estes alterem o valor padrão. Também no IPv6, foi criado o campo que provavelmente necessitará ser anonimizado que é o identificador de fluxo, pois poderá conter dados que identifiquem a origem e o destino do pacote.

Muitos dos demais protocolos da camada de rede e outros de aplicação possuem versões adaptadas para utilizar IPv6. Do ponto de vista deste trabalho, entretanto, essas alterações não trazem novos elementos e nas seções a seguir discutimos apenas a versão associada ao IPv4. Pode-se considerar que os mesmos problemas e soluções se aplicarão no caso do IPv6.

4.1.3.3 ARP

O protocolo ARP permite que uma máquina identifique o endereço físico de um *host* de destino na mesma rede física. Para isso a origem envia uma mensagem *broadcast* ARP pela rede física solicitando o endereço de hardware do equipamento (também chamado de *MAC Address*) que possui determinado endereço IP. Todas as máquinas recebem a mensagem, mas apenas o equipamento com aquele IP responde a mensagem incluindo o seu endereço de hardware.

Para realizar essa tarefa o protocolo ARP possui um formato de mensagem com vários campos, para que ele possa ser útil para diferentes tecnologias de rede onde os campos de endereços podem ter tamanhos variados. O campo TIPO DE HARDWARE identifica o tipo de hardware para o qual o transmissor espera uma resposta, por exemplo, Ethernet. Da mesma forma, o campo tipo de protocolo identifica o protocolo do nível de rede cujo endereço está sendo usado. Assim, a funcionalidade dos campos HLEN e PLEN é permitir a adaptação a várias tecnologias de rede, pois eles especificam o tamanho do endereço do hardware e do endereço de protocolo, respectivamente. O campo OPERAÇÃO é usado para identificar o tipo de operação do protocolo, por exemplo, solicitação, resposta, etc.

Por fim, temos os campos SENDER HA, que identifica o endereço de *hardware* do emissor, SENDER IP que identifica o endereço IP do emissor, TARGET HA que identifica o endereço de *hardware* do destinatário e finalmente, o TARGET IP que identifica o endereço IP do destinatário.

Aspectos Relevantes para a Anonimização

Os campos SENDER HA e TARGET HA são os endereços de *hardware* do emissor e destinatário do pacate no padrão IEEE 802, usados em tecnologias como Ethernet,

redes sem fio e outras, são formados por seis *bytes*. Os três primeiros identificam um lote de endereços que pode ser comprado pelos fabricantes do *hardware*. Sendo assim, com os três primeiros é possível identificar o fabricante do equipamento, o que pode representar uma ameaça à privacidade e segurança caso o *hardware* utilizado seja muito específico. A segunda parte do endereço é um número único dentro da numeração do lote que identifica cada unidade fabricada.

Esse campo costuma ser importante para quaisquer análises, pois pode indentificar cada equipamento, ou identificar erros em produtos de um determinado fabricante. Devido a essas caracteríticas, as técnicas anonimização devem tentar preservar ao máximo essas informações importantes, entretanto, sem revelar o fabricante do equipamento ou o número original de cada dispositivo.

Já os campos SENDER IP e TARGET IP são os endereços IP da origem e do destino da mensagem ARP. Eles identificam o endereço IP de origem e destino do pacote ARP. Como discutido anteriormente, temos um interesse especial por esse campo, pois é o mesmo endereço do protocolo IP, ou seja, é muito importante que esses campos, ao serem anonimizados, não deixem vestígios dos seus valores originais. Se possível, a anonimização desses campos deve seguir o mesmo tipo de anonimização dos endereços dos pacotes do protocolo IP.

4.1.3.4 ICMP

Como os pacotes IP são trocados com base na política de melhor esforço, sem confirmação de entrega ou conexão, quando há algum erro o protocolo IP não possui nenhum recurso para comunicar à origem do pacote que algo está errado. Para isso existe o protocolo *Internet Control Message Protocol* (ICMP), que tem como função informar erros entre os elementos de conexão da rede e permitir a troca de mensagem de controle.

O pacote ICMP fica dentro da área de dados do IP e o formato de seu cabeçalho varia de acordo com o tipo de mensagem que ele está enviando. As mensagens ICMP devem ser tratadas com cuidado, pois além de possuírem os campos do IP, possuem no seu cabeçalho campos como a soma de verificação que é calculada com o próprio pacote. Além disso, as mensagens de erro causadas por um certo pacote IP levam em seu *payload* os 64 primeiros bits daquele pacote. Em outros tipos de mensagem, pode ser incluído o endereço do roteador que a origem deve enviar o pacote IP.

Devido às informações que estão no pacote ICMP e à variadade de tipos de ferramentas de anonimização deve-se ter muito cuidado ao tratar o ICMP. Descartar simplesmente o pacote pode não ser uma opção, pois eles também podem ser importantes para a análise, pois ajudam a identificar problemas na rede. Por outro lado,

o ataque de injeção de dados, explicado na seção 2.3, usualmente se baseia no envio pelo atacante de mensagens ICMP e pode permitir que o adversário consiga identificar determinada máquina. A necessidade de retirar isso pode ser um fator importante que justifique os pacotes que podem configurar um ataque dos *logs* anonimizados.

4.1.3.5 Protocolos de Roteamento

Em uma análise usual de uma rede é comum assumir que os pacotes são sempre roteados adequadamente para seu destino, presumindo que os roteadores conhecem de antemão todos os destinos dos pacotes que trafegam pela rede. Isso na prática não ocorre, e a função de roteamento de pacotes pelo melhor caminho é difícil de ser implementada, exigindo o uso de protocolos de roteamento para se identificar caminhos viáveis.

Para isso, foram criados diversos protocolos de roteamento, dentre eles destacam-se o `BGP`, `OSPF` e `RIP`. Cada um desses protocolos possui técnicas diferentes para tentar obter a informação do melhor caminho por onde o pacote deve ser repassado, montando a chamada tabela de roteamento. Para isso, eles possuem mensagens que podem ser usadas para obter informações de rotas que podem ser utilizadas pelos adversários. Em uma , por exemplo, para obter detalhes de uma topologia de rede de uma organização, o que interfere na segurança da mesma.

Mensagens de roteamento também podem possuir informação de endereços de *hardware* que podem, como visto anteriormente, interferir no anonimato. Entretanto, na maioria das vezes essa informação é relativa a um roteador da rede e não uma máquina de usuário, o que representa um problema menor do ponto de vista de anonimato.

Diante do exposto é interessante que esse tipo de pacote seja removido dos *logs* pelas ferramentas de anonimização.

4.1.4 Camada de Tecnologia de Rede Local

Essa camada, como definida originalmente na arquitetura TCP/IP, costuma ser dividida em camadas de enlace e camada física, tomando emprestadas as definições dessas camadas do modelo OSI/ISO [1]. Entretanto, para os fins deste trabalho adotamos a definição original da arquitetura TCP/IP.

Na camada de rede local encontramos diversas tecnologias, como *token-ring*, FDDI, redes sem fio, enlaces PPP e Ethernet. Apesar dessa grande variedade, Ethernet é a tecnologia de maior penetração no momento atual e é a que tem maiores implicações em termos de questões de segurança e privacidade, por esse motivo ela será o foco desta discussão.

[1] http://www.iso.org/iso/home.htm

Além disso, nessa camada, o principal elemento para questões de anonimização é a noção de endereço físico, normalmente denominado endereço MAC em redes Ethernet. Esse tipo de endereço hoje é padronizado e compartilhado por todas as tecnologias de rede agregadas em padrões IEEE 802. Assim sendo, a discussão desses endereços se aplica a todas as tecnologias de rede dessa família.

O quadro Ethernet possui seis campos, o campo preâmbulo é usado para sincronização do receptor do sinal. Em seguida, temos os campos de endereço de destino e origem do quadro. Como discutido anteriormente, esse endereço possui 6 bytes é atribuído a cada dispositivo de rede.

Após os endereço, aparece o tipo de quadro, que identifica o tipo de protocolo do nível de rede receberá o quadro. Por último, o campo CRC (Código de Redundância Cíclica) tem como função a identificação de erros ocorridos durante a transmissão do quadro.

Aspectos Relevantes para a Anonimização

Como mencionado, os únicos campos que influenciam na privacidade e segurança da rede são os endereços de origem e destino do quadro. Esses campos, conforme a seção 4.1.3.3, são importantes para a pesquisa, pois é possível indentificar cada máquina de uma rede local, ou até mesmo saber se pacotes de determinado fabricante é gerado com algum problema. Assim sendo, se esses campos não forem anonimizados, é possível inferir de qual equipamento saiu determinado pacote e, consequentemente, descobrir a identidade do usuário associado a ela.

