UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Metodologia de Caracterização e Modelagem de Tráfego para Transmissão de Imagens Médicas

Robert Paulo Barbosa e Silva

SÃO CRISTÓVÃO/ SE
2015

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Robert Paulo Barbosa e Silva

Metodologia de Caracterização e Modelagem de Tráfego para Transmissão de Imagens Médicas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Edilayne Meneses Salgueiro

SÃO CRISTÓVÃO/ SE
2015

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

S586m

Silva, Robert Paulo Barbosa e.
Metodologia de caracterização e modelagem de tráfego para transmissão de imagens médicas / Robert Paulo Barbosa e Silva; orientador Edilayne Meneses Salgueiro. – São Cristóvão, 2015.
100 f.: il.

Dissertação (mestrado em Ciência da Computação)– Universidade Federal de Sergipe, 2015.

1. Modelagem. 2. Sistemas de transmissão de dados. 3. Radiografia médica. I. Salgueiro, Edilayne Meneses, orient. II. Título.

CDU 004.93

Robert Paulo Barbosa e Silva

Metodologia de Caracterização e Modelagem de Tráfego para Transmissão de Imagens Médicas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

Profa. Dra. Edilayne Meneses Salgueiro, Presidente
Universidade Federal de Sergipe (UFS)

Prof. Dr. Ricardo José Paiva de Britto Salgueiro, Membro
Universidade Federal de Sergipe (UFS)

Prof. Dr. José Augusto Suruagy Monteiro, Membro
Universidade Federal de Pernambuco (UFPE)

Metodologia de Caracterização e Modelagem de Tráfego para Transmissão de Imagens Médicas

Este exemplar corresponde à redação final da Dissertação de Mestrado, do mestrando **Robert Paulo Barbosa e Silva** para ser aprovada pela Banca examinadora.

São Cristóvão - SE, 28 de Maio de 2015

_____

Profa. Dra. Edilayne Meneses Salgueiro

Orientadora

_____

Prof. Dr. Ricardo José Paiva de Britto Salgueiro

Membro

_____

Prof. Dr. José Augusto Suruagy Monteiro

Membro

Dedicatória

Este trabalho é dedicado a todos que indiretamente ou diretamente contribuíram para a criação, orientação e inspiração na produção desta pesquisa. Dedico o meu mestrado a aqueles que ajudaram a tornar o meu sonho uma realidade, me proporcionando forças para que eu não desistisse de ir atrás do que eu buscava para minha vida. Muitos obstáculos foram impostos para mim durante esses últimos anos, mas graças a vocês eu não fraquejei. Dedico também este trabalho à minha maravilhosa esposa, Cynthia Yamara Bonfim Santos, que sempre me incentivou para a realização dos meus ideais, encorajando-me a enfrentar todos os momentos difíceis. Faço também uma dedicação especial a minha orientadora, Edilayne Salgueiro, por toda a ajuda na elaboração deste trabalho, ao professor Ricardo Salgueiro pelas dicas na elaboração da dissertação e aos meus pais pelo incentivo para sempre continuar estudando. E por fim, aos meus colegas de mestrado que sempre estiveram prontos para ajudar nas dificuldades.

Agradecimentos

Quero agradecer, em primeiro lugar, a Deus, pela força e coragem durante toda esta longa caminhada.

A todos os professores do curso, que foram tão importantes durante toda esta jornada. Em especial a professora Dra. Edilayne Meneses Salgueiro, por seus ensinamentos, paciência e confiança ao longo da orientação das minhas atividades neste mestrado. É um prazer tê-la como orientadora e na banca examinadora.

Ao Professor Dr. Ricardo José Paiva de Britto Salgueiro, com quem partilhei o que era o início daquilo que veio a ser esse trabalho e que forneceu grandes conselhos durante este estudo. Nossas conversas durante as aulas foram fundamentais. Desejei a sua participação na banca examinadora deste trabalho desde o princípio.