4.2 Técnicas de Anonimização de Dados

Diante da necessidade de se manter a privacidade dos dados e a segurança das redes, surgiram as chamadas ferramentas de anonimização de dados, que definem um conjunto de políticas e técnicas desenvolvidas para tentar garantir a privacidade dos usuários de redes e outros serviços, tentando preservar a qualidade das informações necessárias para o desenvolvimento de pesquisas, análises gerenciais e auditorias.

Uma solução aparentemente óbvia para garantir a privacidade é simplesmente excluir dos dados as informações consideradas sensíveis do ponto de vista de privacidade e segurança. Infelizmente, dessa forma pode-se destruir a qualidade dos dados para a pesquisa, pois eles, por exemplo, não poderão ser separados em função de suas origens e

destinos. Diante disso, pode ser necessário, ao invés de excluir as informações, substituí-las por outras que mantenham parte da informação, por exemplo, as características que separam os endereços IP em diferentes máquinas, apesar de não permitir sua identificação. Nesse caso, é necessário garantir que a partir desses identificadores não seja possível deduzir o valor original dos dados.

Para tentar anonimizar os dados garantindo que as informações sensíveis à segurança e ao anonimato sejam eliminadas, foram criadas várias técnicas de anonimização. Veremos a seguir, que há um compromisso envolvendo essas técnicas de anonimização, pois quanto melhor é a anonimização (no sentido do alto grau de dificuldade para reverter a anonimização) pior é a qualidade desses dados para a pesquisa.

4.2.1 Substituição por *Black Marker*

O nome *black marker* foi dado por Slagell et al. [2006]. Essa técnica é implementada pela maioria das ferramentas de anonimização, e tem como principal característica substituir as informações relevantes por um valor constante, equivalendo à exclusão das informações.

Essa técnica tem uma anonimização muito forte, por ser praticamente impossível para um adversário inferir a informação original, pois o único padrão dessa técnica é o valor usado como *black marker*. Por outro lado, essa técnica praticamente inutiliza o dado anonimizado para análise, pois perde-se qualquer correlação entre os dados. Por exemplo, com o uso do *black marker* não é possível correlacionar eventos entre os vários pacotes endereçados para uma mesma rede.

Existem algumas variações dessa técnica como, por exemplo, usar o *black marker* em partes do campo. Em um endereço IP pode-se anonimizar com essa técnica apenas os dois últimos bytes do campo, por exemplo. Outra variação é a técnica chamada de *truncation*, que ao invés de substituir a informação, apenas a elimina.

4.2.2 Substituição Aleatória

Como o próprio nome diz, essa técnica faz uma substituição de valores em todas as ocorrências de um mesmo campo por valores aleatórios, mantendo entretanto, a relação entre valores anonimizados e valores originais. Isto é, cada valor encontrado pela primeira vez é substituído por um valor aleatório; novas ocorrências do mesmo valor original são sempre substituídos pelo mesmo valor já anonimizado. Basicamente o anonimizador se vale de uma tabela que vai sendo preenchida com os valores utilizados.

Essa técnica é interessante porque ela dificulta a identificação dos valores ori-

ginais, pois a anonimização é feita de forma aleatória, mas ainda mantém algumas características importantes para a análise, pois permite que as distribuições de valores em cada campo se mantenham, mas está sujeita a ataques de injeção de dados, descritos na secão 2.3.

4.2.3 Criptografia

A criptografia é utilizada nas mais diversas áreas para a segurança e sua característica é substituir informações existentes por outras, mantendo um mesmo padrão de substituição, mas diferentemente da substituição aleatória, os valores originais dos campos não são substituídos de forma aleatória e sim gerados através de um chave criptográfica. Dessa forma, se a mesma chave for usada várias vezes o valor original sempre será anonimizado pelo mesmo valor, facilitando o processamento paralelo dos arquivos, pois em todos eles os valores serão anonimazados da mesma forma.

Este tipo de anonimização mantém o dado com um certo nível de qualidade para a pesquisa, porque se as informações iguais são sempre substituídas por outras também iguais, é possível fazer uma correlação entre os vários pacotes de um arquivo de log. O problema é que, havendo essa correlação, esse log fica mais suscetível a ataques de injeção de dados, descritos na seção 2.3.

4.2.4 Deslocamento

A técnica de deslocamento consiste em somar ao valor a ser anonimizado um valor fixo (às vezes combinado a um pequeno desvio aleatório), alterando em todo arquivo de *log* determinado campo com a mesma variação. Essa técnica nos lembra a criptografia, pois mantém o mesmo valor para os campos iguais, mas diferentemente daquela, os valores são sempre gerados a partir de um valor fixo somado ao valor do campo. Essa anonimização geralmente é utilizada em campos relativos a tempo.

4.2.5 Preservação de prefixos

A técnica de preservação de prefixos consiste em usar uma substituição aleatória ou com criptografia, porém preservando as relações entre prefixos dos dados originais. Em alguns campos da arquitetura TCP/IP o mesmo campo pode conter duas informações diferentes, por exemplo, no campo `endereço IP`, os primeiros *bytes* identificam a rede a qual pertence determinada máquina e os *bytes* finais identificam a máquina univocamente na rede.

Segundo Slagell et al. [2006] dois endereços anonimizados que compartilham um prefixo de n bits só foram anonimizados com a técnica de preservação de prefixos, se e somente se os endereços originais (não anonimizados) compartilham o prefixo de n bits.

4.3 Anonimização de Endereços IP

O endereço IP, por identificar univocamente uma equipamento na Internet, é o campo mais visado em ataques, pois identificando uma máquina, um adversário pode identificar um servidor para invadir, ou pode ainda identificar individualmente um usuário, pois muitas das vezes uma máquina é utilizada apenas por uma pessoa. Além disso, endereços IP identificam também a rede de origem de um pacote e o fluxo de pacotes de uma conexão. Por isso, ele também é o campo mais utilizado para as pesquisas e análises. Assim sendo, quanto mais essa informação for preservada, maior a qualidade dos dados.

Devido a esse compromisso, o endereço IP tornou-se o campo de maior visibilidade entre os pesquisadores de técnicas de anonimização, pois é muito importante anonimizá-lo de forma segura e ao mesmo tempo garantindo que suas características principais se matenham para que a pesquisa não seja prejudicada.

A anonimização de endereço IP usando a técnica de *black marker* é muito segura com relação à identificação da rede ou dos endereços que aquele *log* contém. Mas tem um efeito drástico quanto à análise, pois sem a idenficação de origem e destino dos pacotes, não se consegue extrair quase nenhuma informação dos dados.

Já a substituição aleatória e a criptografia são utilizadas na anonimização de endereço IP para tentar manter as características que ele possui, sem afetar a privacidade e segurança da rede. A grande diferença entre os dois métodos é que quando se usa a mesma chave na criptografia, a transformação de determinada informação sempre terá o mesmo resultado; diferentemente da substituição aleatória que gera uma valor diferente a cada vez que é utilizada. Sendo assim, na substituição por criptografia é possível identificar que dois pacotes foram gerados por um mesma origem. O problema é que, dependendo do tipo de chave sendo utilizada para anonimizar ou até mesmo o tipo de ataque, é possível inferir o valor original do dado.

Na anonimização de endereços IP a técnica mais recomendada é a preservação de prefixos. Isso se deve à importância do endereço IP para as pesquisas e o fato dele ser normalmente tratado de um forma hierarquizada com base em prefixos. Essa técnica consiste em manter entre dois endereços que possuem um determinado número de bits

iniciais iguais, a mesma quantidade de bits iguais após a anonimização. Sendo assim, os bits iniciais que dois endereços IP compartilham são sempre transformados de forma idêntica [Burkhart et al., 2008a].

Xu et al. [2002] define formalmente a anonimização com preservação de prefixo:

> "Dois endereços IP a = a_1 a_2...a_n e b = b_1 b_2...b_n compartilham um prefixo de kbits ($0 \leq k \leq n$), se a_1 a_2...a_k = b_1 b_2...b_k e $a_{k+1} \neq b_{k+1}$, onde k < n. Uma função de anonimização é dita ser de preservação de prefixo, se dado dois endereços IP a e b estes compartilham um prefixo de *kbits*, F(a) e F(b) também compartilham um prefixo de *kbits*".