Ao professor Dr. José Augusto Suruagy Monteiro, por ter aceitado a participação nesta banca e por ter analisado de forma coerente este trabalho.

Aos meus pais, irmã, minha esposa e a toda minha família que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa da minha vida.

Aos amigos e colegas, pelo incentivo e pelo apoio constante.

Ao Curso de Mestrado da UFS, e às pessoas com quem convivi neste espaço ao longo desses anos. A experiência de estudo com vocês foram as melhores da minha formação.

A todos aqueles que de alguma forma estiveram e estão próximos de mim, fazendo esta vida valer cada vez mais a pena.

Por fim, agradeço a todos que produziram e produzem estudos científicos relevantes, permitindo assim o embasamento das atuais e futuras pesquisas.

Resumo

Caracterização e modelagem de tráfego em transmissões de imagens médicas é uma atividade importante para a gestão de redes corporativas. Com a popularização do uso de equipamentos de diagnóstico radiológico, uma grande quantidade de informação sobre o paciente passou a estar disponível. Essa nova forma digital de imagens de alta definição vem sendo integrada às redes de computadores ao longo dos anos, com uso do protocolo DICOM. Recentemente, uma grande quantidade de equipamentos médicos radiológicos passou a transmitir imagens médicas em redes de computadores, gerando assim, novos desafios para operações de monitoramento e gerenciamento das redes. Este trabalho apresenta uma metodologia para a modelagem de tráfego DICOM. Foram realizadas medições de tráfego na rede e coletas diretas nos equipamentos radiológicos de um hospital de pequeno porte para análise. As medições sobre o tráfego da rede foram efetuadas para caracterizar o comportamento deste tráfego, identificando assim, a sua forma e composição na rede. As coletas nos equipamentos radiológicos foram executadas ao longo de um ano para modelagem da fonte de tráfego. A modelagem da fonte de tráfego foi realizada com uso de técnicas estatísticas de ajuste de curvas, para modelar a distribuição de tamanho dos arquivos de imagens. Análises de testes de aderência apontaram a distribuição de Dagum como a de melhor aproximação nestes resultados. Deste modo, o modelo de fonte de tráfego sugerido por esse trabalho pode ser utilizado em experimentos de simulação e em projetos de expansão da rede.

Abstract

Characterization and traffic modeling in medical image transmission is an important activity for corporate network management. With popularization of medical imaging a lot of information about the patients became available. This new form of digital high-definition images has been integrated into computer networks over the years, using the DICOM protocol. Recently, a large amount radiological equipament has been used for medical image generation transmission, thus causing new challenges for monitoring and management of network operations. This paper presents a methodology for modeling DICOM traffic. Traffic measurements were carried out on the network and direct collections in radiological equipment of a small hospital for analysis. Measurements on the network traffic were performed to characterize the behavior of the traffic, thereby identifying the model and composition of the network. The traffic source capture were executed over a year to traffic from real medical diagnosis. The modeling of the traffic source was performed using curve fitting statistical techniques, to model the distribution of image file sizes. Compliance tests analysis showed the Dagum distribution as the best approach in these results. Thus, the traffic source model suggested by this work can be used in simulation experiments and network expansion projects.

Lista de Figuras

2.1 Infraestrutura Básica PACS 9
2.2 PACS Centralizado 10
2.3 PACS Descentralizado 11
2.4 Arquitetura DICOM 12
2.5 Envio de imagem de um TC para uma estação de trabalho com DICOM . . 14

3.1 Metodologia de Classificação de Pacotes 21
3.2 Função de Densidade de Probabilidade (pdf) - Lognormal 26
3.3 Função de Densidade de Probabilidade (pdf) - Exponencial 26
3.4 Função de Densidade de Probabilidade (pdf) - Weibull 27
3.5 Função de Densidade de Probabilidade (pdf) - Gamma 27
3.6 Função de Densidade de Probabilidade (pdf) - Pareto 28
3.7 Função de Densidade de Probabilidade (pdf) - Dagum 29