Dessa maneira, a preservação de prefixo permite que os endereços IP mantenham a relação hierárquica entre si após a anonimização, permitindo a identificação de fluxos de pacotes de uma mesma rede, o que torna esse tipo de anonimização mais útil para os pesquisadores que utilizam essa informação, por exemplo, para entender o comportamento do roteamento [Xu et al., 2001].

Entretanto, o compromisso existente entre a qualidade da anonimização e a qualidade da informação restante, faz com que essa maior permanência de informação nos *logs* implique em uma maior fragilidade na privacidade dos dados e na segurança da rede, pois os dados ficam mais suscetíveis a ataques. Por exemplo, com a separação das máquinas ataques como *fingerprinting* [Ribeiro et al., 2008], ataques de injeção de dados [Slagell & Yurcik, 2004] e ataques que inferem as máquinas através do seu comportamento [Coull et al., 2007], se tornam mais visíveis em certos casos.

4.4 Ferramentas de Anonimização

Como vimos na seção 2.6, existem diversas ferramentas de anonimização. Algumas ferramentas são mais simples e aplicam técnicas de anonimização apenas em alguns campos, por exemplo, no endereço IP. Outras, além de anonimizar uma quantidade maior de informação, são mais flexíveis, permitindo até a implementação de novas funções que podem ser adicionadas à ferramenta. Nessa seção discutimos com mais detalhes os principais trabalhos nessa área.

4.4.1 Tcpdpriv

Desenvolvido em 1996, o `tcpdpriv` [Minshall, 1996] é uma das mais antigas e conhecidas ferramentas de anonimização. Ela anonimiza os dados coletados diretamente da

interface de rede ou através de arquivos de saída do `tcpdump`. Sua importância é tão grande que ela é citada em quase todos artigos sobre anonimização de dados.

`Tcpdpriv` remove os *payloads* dos pacotes UDP e TCP e remove toda a área de dados do protocolo IP para outros tipos de protocolos. Quanto aos demais protocolos, ela é capaz de gerar diversos níveis de anonimização, desde anonimização colocando zero nos valores dos campos (*black marker*) até a anonimização mantendo o mesmo endereço IP anonimizado para as várias ocorrências de um endereço IP no arquivo original e a técnica de preservação de prefixos. Além disso, para evitar a identificação do sistema operacional que gerou o pacote ela não preserva o campo de opções de TCP.

Xu et al. [2002] fazem uma crítica à técnica de preservação de prefixos implementada pelo `tcpdpriv` pois, segundo os autores, a chave geradora dos endereços IP anonimizados é aleatória então o `tcpdpriv` gera um tabela que relaciona os endereços IP originais e seus endereços anonimizados correspondentes. Essa técnica faz com que um mesmo endereço IP seja anonimizado com endereços diferentes a cada vez que a ferramenta é executada. Como *logs* de tráfego em geral são grandes eles tendem a ser armazenados em vários arquivos. O `tcpdpriv` não permite que se faça uma anonimização simultânea de várias partes de um mesmo *log*. A proposta de Xu et al. [2002] para solucionar esse problema é gerar os novos endereços através de uma chave criptográfica, ou seja, se a chave for mantida, um endereço IP será sempre transformado para um mesmo valor.

4.4.2 Crypto-Pan

Xu et al. [2002] implementam a técnica de anonimização de preservação de prefixo mencionada utilizando uma árvore *trie*, onde o endereço IP é transformado em uma seqüência de *bits* e cada bit é representado por um nodo na árvore. É usado então um algoritmo de criptografia que seleciona determinados nodos do *trie* para terem o seu valor invertido. Esse algoritmo é baseado em uma chave que poderá ser usada em diversas anonimizações, permitindo que se façam anonimizações paralelas, gerando valores iguais para os endereços anonimizados.

4.4.3 Tcpmkpub

Diante da tarefa de disponibilizar os dados do Lawrence Berkely National Lab (LBNL), um centro de referência em pesquisas com protocolos TCP/IP no início da Internet, Pang et al. [2006] enfrentaram dois problemas: o primeiro é que não encontraram nenhuma política de anonimização que atendesse suas necessidades e, após desenvolve-

rem a sua política, não encontraram nenhuma ferramenta que anonimizasse os *traces* de acordo com tal política. Foi então que desenvolveram o `tcpmkpub` para implementar a política de anonimização que tinham desenvolvido. Seu objetivo era fazer uma anonimização que mantivesse um equilíbrio entre retirar as informações no limite da segurança e privacidade, e manter o máximo de informação importante para a análise de tráfego.

O `tcpmkpub` é uma ferramenta de análise de dados coletados pelo tcpdump que não prevê anonimização de dados *online*. Ela disponibiliza aos usuários um *framework* para manipulação de pacotes de rede a partir de regras especificando cada um dos campos do cabeçalho, seu tamanho em bytes e a ação de anonimização desejada para o campo.

O nome do campo é formado através de um padrão de nomes para cada um dos campos: geralmente, o nome do protocolo, seguido do caracter *underline*, seguido de um nome indicativo do campo; por exemplo, IP_src representa o campo endereço de origem do protocolo IP.

O campo ação representa qual procedimento a ferramenta deve executar para aquele campo. O `tcpmkpub` fornece três opções: a primeira é *KEEP*, que mantém os dados originais do arquivo de *log*; outra opção é ZERO, que limpa as informações do campo (*black marker*). A terceira opção é fornecer o nome de uma função em `C++` que será responsável pela transformação do campo.

O `tcpmkpub` disponibiliza algumas dessas funções `C++` para alguns campos, como o endereço IP, opções de IP, etc; cada uma dessas funções executa um método de anonimização específico para os campos. Além disso, como discutido anteriormente, a ferramenta prevê a adição de novas funções em `C++`, permitindo que o usuário implemente a sua própria política de anonimização para determinado campo.

4.4.4 Framework for Log Anonymization and Information Management (FLAIM)

Possuir uma grande variedade de algoritmos de anonimização, ter suporte a muitos formatos de *log*, suportar vários níveis de anonimização e ter uma arquitetura modular extensível; estas são as quatros propriedades que, segundo Slagell et al. [2006], uma ferramenta de anonimização de logs deve possuir. O `FLAIM` foi desenvolvido para atender a todos esses aspectos.

A primeira característica preconiza ter uma grande variedade de algoritmos de anonimização. Como vimos na seção 4.2, devido ao compromisso entre privacidade e qualidade dos dados, foram desenvolvidas várias técnicas de anonimização, nas quais

algumas priorizam a privacidade sem se preocupar em manter a qualidade dos dados para a análise posterior e outras se preocupam em atingir um nível de privacidade que mantenha as informações relevantes para a análise.

O `FLAIM` possui, para cada campo do cabeçalho, diversos algoritmos de anonimização como, por exemplo, *black marker*, permutação randômica e preservação de prefixo. Além disso, ele possui o suporte a vários tipos de logs como, por exemplo, arquivos `tcpdump` e `netflow`. Além disso, a ferramenta possui a capacidade de se adaptar à política de anonimização do usuário, pois determinado usuário pode querer que um campo seja anonimizado e outro não; que um campo seja anonimizado utilizando o algoritmo *black marker* e que outro seja anonimizado utilizando a técnica de preservação de prefixo.

Por fim, FLAIM possui a capacidade de receber novas funcionalidades através de novos módulos de funções. Isto é, além dos módulos que ele já disponibiliza, é possível adicionar novos, que ampliem a sua capacidade de receber outros de tipos de arquivos de *logs* ou até mesmo adicionar novos algoritmos de anonimização.

O `FLAIM` consegue trabalhar com arquivos de *logs* estáticos ou com dados coletados *online*. Para anonimizar um *log* de acordo com sua política de anonimização, essa ferramenta utiliza um arquivo em formato XML onde o usuário pode especificar, para cada campo dos cabeçalhos da arquitetura TCP/IP, que tipo de anonimização deve ser executada.

Abaixo um exemplo de um do arquivo de configuração do FLAIM que exemplifica diversas técnicas de anonimização, discutidas logo em seguida.

```
<policy>
     <field name="IPV4_DST_IP">
          <BinaryPrefixPreserving>
               <passphrase>abracadabra</passphrase>
          </BinaryPrefixPreserving>
     </field>

     <field name="IPV4_SRC_IP">
          <BinaryBlackMarker>
               <numMarks>8</numMarks>
               <replacement>0</replacement>
          </BinaryBlackMarker>
     </field>

     <field name="TS_SEC">
```

```
        <RandomTimeShift>
            <lowerTimeShiftLimit>60</lowerTimeShiftLimit>
            <upperTimeShiftLimit>600</upperTimeShiftLimit>
            <secondaryField>NONE</secondaryField>
        </RandomTimeShift>
    </field>

    <field name="SRC_MAC">
        <BinaryRandomPermutation/>
    </field>

    <field name="DST_MAC">
        <BinaryBlackMarker>
            <numMarks>24</numMarks>
            <replacement>0</replacement>
        </BinaryBlackMarker>
    </field>

</policy>
```

O exemplo acima define seis campos a serem anonimizados, a configuração de cada campo é delimitada pelos códigos `<field NomedoCampo>` e `</field>`. O primeiro campo apresentado é o endereço IP de destino (IPV4_DST_IP) e é determinado a utilização da técnica de preservação de prefixo utilizando a palavra "abracadabra" como chave. Em seguida, temos o campo IP de origem (IPV4_SRC_IP) que será anonimizado a técnica de *black marker*, os códigos <numMarks> e <replacement> determinam o número de bits que serão anonimizados e o valor que será utilizado, respectivamente.

Em seguida temos o campo dos segundos do *timestamp* (TS_SEC). Para ele se utilizará a técnica de deslocamento com um componente aleatório: o deslocamento será um valor aleatório gerado entre 60 e 600. Depois temos o campo endereço de *hardware* de origem (SRC_MAC) que utilizará a técnica de permutação. Já o campo endereço de *hardware* de destino (DST_MAC) será anonimizado com a técnica de *black marker* em 24 bits, com o valor zero.

4.5 Conclusão

Neste capítulo foram analisados cada um dos campos dos cabeçalhos dos principais protocolos da arquitetura TCP/IP, sendo identificados os campos que podem compro-

meter a privacidade e/ou segurança da rede. Foram vistas as principais técnicas de anonimização de *logs*, tendo maior destaque a anonimização de endereços IP. Por fim, vimos algumas das ferramentas de anonimização e suas as técnicas que elas utilizam.

Capítulo 5

Metodologia Proposta

A metodologia proposta tem como objetivo auxiliar os administradores de rede na tarefa de coleta e disponibilização de dados de rede. Essa disponibilização pode ser solicitada pelos pesquisadores de uma empresa ou universidade ou até mesmo por uma ordem judicial, que solicitarão os dados de acordo com uma política de anonimização que os auxiliarão em alguma pesquisa ou processo.

Diante desse pedido, que provavelmente pode vir acompanhado com uma ferramenta de anonimização a ser utilizada ou das exigências sobre quais dados podem ou não ser ocultados (e como isso pode ser feito), o administrador de rede provavelmente não terá a certeza que a sua expectativa de anonimização será atingida e, principalmente, não saberá se os dados disponibilizados constituirão uma ameaça à segurança da sua rede ou se permitirão algum tipo de quebra da privacidade dos seus usuários.

Devido à grande quantidade de dados que esses *logs* de rede podem gerar, analisar esses dados manualmente para conferir se os dados foram anonimizados de acordo com a política solicitada, se torna uma tarefa impossível. Por isso, propomos o desenvolvimento de uma ferramenta que compara o arquivo original com o arquivo anonimizado e faça uma análise de quais dados foram anonimizados e qual o método utilizado. A seguir discutimos as características dessa ferramenta.

5.1 Arquitetura

A ferramenta proposta vem suprir a necessidade de automatizar a tarefa de análise dos arquivos anonimizados, conforme mostra a figura 5.1, a ferramenta desenvolvida tem como entrada o arquivo de *log* original e o mesmo arquivo anonimizado pela ferramenta de anonimização sugerida.

Em seguida, a ferramenta compara os pacotes encontrados nos dois arquivos. Essa

comparação avalia se os campos considerados sensíveis do ponto de vista da segurança da rede e/ou privacidade sofreram algum tipo de alteração. Se for detectada alguma mudança é provável que determinado campo sofreu algum tipo de anonimização, que precisa então ser qualificada.

Finalmente, um relatório deve ser produzido, descrevendo as conclusões da análise, identificando os campos que foram ou não anonimizados e as formas de anonimização utilizadas. Com base nessa informação o relatório deve também apresentar uma discussão do possível impacto de cada transformação aplicada (ou falta de tal transformação) para as políticas de privacidade e segurança da organização.

Idealmente, uma ferramenta que implemente essa arquitetura deveria ser configurável e extensível em cada uma dessas etapas. Novos módulos que verifiquem padrões específicos ou que analisem um novo protocolo não previsto originalmente deveriam poder ser adicionados de forma simples. Arquivos de configuração poderiam determinar exatamente quais testes deveriam ser aplicados e até definir de forma completa uma política de anonimização desejada. Nesse caso, o relatório poderia ser simplificado para indicar apenas se o arquivo atende ou não às exigências da política proposta. Na seção **??** será apresentado um protótipo desenvolvido com base na metodologia proposta, que realiza um conjunto de testes pré-definidos para demonstrar o conceito.

5.2 Fases da Metologia

A metodologia proposta para a ferramenta de análise de anonimização de *logs* propõe que o pacote seja analisado em todas as camadas da arquitetura TCP/IP (Tecnologia de Rede Local, Rede, Transporte e Aplicação).

5.2.1 Identificação dos Pares dos Pacotes

Primeiramente, os dois arquivos de entrada podem não conter exatamente os mesmos pacotes, pois filtros podem ser aplicados retirando completamente certos pacotes do arquivo original. Isso pode ser feito para restringir o foco, no arquivo a ser disponibilizado, a pacotes que atendam um certo critério (como "manter apenas tráfego HTTP"), ou por que certos pacotes podem ser considerados sensíveis demais para serem distribuídos (como pacotes de protocolos de roteamento, em certos casos). Nesse caso, deve-se fazer uma comparação entre os pacotes de cada arquivo para identificar o par de pacote a ser analisado. Para isto, a primeira análise deverá ser feita no campo de marca de tempo (*timestamp*) que o tcpdump inclui no arquivo.

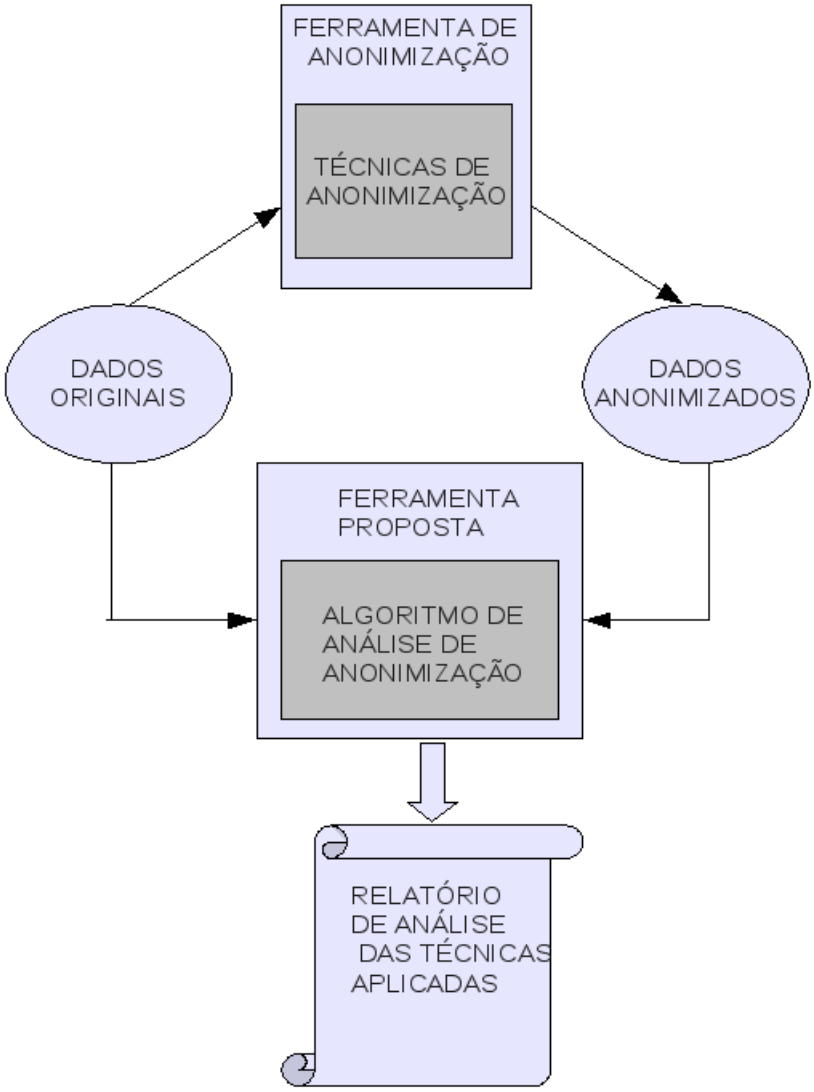


Figura 5.1. Funcionamento da Ferramenta Proposta

Não é suficiente, entretanto, realizar uma comparação direta dos valores desse campo nos dois arquivos, por dois motivos importantes: primeiro, em redes rápidas, devido à resolução limitada do campo de *timestamp* do tcpdump, diversos pacotes podem ter o mesmo tempo associado a eles; segundo, em certos casos, como discutido anteriormente, o próprio *timestamp* pode ser anomizado, seja com um deslocamento simples (igual) de todos os tempos, ou por um deslocamento com um componente aleatório.