4.1 Metodologia de Caracterização e Modelagem de Tráfego 42

5.1 Tamanho médio dos pacotes comparando 4 amostras DICOM com outros protocolos 49
5.2 Vazão do tráfego da rede comparando 4 amostras DICOM com outros protocolos 50
5.3 Ambiente de Rede do Estudo 52
5.4 Histograma de distribuição da Quantidade de Pacotes por Tamanho 71
5.5 Histograma de distribuição da Quantidade de Pacotes DICOM por Tamanho 72
5.6 Comparação entre Distribuição Real e Probabilísticas após Ajuste dos Parâmetros 73

5.7 CDF de Comparação entre Distribuição Real e Probabilística 73
5.8 pdf de Comparação entre Distribuição Real e Probabilística 74
5.9 P-P Plot de Comparação entre Distribuição Real e Probabilística 74
5.10 Simulação do Distribuição Dagum com Transmissão de até 10 Modalidade 75

Lista de Tabelas

5.1 Identificação de Tráfego por Fluxos Preliminar 51
5.2 Coletas de Tráfego Realizadas para Classificação 54
5.3 Data e Tamanho dos Exames X Dias das Coletas 55
5.4 Estatística Geral das Coletas 55
5.5 Estatística DICOM das Coletas 56
5.6 Pacotes por Protocolo 58
5.7 Bytes por Protocolo 59
5.8 Quantidade de Fluxo Geral por Tempo 60
5.9 Quantidade de Fluxo DICOM por Tempo 61
5.10 Identificação de Tráfego Geral (Fluxos Elefantes) 61
5.11 Identificação de Tráfego DICOM (Fluxos Elefantes) 62
5.12 Características da fonte de dados Ultrassom (US) por período 63
5.13 Parâmetros Utilizados nas Distribuições Probabilísticas 64
5.14 Qualidade do Ajuste entre as Distribuições Real e Probabilística 65
5.15 Parâmetros da Simulação 67
5.16 Vazão média na simulação com até 10 modalidades 68
5.17 Pacotes Perdidos e Atraso na simulação com até 10 modalidades 69

A.1 Distribuição Total da Quantidade de Pacotes por Tamanho 86
A.2 Distribuição DICOM da Quantidade de Pacotes por Tamanho 87

Lista de Siglas

ADSL	*Asymmetric Digital Subscriber Line*	15
DICOM	*Digital Imaging and Communications in Medicine*	1
Diffserv	*Differentiated Services*	33
FTP	*File Transfer Protocol*	15
HD	*Hard Disk*	9
HTTP	*Hypertext Transfer Protocol*	35
IP	*Internet Protocol*	13
ISP	*Internet Service Provider*	36
MSD	*Multi-Series DICOM*	13
NEMA	*American National Association of Electric Machines*	8
OSI	*Open Systems Interconnection*	12
PACS	*Picture Archiving and Communication System*	1
QoS	*Quality of Service*	3
RAID	*Reduntant Array of Inexpensive Disks*	10
RIS	*Radiology Information System*	15
RSNA	*Radiology Society of North America*	8
RTFM	*Realtime Traffic Flows Measurement*	23
S-MIME	*Secure Multipurpose Mail Extension*	15
SFD	*Single-Frame DICOM*	13
SMTP	*Simple Mail Transfer Protocol*	35
SSH	*Secure Shell*	38
TCP	*Transmission Control Protocol*	13
TE	*Traffic Engineering*	21

TLS	*Transport Layer Security*	15
UDP	*User Datagram Protocol*	39
VOIP	*Voice Over Internet Protocol*	38
VPN	*Virtual Private Network*	15
WIMAX	*Worldwide Interoperability for Microwave Access*	38
WRR	*Weighted Round Robin*	39

Sumário

Lista de Siglas **viii**

1 Introdução **1**

1.1 Problemática e Hipótese 3

1.2 Objetivos da Dissertação 4

1.3 Justificativa 5

1.4 Organização da Dissertação 6

2 Tecnologias de Imagens Médicas **7**

2.1 Infraestrutura para Tecnologias de Imagens Médicas 7

2.2 Sistema de Armazenamento 8

2.3 Transmissão de Imagens Médicas 10

2.4 Medições em Ambiente com PACS e DICOM 14

3 Caracterização e Modelagem de Tráfego **17**

3.1 Caracterização de Tráfego 17

3.1.1 Monitoramento 18

3.1.2 Classificação 20

3.1.3 Identificação de Tráfego 21

3.1.4 Metodologias de caracterização 22

3.2 Modelagem de Tráfego 23

3.3 Trabalhos Relacionados 30

3.3.1 Caracterização 30

3.3.2 Métodos de Caracterização 35

3.3.3 Modelagem de Tráfego 37

4 Metodologia Proposta **40**

4.1 Política de Transferências DICOM 41

4.2 Caracterização de Tráfego 43

4.2.1 Forma de Coleta de Dados 44

4.2.2 Filtragem e Classificação 44

4.3 Modelagem da Fonte de Dados 44

4.4 Planejamento da Avaliação de desempenho 46

5 Caracterização e Modelagem de Tráfego DICOM **48**

5.1 Caracterização Realizada 48

5.1.1 Estudos Preliminares 48

5.1.2 Estudo de Caso 50

5.1.3 Coletas de tráfego 53

5.1.4 Classificação dos Dados 56

5.1.5 Identificação do Tráfego 58

5.2 Modelagem de Tráfego DICOM 60

5.2.1 Coleta de Dados da Modalidade 60

5.2.2 Caracterização da Fonte 62

5.2.3 Modelagem da Distribuição 63

5.3 Avaliação de Desempenho 65

6 Conclusão **76**

6.1 Contribuições 77

6.2 Trabalhos Futuros 77

Referências **78**

A Resultados Complementares **86**

Capítulo 1

Introdução

Com o surgimento das tecnologias de diagnóstico por imagem e a popularização dos equipamentos de diagnóstico, uma grande quantidade de informações sobre o paciente passou a estar disponível. Essas informações eram descartadas ou guardadas em arquivos físicos, que em ambos os casos, não facilitavam o trabalho dos profissionais da área da saúde para acesso ao histórico do paciente no momento da consulta, algo importante na busca do melhor diagnóstico ou tratamento. Além disso, criava-se um transtorno enorme para as instituições de saúde, visto que uma grande área de armazenamento destes documentos se fazia necessária, gerando altos custos.

Com o advento das novas tecnologias, sobrevieram formas digitais de armazenamento e transmissão das imagens médicas e das informações dos pacientes. Essas novas formas de guarda e distribuição possibilitaram a redução do custo de conservação das imagens e uma maior facilidade de acesso a essas informações. Por outro lado, também surgiram novos desafios para a área de tecnologia da informação como a manutenção destes dados armazenados, o processamento destes dados em formato digital e o gerenciamento das transmissões dessas informações.

Diante destes desafios, várias formas de armazenamento, gerenciamento e transmissão de imagens médicas surgiram com o intuito de se tornarem um padrão de mercado, mas os modelos que se destacaram como padrões foram o **PACS** (*Picture Archiving and Communication System*) (MARQUES; SALOMÃO, 2009), que é muito utilizado em instituições de saúde para armazenamento, gerenciamento e recuperação das imagens em formato digital, e o protocolo **DICOM** (*Digital Imaging and Communications in Medicine*) (BIDGOOD, 1997),