Para resolver o primeiro problema, da unicidade dos tempos, outros campos de pouco interesse para a segurança e o anonimato podem ser usados na diferenciação dos pacotes, como os identificadores dos protocolos de enlace (rede local), rede e transporte,

que usualmente são mantidos. Para o problema do deslocamento dos *timestamps*, é necessário aplicar-se um algoritmo de casamento de padrões temporais, que busque identificar um casamento entre os pacotes que valide os intervalos de tempo entre eles. Isso é possível, desde que o deslocamento seja simples (onde basta encontrar o valor de deslocamento aplicado para que todos os pacotes se alinhem) ou que o valor aleatório adicionado seja pequeno em relação à maioria dos intervalos. (Esse é normalmente o caso, pois normalmente deseja-se apenas ocultar o momento exato em que o *log* foi coletado.

5.2.2 Camada de Tecnologia de Rede Local

Depois de identificar os pares de pacotes correspondentes, deve-se primeiro analisar o pacote no nível de rede local e verificar se o pacote é Ethernet (ou, basicamente, qualquer protocolo ISO 802); caso não seja, a ferramenta deve possuir regras para avaliação do protocolo específico indicado. Em um primeiro momento, recomenda-se apenas a contabilização desses pacotes, já que a maior parte do tráfego é hoje coletada em redes desse tipo. No caso de pacotes Ethernet, deve-se verificar os endereços de origem e destino dos quadros, verificando quantos pacotes foram anonimizados e o método utilizado para isso.

Além disso, caso o pacote seja ARP deve-se fazer uma verificação se os endereços foram anonimizados de forma idêntica aos endereços de *hardware* do pacote desse protocolo, caso um mapeamento de endereços preciso (mesmo que anonimizado) seja desejável.

5.2.3 Camada de Rede

Na camada de rede deve-se testar o tipo de pacote, ARP, IP, ICMP ou algum protocolo de roteamente. Caso o pacote seja ARP, como mencionado na seção anterior, deve-se verificar se há consistência entre a anonimização dos endereços de *hardware* com o endereços do pacote. Caso o pacote seja ICMP ou de algum protocolo de roteamento, a ferramenta deverá alertar sobre os riscos da presença desse tipo de pacote no arquivo anonimizado.

Se o pacote for IP, a ferramenta terá que analisar se houve algum tipo de anonimização em alguns dos campos. Para os campos TIPO DE SERVIÇO, COMPRIMENTO TOTAL, IDENTIFICAÇÃO, bit de não fragmentação do campo FLAGS e o campo TEMPO DE VIDA (TTL), deve-se fazer apenas uma verificação de quantos pacotes não tiveram esses campos anonimizados, pois, como vimos, esses campos são utilizados para

descobrir o sistema operacional da máquina.

Em seguida temos o campo *checksum*. Pode ser importante analisar se o pacote original tinha algum erro e se o checksum do pacote anonimizado foi alterado para mantê-lo correto (ou errado) como no pacote original, após a anoimização. Como discutido anteriormente, este campo é calculado com base em dados de outros campos do cabeçalho, o que pode em certos casos permitir a um atacante recompor dados originais que deveriam ter sido removidos.

Por último, deve-se analisar o endereço IP de origem e de destino. Essa análise deve ser minuciosa, devido à importância desses campos na identificação de algum usuário da rede. Se for verificada a existência de anonimização a ferramenta deverá identificar o método utilizado. Essa identificação pode exigir a coleta de dados sobre todos os endereços encontrados no *log*.

5.2.4 Transporte

Na camada de transporte, primeiramente deve ser identificado se o pacote é TCP ou UDP (um outro protocolo exigiria regras de processamento particulares e deveria ser claramente identificado no relatório). Caso ele seja UDP, deve ser verificado se houve anonimização do campo SOMA DE VERIFICAÇÃO (pela possibilidade de recuperação de informação sensível em alguns casos) e dos campos PORTA DE ORIGEM E DESTINO. Estes últimos podem ser importantes em análise de tráfego por protocolo, mas até essa informação pode ser anonimizada em certos casos, pois a identificação de protocolos usados pode levar à identificação de servidores ativos que podem ser atacados, constituindo-se em uma ameaça de segurança em certos casos.

Já no TCP, além dos mesmos testes vistos no UDP, deve-se testar também os campos número de seqüência, janela e opções do TCP para verificar se houve anonimização, pois esses campos são utilizados pelos ataques de identificação de sistema operacional, entre outros.

5.2.5 Aplicação

Finalmente, na camada de aplicação deve-se testar se há algum *payload* no pacote anonimizado, pois os dados contidos nele podem revelar informações privadas dos usuários. Caso exista algum pacote com payload, mesmo que seja apenas uma fração dos dados originais, deve ser alertado do grande risco que essa informação pode trazer à privacidade das pessoas, caso o arquivo de *logs* seja disponibilizado para a análise. Uma análise mais detalhada de dados de aplicação é normalmente de difícil implementação,

pela grande variedade de aplicações possíveis e da interpretação dos dados de cada uma em termos de anonimato.

5.2.6 Análise da anonimização de endereços

Como discutido anteriormente, para os diversos tipos de endereços encontrados na arquitetura TCP/IP (como endereços de rede local, IP, números de portos) a análise do padrão de anonimização adotado exige a coleta de informações sobre todos os endereços encontrados. Idealmente, deve-se montar um mapeamento entre endereços encontrados no arquivo original e os endereços a eles associados no arquivo anonimizado. A partir daí, diversas observações devem ser feitas.

- Se algum endereço for encontrado no arquivo anonimizado sem transformação, isso deve ser claramente indicado no relatório, pois pode consitituir uma falha do processo de anonimização.

- Se as relações entre os dois mapeamentos forem de um para muitos, os dados de identificação de máquinas individuais provavelmente foram removidos do arquivo. É importante, entretanto, uma verificação cuidadosa para determinar se o mesmo padrão se aplica a todos os endereços e que não há endereços que recebem tratamento diferenciado.

- Se for observada uma relação 1:1 entre os dois conjuntos, o processo de anonimização não utilizou a técnica de *blackmarker*. Isso pode ser útil na análise de comportamento de máquinas, mas pode constituir uma ameaça às políticas em alguns casos. Para se verificar se foi utilizada uma técnica de preservação de prefixos, pode-se montar um *trie* binário para cada um dos dois conjuntos e verificar a equivalência da topologia de ambos (e da localização das chaves) [Xu et al., 2002].

- É interessante verificar-se a distribuição estatística dos endereços encontrados nos dois conjuntos, por exemplo, pode-se utilizar a função de distribuição cumulativa (CDF) para identificar a frequência com que os endereços aparecem em cada conjunto.

- Caso se conheça o prefixo da rede da organização onde o tráfego foi coletado (normalmente disponível se a ferramenta for aplicada no momento da coleta) é interessante fazer o tratamento separado dos endereços da própria organização e dos endereços externos. Algumas ferramentas podem usar técnicas diferentes em cada caso; uma ferramenta maliciosa poderia mascarar alguns endereços e

não outros, por exemplo, para tentar extrair informações que comprometam a segurança da rede.

Diversas análise mais sofisticadas são ainda possíveis sobre os conjuntos de endereços coletados, como técnicas de avaliação da qualidade da informação disponível, técnicas de análise de correlação e similares, já discutidas anteriormente.

5.3 Protótipo

Durante o trabalho foi desenvolvido um protótipo da ferramenta proposta, seguindo os passos básicos da metodologia. O objetivo nesse caso era verificar a viabilidade de certos tipos de processamento, identificar os pontos mais complexos do processamento e colocar em prática os conceitos envolvidos.

Para o desenvolvimento, foram analisadas diversas plataformas para manipulação de arquivos de *log* de tráfego de redes considerando o formato `pcap` usado pelo `tcpdump`, hoje considerado um padrão para essa área. Existem diversas ferramentas que fazem a análise desses arquivos, mas todas com objetivos já bastante específicos que não poderiam se alteradas para os nossos objetivos. Procuramos então bibliotecas de programação que simplificassem o desenvolvimento de uma nova ferramenta. Apesar de haver bibliotecas até para a linguagem C para esse fim, a característica hierárquica, em camadas, da arquitetura TCP/IP, faz com que o processamento dos diversos protocolos encapsulados nos pacotes coletados seja mais simples em uma linguagem orientada a objetos.

Bibliotecas orientadas a objetos para processamento de arquivos no formato `PCAP` se aproveitam do fato de que cada entrada do arquivo possui certos campos em comum, presentes em todos os pacotes (os campos de controle criados durante a coleta e os campos do cabeçalho do nível de rede local). Uma classe básica descreve então apenas esses campos e permite seu acesso direto a partir das entradas do arquivo. Com base nas informações dos protocolos dos níveis inferiores pode-se identificar o tipo do protocolo de cada camada superior. Para se analisar então os campos do protocolo de um novo nível, basta que se utilize então uma classe derivada da classe original, porém mais especializada para identificar os campos específicos do protocolo. Dessa forma a cada protocolo processado identifica-se o tipo do protocolo superior e promove-se o objeto contendo o pacote extraído do arquivo para uma classe mais específica que detalha cada protocolo.

Bibliotecas com hiearquias de classes desse tipo existem para diversas linguagens, como `Perl`, `Python`, `Ruby`, `C++` e `Java`, entre outras (e, muitas vezes, diversas bibliotecas

diferentes para cada linguagem). Inicialmente experimentamos com bibliotecas para `Python` [pylibpcap, 2009] e `Ruby` [rubypcap, 2009], mas a decisão final foi adotar a Java, com a biblioteca `jpcap` [Jpcap, 2009] para o desenvolvimento do protótipo. Essa combinação ofereceu o melhor compromisso entre aspectos como documentação, poder de expressão e simplicidade de utilização.

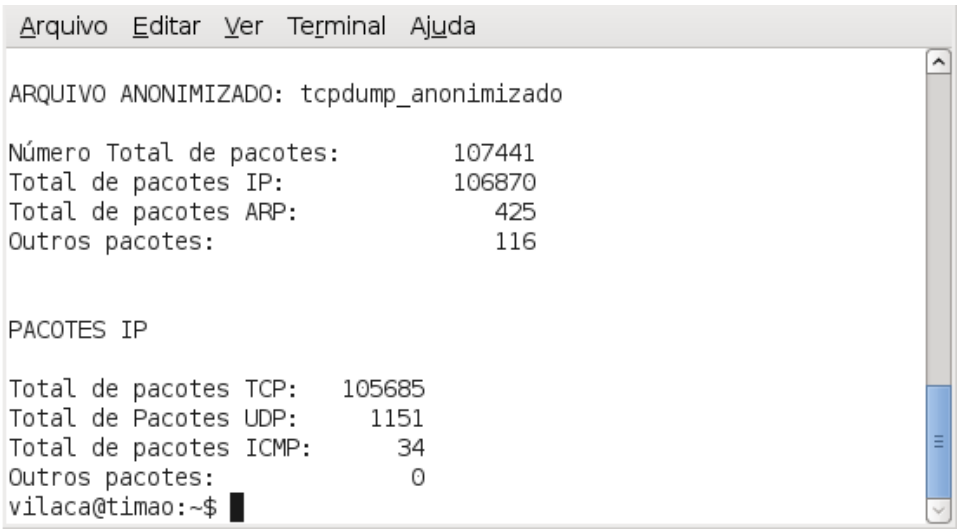
O protótipo desenvolvido segue a metodologia proposta e tem como entrada os nomes dos arquivos a serem analisados. O primeiro teste que é executado é o que identifica se os pacotes tratados nos dois arquivos são os mesmos e ele faz isso com base no campo de tempo do `tcpdump`. A seguir, é mostrada a parte da função desenvolvida que identifica a diferença de segundos existente entre os pacotes iguais dos dois arquivos, retornando esse valor para o programa, além do tempo, o tipo de pacote e o número de sequência podem ser verificados para garantir maior confiabilidade no resultado.

```
import jpcap.packet.*;
public class DiferencaSegundos2
{
  public long dif_sec;
  public void CalculaDiferenca(JpcapCaptor jpcap_real, JpcapCaptor jpcap_anon)
throws Exception
  {
    while(true)
    {
      Packet packet_anon=jpcap_anon.getPacket();
      if(packet_anon==null || packet_anon==Packet.EOF) break;
while(true)
{
  Packet packet_real=jpcap_real.getPacket();
  if(packet_real==null || packet_real==Packet.EOF) break;
  if(packet_anon.sec==packet_real.sec)
  {
    if(packet_anon.usec==packet_real.usec)
    {
      dif_sec = 0;
      break;
    }
    else break;
  }
  if(packet_anon.sec!=packet_real.sec)
  {
```

```
if(packet_anon.usec==packet_real.usec)
{
  dif_sec = packet_anon.sec-packet_real.sec;
  break;
}
```

Também são verificados em todas as camadas da arquitetura TCP/IP se os campos identificados como campos passíveis de recuperação de dados privados ou que comprometam a segurança da rede, vistos na seção 4.1, foram realmente anonimizados. Isso é feito comparando cada um desses campos do arquivo original e do arquivo anonimizado, e ao final é gerado um relatório com as estatíticas dos campos anonimizado, conforme a figura 5.2.

a)



b)

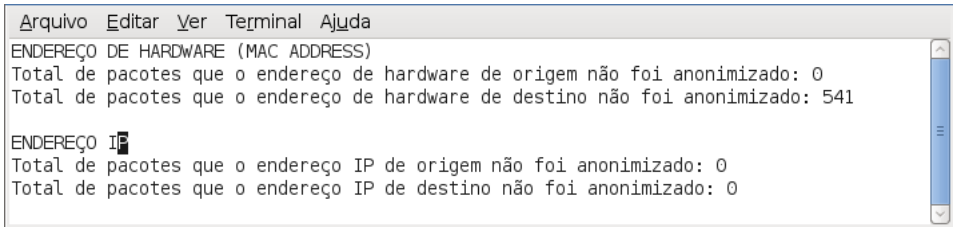


Figura 5.2. Relatórios do Protótipo: a) Quantidade de pacotes por protocolo; b) Endereços de hardware e endereços IP não anonimizados

Além disso, o protótipo analisa mais detalhadamente o endereço IP, que é o campo

com maior risco de identificação de usuário. Ele identifica se houve anonimização nos endereços IP e se foi utilizada anonimização por *black marker* e relata a quantidade de endereços que foram anonimizados utilizando essa técnica. A seguir, é mostrado uma parte do código que faz a análise do endereço IP para identificar se houve ou não anonimização nesse campo. Nesse código vemos que os endereços IP são armazenados em uma coleção de hash, tornando mais fácil a identificação de ocorrências de anonimizações diferentes para o mesmo endereço IP.

```
if (dlp_anon.frametype == EthernetPacket.ETHERTYPE_IP)
{
  IPPacket ipp_anon = (IPPacket)packet_anon;
  IPPacket ipp_real = (IPPacket)packet_real;

  if (ipp_anon.src_ip.equals(ipp_real.src_ip))  //IP SRC nao foi anonimizado
  {
    qtd_ipsrc_naoanon=qtd_ipsrc_naoanon+1;
  }
  else      //IP SRC foi anonimizado
  {
    ipsrc_real = ipp_real.src_ip.getHostAddress();
    ipsrc_anon = ipp_anon.src_ip.getHostAddress();
    Collection<String> col = new HashSet<String>();
  }
```

Ao final, além das estatíticas de análise dos arquivos, o protótipo imprime um relatório indicando para cada campo que não foi anonimizado quais os riscos envolvidos na disponibilização daquela informação.

5.4 Conclusão

Nesse capítulo foi proposta uma metodologia que inclui o desenvolvimento de uma ferramenta que analisaria o arquivo de *log* original e o arquivo após a anonimização, o principal objetivo dessa ferramenta é facilitar a tarefa do administrador de rede que precisa disponibilizar arquivos de *logs* para a pesquisa verificando se a informação contida em determinados campos que pode afetar a segurança da rede e/ou a privacidade dos usuários foi anonimizado ou não.