adotado pelos grandes fabricantes de equipamentos médicos para transmissões através das redes de computadores, possibilitando assim, uma padronização nas consultas, recuperações e transmissões dos dados. Esses padrões dissiparam problemas anteriores como a falta de paradigmas de arquivamento, dificuldades de indexação das imagens, problemas de comunicação entre equipamentos de fabricantes diferentes e dificuldades de acesso ao histórico das imagens médicas em equipamento de fabricantes diferentes ou em modelos de equipamento de diagnóstico com finalidades distintas. Ademais, a padronização da forma de armazenamento e da comunicação no mercado de saúde possibilitou a transmissão destes dados utilizando as redes de dados, surgindo assim, mais um serviço convergente nas redes. Entretanto, o impacto destes novos dados na rede, o seu comportamento nas transmissões e em relação aos demais tráfegos são alguns dos temas relevantes a serem estudados atualmente. O desconhecimento do comportamento do tráfego de imagens médicas em relação aos demais tráfegos e o impacto do volume destes novos dados, podem acarretar em problemas futuros como, por exemplo, redução na taxa de vazão, perda de dados e atrasos, para alguns ambientes de rede podendo inclusive até inviabilizar o funcionamento de alguns serviços.

Estudos realizados nos Estados Unidos indicavam que só em 2014 seriam executados cerca de um bilhão de exames por imagens, gerando aproximadamente 100 petabytes de dados para armazenamento e transmissão (ROSTROM; TENG, 2011). Para o período de 2014 a 2015, esse volume tende a ser ainda maior com valores bem superiores a 100 petabytes (GULATI, 2015). Esse volume de dados abriu espaço para novas pesquisas na área de engenharia de tráfego de rede como, por exemplo, o monitoramento, projeto e implementação do ciclo de operação das redes utilizando informações de caracterização de tráfego e a modelagem do tráfego para, através de simulação, realizar mudanças ou ampliações conscientes nas redes (LEE; LEVANTI; KIM, 2014).

O conhecimento do comportamento do tráfego das imagens médicas passam inicialmente pela caracterização destes novos fluxos de dados na rede. É importante o conhecimento do comportamento deste tráfego para que não haja uma degradação dos serviços já existentes. Além disso, este tipo de tráfego vem se expandindo para outros tipos de ambientes como rede sem fio ou computação nas nuvens, ampliando a importância do conhecimento deste novos fluxos de dados. A compreensão do comportamento deste tráfego nas redes locais poderá servir de comparação ou de parâmetros para futuras caracterizações de outras áreas

da saúde e da indústria.

No estudo (WAMSER, 2011) sobre caracterização do acesso à internet em redes sem fio residenciais, inesperadamente constatou-se 73 fluxos de pacotes de dados com volume correspondente a 1,25 TB referindo-se a tráfegos médicos. Esse grande volume de dados em uma quantidade pequena de fluxos de pacotes expõe a importância que estes dados podem ter em uma rede. Logo, o conhecimento do comportamento dos mesmos pode ser de suma importância.

Segundo Martins (2008), a caracterização de tráfego em si pode contribuir de forma significativa na melhoria das tarefas de gestão de rede, alocação de recursos, planejamento, design, controle de **QoS** (*Quality of Service*), segurança e detecção de intrusões.

Além da caracterização, a modelagem de tráfego também pode contribuir no processo de entendimento do funcionamento deste tipo de projeto. A modelagem procura replicar o comportamento do tráfego, permitindo o uso desta réplica em simulações de desempenho, na busca de melhoria da performance do mesmo, através da repetição dos modelos propostos em outras escalas, dimensões e parâmetros (THYAGO ANTONELLO; CUNHA et al., 2008). Este tipo de pesquisa pode possibilitar a análise de modificações ou expansões futuras das novas infraestruturas propostas, antes que as implementações sejam realizadas.

O esclarecimento das características dos fluxos de dados de imagens médicas, em conjunto com outros fluxos existentes na rede, podem aclarar a viabilidade deste tipo de dados sobre as redes atuais. A modelagem da fonte de dados de imagens médicas, pode possibilitar a simulação do comportamento das redes para o gerenciamento e expansão de infraestrutura que a contemplem.

1.1 Problemática e Hipótese

Problema

Progressivamente, novos equipamentos da área de saúde vêm sendo integradas às redes de computadores. O conhecimento sobre as características dos dados gerados por estes equipamentos nas redes de computadores ainda são muito escassos, existindo com isso problemas ainda não estudados a fundo sobre o seu comportamento, possibilitando assim, decisões equivocadas na gestão das redes.