Em seguida, é feita uma análise em cada camada da arquitetura TCP/IP, indicando quais campos devem ser comparados e o qual o tipo de análise deve ser feito. Por fim, é apresentado um protótipo dessa ferramenta que tenta comprovar a eficiência e utilidade da ferramenta proposta.

Capítulo 6

Conclusão e Trabalhos Futuros

O uso da Internet cresce a cada dia e ao lado desse aumento cresce também a necessidade, por parte de auditores e pesquisadores, de usar os *logs* de tráfego de rede para propor novas soluções ou analisar situações que coloquem em risco a rede de uma empresa ou até mesmo o bom funcionamento da Internet. Para que essas pesquisas sejam mais confiáveis o ideal é que se utilizem dados diversificados e para isso é necessário a troca desses dados entre as entidades de pesquisa.

Por outro lado, cresce também a preocupação com a circulação de dados com informações privadas, pois é cada vez maior o número de banco de dados com informações pessoais nas empresas. Diante disso, em um primeiro momento, diversos países começaram a legislar sobre o tratamento e a troca dos dados pessoais, regulamentando o uso e manutenção dos mesmos. Com aumento de fraudes e crimes praticados através da Internet, esses países passaram a se preocupar com a manutenção e uso desses dados para ajudar a solucionar esses delitos.

Dessa forma, os administradores de redes enfrentam um dilema, onde a necessidade de uso e troca de dados de conexão se torna cada dia maior e por outro lado as legislações limitam cada vez mais a divulgação de dados que contenham informações pessoais.

Diante disso, este trabalho apresentou um estudo no qual foram analisadas as características técnicas do tráfego de rede IP sob a ótica da privacidade e segurança de rede, e também foi feita uma pesquisa e análise das legislações vigentes em alguns países e, é claro, no Brasil, sobre o controle dos dados em geral e mais especificamente sobre os dados de conexões de rede. Após isso, foi proposta uma metologia e apresentada um protótipo de uma ferramenta que auxilie o profissional a identificar se os dados que pretende disponibilizar foram anonimizados da forma desejada.

Este trabalho, em um primeiro momento apresentou, a dificuldade em lidar por

um lado com a necessidade de disponibilização de dados para a pesquisa e por outro com a preocupação de fragilizar a segurança da rede e expor a outrem dados privados dos usuários da rede. Depois foram mostrados os principais conceitos que envolvem o assunto, por exemplo, privacidade, arquitetura TCP/IP, coleta de dados e anonimização.

Em seguida, foram analisadas as legislações existentes em alguns dos principais países do mundo, demonstrando o processo histórico de formação da legislação atual. No Brasil, além dessa visão, foram mostradas mais detalhadamente as leis existentes e também as tendências de evolução das mesmas diante dos principais projetos de lei em tramitação. Foi visto que as primeiras legislações se mostraram preocupadas em proteger a privacidade das pessoas, mas com o crescimento da criminalidade através da Internet a tendência atual é o surgimento de leis que tenham um maior controle no uso da rede, mas mantendo a preocupação de preservar, dentro de um certo limite, a privacidade nas comunicações eletrônicas. No Brasil, foi mostrado que apesar de não termos uma legislação específica para a comunicação eletrônica de dados, as leis existentes de certa forma já resguardam a privacidade dos usuários, principalmente se equipararmos a comunicação de dados à comunicação telefônica com as sua legislação e regulamentos existentes. Também foi mostrado que, diante do projeto aprovado em 1^o turno no Senado Federal, a tendência no Brasil é que a nossa legislação evolua seguindo o modelo europeu, mantendo, é claro, algumas particularidades nacionais.

Após isso, foram descritas todas as estruturas existentes da arquitetura de maior uso na Internet, o TCP/IP, juntamente com uma análise dos campos dos protocolos que podem influenciar na privacidade ou segurança de uma rede. Identificando de acordo com a bibliografia pesquisada, todos os campos que de alguma forma interferem na identificação e privacidade das pessoas, bem como na segurança da rede. Em seguida, vimos que diante da necessidade de disponibilização dos arquivos de *logs*, foram criadas diversas técnicas e ferramentas que tornam os dados anônimos, tentando preservar a utilidade desses dados para as análises.

Finalmente, diante de toda a pesquisa feita, foi proposta uma metodologia e em seguida apresentamos um protótipo de uma ferramenta baseada nessa metodologia, para auxiliar os administradores a disponibilizar os arquivos de dados de conexão, informando se o dado foi anonimizado, quais informações foram anonimizadas e qual a técnica utilizada e seus riscos. Em seguida, foi apresentado o protótipo de uma ferramenta desenvolvida com base nos preceitos da metodologia proposta.

Como trabalhos futuros, uma possibilidade é o aprofundamento do estudo da legislação na área, especialmente considerando-se que estamos em um período de bastante atividade no congresso em assuntos relacionados à área. A evolução do entendimento

das questões envolvidas no uso da Internet entre os membros da sociedade pode fazer com que leis mais específicas sejam desenvolvidas, nos casos ainda não previstos na legislação atual.

A expansão do IPv6 é outro aspecto que deve ser considerado. Apesar de conceitualmente não haver elementos significativamente novos na operação da rede com IPv6, os endereços, por exemplo, ganham novo formato e serão distribuídos de forma ainda a ser completamente definida. Isso pode levantar problemas com a manutenção de anonimato em relação a endereços que hoje ainda não existem.

O protótipo desenvolvido é apenas uma prova de conceito para ferramenta e a metodologia de verificação propostas neste trabalho. Uma linha clara de ação seria o desenvolvimento de uma ferramenta completa, aproveitando melhor recursos de configuração e extensão dinâmicas para criar uma ferramenta que possa ser distribuída para uso pela comunidade. Em particular, seria necessário desenvolver um formato (linguagem) para a descrição do que seriam políticas aceitáveis de anonimização e divulgação de dados, de forma que a ferramenta, ao invés de gerar um relatório final com recomendações de pontos a serem considerados pelo administrador, gerasse um relatório simplificado, simplesmente indicando quais pontos da política estariam sendo observados/violados pela anonimização sendo considerada.

Referências Bibliográficas

Allman, M. & Paxson, V. (2007). Issues and etiquette concerning use of share measurement data. In ACM, editor, *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 135–140.

Arthur, D. & Panigrahy, R. (2006). Analyzing bittorrent and related peer-to-peer networks. In *SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 961--969, New York, NY, USA. ACM Press.

Bianchi, G.; Boschi, E.; Gaudino, F.; Koutsoloukas, L.; Lioudakis, G.; Rao, S.; Ricciato, F.; Schmoll, C. & Strohmeier, F. (2008a). Privacy-preserving network monitoring: Challenges and solutions. Disponível em http://www.salzburgresearch.at/research/gfx/mobsum08-cameraready.pdf.

Bianchi, G.; Teofili, S. & Pomposini, M. (2008b). New directions in privacy-preserving anomaly detection for network traffic. In *Proceedings of the 1st ACM workshop on Network data anonymization*, pp. 11–18. ACM.

Bishop, M.; Crawford, R.; Bhumiratana, B.; Clark, L. & Levitt, K. (2006). Some problems in sanitizing network data. In Society, I. C., editor, *Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pp. 307–312.

Blanton, E. (Acessado em 01/09/2009). Tcpurify disponível em http://irg.cs.ohiou.edu/ eblanton/tcpurify/.

Burkhart, M.; Brauckhoff, D. & May, M. (2008a). On the utility of anonymized flow traces for anomaly detection. In *Proceedings of the 19th ITC Specialist Seminar on Network Usage and Traffic (ITC SS)*.

Burkhart, M.; Brauckhoff, D.; May, M. & Boschi, E. (2008b). The risk-utility tradeoff for ip address truncation. In ACM, editor, *Proceedings of the 1st ACM workshop on Network data anonymization*, pp. 23–30.

Congresso Nacional, B. (1940). Código penal brasileiro.

Congresso Nacional, B. (1988). Constituição da república.

Congresso Nacional, B. (1996). Lei 9296 de 1996. Regulamenta o inciso XII do artigo 5o da Constituição da República de 1988.

Coull, S.; Wright, C.; Monrose, F.; Collins, M. & Reiter, M. (2007). Inferring sensitive information from anonymized network traces. In *Proceedings of the 15th Annual Network & Distributed System Security Symposium (NDSS 07)*, pp. 35–47.