O uso elevado de carga de dados de imagens médicas nas redes de computadores ampliou ainda mais os desafios de gestão e manutenção da qualidade dos serviços de rede nas corporações. Como pode ser visto em (WAMSER, 2011), o uso de dados médicos em redes de computadores já é uma realidade constante. O alto volume dos dados gerados por esses fluxos de dados podem se tornar um grande problema para as redes.

As transmissões das imagens médicas em conjunto com outros tráfegos de rede como, por exemplo, voz sobre IP, tráfego HTTP, imagens de câmeras de segurança e dados de sistemas hospitalares podem modificar negativamente o comportamento dos serviços já existentes devido ao grande volume de dados, como o encontrado em (GULATI, 2015). O acréscimo de novos dados podem, em alguns casos, inviabilizar o funcionamento de alguns serviços. A falta de entendimento do comportamento destes dados, dependendo do ambiente utilizado, pode acarretar em queda consideráveis da qualidade dos serviços da rede.

A implantação de novos serviços integrados à rede sem a devida análise do seu comportamento pode acarretar em inserção de novos problemas na rede que antes não existiam como, por exemplo, mal dimensionamento da infraestrutura de suporte ao novo serviço ou configuração equivocada da largura de banda mínima para um determinado serviço, além de criar dificuldades na resolução deles devido ao desconhecimento do seu comportamento. Por isso, se faz necessário o uso de alguma forma de análise da repercussão deste novos serviços no ambiente. A descoberta das pertubações inerentes aos novos serviços podem evitar o surgimento de problemas inesperados na inserção dos mesmos.

Hipótese

O conhecimento das características do tráfego, aliado à sua modelagem, podem possibilitar uma melhor compreensão do comportamento dos dados de imagens médicas, permitindo assim, a inserção de novas diretrizes de gestão na rede para melhoria do desempenho.

1.2 Objetivos da Dissertação

O ponto central deste trabalho é a caracterização e modelagem do tráfego de imagens médicas em redes de computadores hospitalares. A finalidade da caracterização é identificar o comportamento do tráfego DICOM. O modelo de tráfego resultante deste trabalho pode ser usado como entrada em experimentos de avaliação de desempenho ou planejamento de

capacidades como a ampliação da infraestrutura da rede voltada para a saúde.

Objetivos Específicos

Para alcançar os objetivos gerais alguns objetivos específicos devem ser atingidos:

- Definição de metodologia para caracterização e modelagem de tráfego DICOM;
- Caracterização do tráfego DICOM em redes de computadores;
- Definição de um modelo de distribuição probabilística da fonte de dados de imagens médicas.

1.3 Justificativa

As tecnologias na área de saúde estão fazendo uma revolução no tratamento e diagnóstico dos pacientes nas instituições de saúde. As imagens médicas fazem parte desta revolução, sendo a recuperação e distribuição destas imagens de suma importância para a obtenção de um rápido diagnóstico ou para uma melhor consulta do histórico do paciente, tornando assim estas imagens um item importante para um bom tratamento dos enfermos. Diante deste cenário, a distribuição destas imagens através das redes de computadores se tornou importantíssima, sendo ao mesmo tempo, em alguns casos, um possível problema para a administração das redes, visto que outros serviços estão em funcionamento nela e esses novos serviços podem provocar possíveis problemas na qualidade dos serviços antigos. Devido a essa dualidade de situações, se faz necessário o aumento das pesquisas sobre os impactos destes tráfegos nas redes de computadores, procurando assim manter ou ampliar a possibilidade de manutenção da qualidade dos serviços.

Além da manutenção da qualidade dos serviços já existentes na rede, o aumento do número de exames realizados, a aquisição de novos equipamentos ou modalidades de exames e a ampliação do compartilhamentos destes dados entre várias unidades de saúde, através das redes, são fatores que ensejam a realização de monitoramento de tráfego.

Este trabalho pretende, além dos pontos citados anteriormente, despertar o interesse de profissionais e estudantes da área de tecnologia no que diz respeito à pesquisa contínua sobre as características de tráfegos e a sua modelagem. Além disso, procura-se explanar a real influência que o tráfego de imagens médicas proporciona nas redes atuais, auxiliando os

administradores de rede nas suas decisões e revelando mais informações sobre este tipo de tráfego. Buscar-se-á saber até que ponto este tráfego compõe a rede e influencia no seu funcionamento. Ao mesmo tempo, esta pesquisa busca construir um material escrito que auxilie outros pesquisadores nas investigações de outros tráfegos de redes.

1.4 Organização da Dissertação

O restante do texto está organizado nos seguintes capítulos:

- O Capítulo 2 apresenta o padrão de armazenamento de imagens médicas e o protocolo de tráfego de imagens médicas com seus componentes;
- O Capítulo 3 apresenta técnicas usadas para caracterização de tráfego e modelagem de dados;
- O Capítulo 4 apresenta a metodologia proposta por essa pesquisa;
- O Capítulo 5 apresenta os resultados da caracterização e a análise dos dados com a modelo de fonte de dados de uma modalidade;
- E finalmente, no Capítulo 6 são discutidas as conclusões encontradas e trabalhos futuros.

Capítulo 2

Tecnologias de Imagens Médicas

Neste capítulo são abordados os conceitos teóricos sobre as tecnologias de imagens médicas, sendo elas, respectivamente, a infraestrutura necessária para uso da tecnologia de imagens médicas integrada às redes de computadores (Seção 2.1), entendimento do conceito de PACS (Seção 2.2) e o protocolo de comunicação para transferência de imagens médicas DICOM (Seção 2.3).

2.1 Infraestrutura para Tecnologias de Imagens Médicas

Para a implantação de uma infraestrutura de gerenciamento de imagens médicas, são necessários os seguintes recursos: equipamentos de diagnóstico por imagem conhecidos como modalidades radiológicas, um ambiente de armazenamento para as imagens e as informações dos pacientes, e um protocolo de comunicação para transmissão das imagens médicas nas redes de computadores.

Como exemplo de modalidade radiológica, podemos citar a ultrassonografia (US), Raio X Digital (X-Ray), Tomografia computadorizada (TC), Ressonância Magnética (RM) entre outros. Modalidade é o jargão técnico utilizado para denominar equipamentos individuais além dos tipos de sistema de exames existentes no ambiente de radiologia.

No quesito armazenamento, o sistema comumente utilizado é o PACS que utiliza mecanismos eficientes para guarda e recuperação das imagens nos *hardwares* de arquivamento.

No caso do protocolo de comunicação, o DICOM é o mais utilizado. O papel deste protocolo no ambiente de gestão de imagens médicas é realizar a transmissão dos dados entre

o ambiente de armazenamento e as modalidades radiológicas, movimentando, buscando, recuperando ou entregando as imagens obtidas nas modalidades para guarda no ambiente de armazenamento ou realizando essas mesmas funções entre o ambiente de armazenamento e as estações de trabalho.

2.2 Sistema de Armazenamento

O armazenamento de imagens e informações clínicas surgiu por volta de 1980, após a introdução do uso de dados no formato digital. Naquela época, um novo formato de armazenamento e recuperação de imagens médicas nos ambientes hospitalares se fazia necessária. O conceito de PACS surgiu desta necessidade. PACS é um sistema de arquivamento e comunicação de imagens e dados digitais dos pacientes, voltado para o diagnóstico por imagem, permitindo assim um ponto de acesso às imagens médicas em formato digital em qualquer setor do hospital (SIBARANI, 2012). PACS foi criado por um consórcio integrado pela **NEMA** (*American National Association of Electric Machines*) (ASSOCIATION, 2015), **RSNA** (*Radiology Society of North America*) (RSNA, 2015) e um conjunto de empresas e universidades dos Estados Unidos da América.