Coull, S. E.; Wright, C. V.; Keromytis, A. D.; Monrose, F. & Reiter, M. (2008). Taming the devil: Techniques for evaluation anonymized network data. In *Proceedings of the 16th Annual Network & Distributed System Security Symposium (NDSS'08)*.

Delmanto, C.; Delmanto, R. & Júnior, R. D. (1998). *Código Penal Comentado*. Renovar.

Ferreira, A. B. d. H. (2008). *Mini Aurélio*. Positivo.

Gattani, S. & Daniels, T. E. (2008). Reference models for network data anonymization. In ACM, editor, *Proceedings of the 1st ACM workshop on Network data anonymization*, pp. 41–48.

Gomes, L. F. & Cervini, R. (1997). *Interceptação Telefônica - Lei 9296, 24-07-96*. Revista dos Tribunais.

Greco Filho, V. (1996). *Interceptação Telefônica*. Saraiva.

Hussain, A.; Heidemann, J.; Bartlett, G.; Papadopoulos, C.; Pryadkin, Y. & Bannister, J. (2006). Experiences with a continuous network tracing infrastructure. In *ACM SIGCOMM 05 Workshops*.

Ipsumdump (Acessado em 01/09/2009). Disponível em http://www.cs.ucla.edu/ kohler/ipsumdump/.

Jesus, D. E. d. (1997). *Direito Penal*, volume 2. Saraiva.

Jpcap (Acessado em 01/09/2009). Disponível em http://netresearch.ics.uci.edu/kfujii/jpcap/doc/.

Keardsri, W.; Teng-amnuay, Y. & Prathombutr, P. (2009). Defining privacy leves for ip address anonymization. In *13o International Symposium on Computational Science and Engineering (ANSCSE 13)*.

Kelly, D. J.; Baldwin, R. O.; Raines, R. A.; Grimaila, M. R. & Mullins, B. E. (2008). A survey of state-of-the-art in anonymity metrics. In *NDA 08*.

King, J.; Lakaraju, K. & Slagell, A. (2009). A taxomony and adversarial model for attacks against network log anonymization. In *SAC 09*.

Kohno, T.; Broido, A. & Claffy, K. C. (2005). Remote physical device fingerprinting. In *Proceedings of the IEEE Symposium on Security and Privacy*.

Koukis, D.; Antonatos, S. & Anagnostakis, K. G. (2006). On the privacy risks of publishing anonymized ip network traces. In *Proceedings of the Conference on Communications and Multimedia Security*.

Kuenning, G. & Miller, E. L. (2003). Anonymization techniques for urls and filenames. In *Technical Report UCSC-CRL-03-05, University of California*.

Luo, K.; Li, Y.; Ermopoulos, C.; Yurcik, W. & Slagell, A. (2006). Scrub-pa: A multi-level multi-dimensional anonymization tool for process accounting. In *Technical Report cs.CR/0601079, ACM Computing Research Repository (CoRR)*.

Mello, M. A. (1992). Voto sobre petição 577 - acórdão stf. http://www.stf.jus.br/portal/jurisprudencia/listarJurisprudencia.asp?s1=Pet-QO.SCLA.+E+577.NUME.&base=baseAcordaos.

Minshall, G. (1996). Tcpdpriv disponível em http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html em 01/09/2009.

Netflow (Acessado em 01/09/2009). Disponível em http://www.cisco.com/web/go/netflow.

Nmap (Acessado em 01/09/2009). Disponível em http://nmap.org/book/osdetect.html.

Ntop (Acessado em 01/09/2009). Disponível em http://www.ntop.org/.

Ohm, P.; Sicker, D. & Grunwald, D. (2007). Legal issues surrounding monitoring during network research. In ACM, editor, *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 141–148.

ONU (1950). Declaração universal dos direitos humanos.

Pang, R.; Allman, M.; Paxson, V. & Lee, J. (2006). The devil and packet trace anonymization. In *ACM SIGCOMM Computer Communication Review Archive*, volume 36, pp. 29--38.

Pang, R. & Paxson, V. (2003). A high-level programming environment for packet trace anonymization and transformation. In *Proceedings of the 2003 Conference on Applications, technologies, Architectures, and Protocols for Computer Communications*, pp. 339 – 351.

Parlamento Europeu, U. E. (1981). Convenção 108. Relativa ao tratamento de dados pessoais e à proteção da privacidade no setor das telecomunicações.

Parlamento Europeu, U. E. (1995). Directiva 95/46/ce. Relativa a proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados.

Parlamento Europeu, U. E. (1997). Directiva 97/66/ce. Relativa ao tratamento de dados pessoais e à proteção da privacidade no setor das telecomunicações.

Parlamento Europeu, U. E. (2001). Regulamento 45/ce. Relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais pelas instituições e pelos órgãos comunitários e à livre circulação desses dados.

Parlamento Europeu, U. E. (2002). Directiva 2002/58/ce. Relativa ao tratamento de dados pessoais e à proteção da privacidade no sector das comunicações eletrônicas (Diretiva relativa à privacidade e às comunicações eletrônicas).

Parlamento Europeu, U. E. (2006). Directiva 2006/24/ce. Relativa à conservação de dados gerados ou tratados no contexto da oferta de serviços de comunicações eletrônicas publicamente disponíveis ou de rede públicas de comunicações, e que altera a Directiva 2002/58/CE.

Peterson, L. L. & Davie, B. S. (2003). *Redes de Computadores uma Abordagem de Sistemas*. Editora Campus, tradução da 3a edição.

Pinheiro, P. P. (2008). *Direito Digital*. Saraiva, 2a edição.

pylibpcap (Acessado em 01/09/2009). Disponível em http://sourceforge.net/projects/pylibpcap/.

Rabinovich, M. & Spatscheck, O. (2002). *Web Caching and Replication*, chapter Basic Mechanisms for Request Distribution, pp. 231--246. Addison-Wesley.

Ramaswamy, R. & Wolf, T. (2007). High-speed prefix-preserving ip address anonymization for passive measurement systems. In *IEEE/ACM Transactions on Networking (TON)*.

Ribeiro, B.; Chen, W.; Miklau, G. & Towsley, D. (2008). Analyzing privacy in enterprise packet trace anonymization. In *Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS 08)*.

rubypcap (Acessado em 01/09/2009). http://www.goto.info.waseda.ac.jp/ fukusima/ruby/pcap-e.html.

Senado Federal, B. (2008). Projeto de lei no 494. Disciplina a forma, os prazos e os meios de preservação e transferência de dados informáticos mantidos por fornecedores de serviço a autoridades públicas, para fins de investigação de crimes praticados contra criança e adolescentes, e dá outras providências.

Senado Federal, B. (2009). Projeto substitutivo aos pls 76/2000, pls 137/2000 e plc 89/2003.

Shanmugasundaram, K. (2003). Fornet: A distributed forensic network. In *Proceedings of the Second International Workshop Mathematical Methods, Models and Architectures for Computer Networks Security*.

Silva, J. A. d. (1997). *Curso de Direito Constitucional Positivo*. Malheiros Editores.

Slagell, A.; Lakkaraju, K. & Luo, K. (2006). Flaim: A multi-level anonymization framework for computer and network logs. In *Proceedings of the 20th Large Installation System Administration Conference (LISA'06)*.

Slagell, A. & Yurcik, W. (2004). Sharing computer network logs for security and privacy: A motivation for new methodologies of anonymization. In *Proceedings of the Workshop on the Value of Security Through Collaboration (SECOVAL)*.

Spangler, R. (2003). Analysis of remote active operating system fingerprinting tools. Disponível em http://www.packetwatch.net/documents/papers/osdetection.pdf.

Steding-Jessen, K.; Vijaykumar, N. L. & Montes, A. (2008). Uso de *Honeypots* de baixa interatividade para o estudo do abuso de *Proxies* abertos para o envio de *Spam*. *INFOCOMP Journal of Computer Science*, xx(yy).

Tcpdump & libpcap (Acessado em 01/09/2009). Disponível em http://www.tcpdump.org.

Warren, S. D. & Brandeis, L. D. (1890). The right to privacy. harvard law review 4.

Wireshark (Acessado em 01/09/2009). Disponível em http://www.wireshark.org/.

Xu, J.; Fan, J.; Ammar, M. & Moon, S. B. (2001). On the design and performance of prefix-preserving ip traffic trace anonymization. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*.

Xu, J.; Fan, J.; Ammar, M. & Moon, S. B. (2002). Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. In *Proceedings of the 10th IEEE International Conference on In Network Protocols*.