Os requisitos de PACS, de acordo com a (ASSOCIATION, 2015) são a oferta de imagens em estações remotas, armazenamento de dados para recuperação, em curto ou longo prazo, comunicação em rede e sistema de interfaceamento e conexão com as modalidades radiológicas. A Figura 2.1 demonstra um exemplo deste ambiente, onde imagens das modalidades são transmitidas para o PACS e posteriormente consultadas pelos clientes.

O PACS em conjunto com outros sistemas de informações radiológicas tem o objetivo de eliminar o filme, completamente ou em grande parte das instituições de saúde, substituindo-o por sistemas eletrônicos. O servidor PACS é um dos componentes fundamentais da arquitetura de radiologia. Ele pode ser dividido em dois componentes: o controlador e o armazenador.

O primeiro componente é o controlador PACS, que organiza os dados e a comunicação no PACS, utilizando hardware e software. A forma de comunicação pode ser centralizada ou descentralizada. A comunicação centralizada ou sob demanda, Figura 2.2, trabalha com o envio direto das imagens médicas para o servidor PACS e dele para as estações de trabalho.

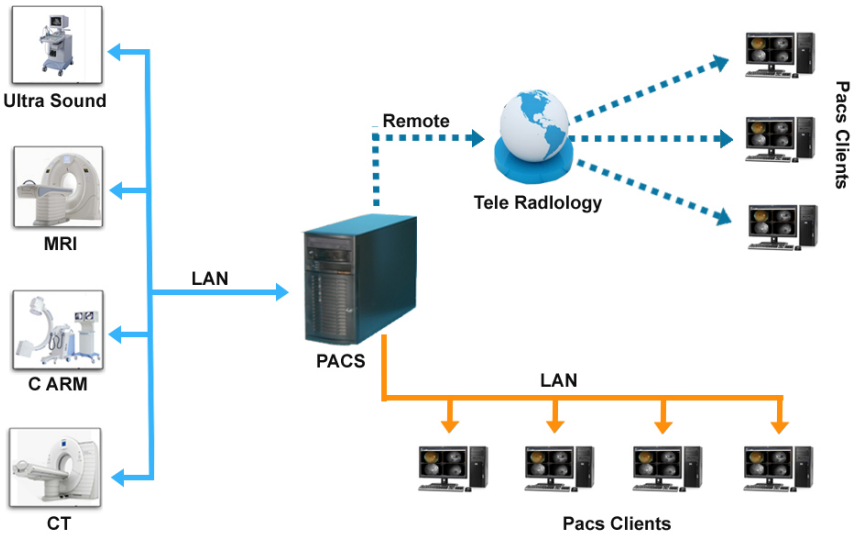

Figura 2.1: Infraestrutura Básica PACS

Já a descentralizada ou roteada, Figura 2.3, trabalha de forma inversa, enviando as imagens diretamente para a estação de trabalho e depois da estação de trabalho para o servidor PACS ou outras estações de trabalho.

As duas arquiteturas têm vantagens e desvantagens, sendo que na arquitetura centralizada as principais vantagens são a organização e o melhor gerenciamento dos dados. Já as principais desvantagens desta arquitetura são a sua maior dependência da rede e do servidor centralizador.

Em contrapartida, na arquitetura descentralizada ocorre justamente o inverso, não há um sistema centralizado de armazenamento. Essa estrutura dificulta a recuperação das imagens em outra ocasiões em que elas sejam necessárias, visto que as imagens estão distribuídas em vários computadores, causando assim uma desorganização e uma não padronização da forma de armazenamento. Entretanto, o acesso às informações ainda estarão disponíveis mesmo nos casos em que a rede ou o servidor de armazenamento não esteja presente (MARQUES; SALOMÃO, 2009).

O segundo componente, o armazenador, é responsável por guardar de forma segura e íntegra os dados das imagens recebidas (MARQUES; SALOMÃO, 2009). É possível utilizar várias técnicas de armazenamento e diversos tipos de hardware. O dispositivo mais utilizado para armazenamento é o **HD** (*Hard Disk*) e a principal técnica de armazenamento usada é

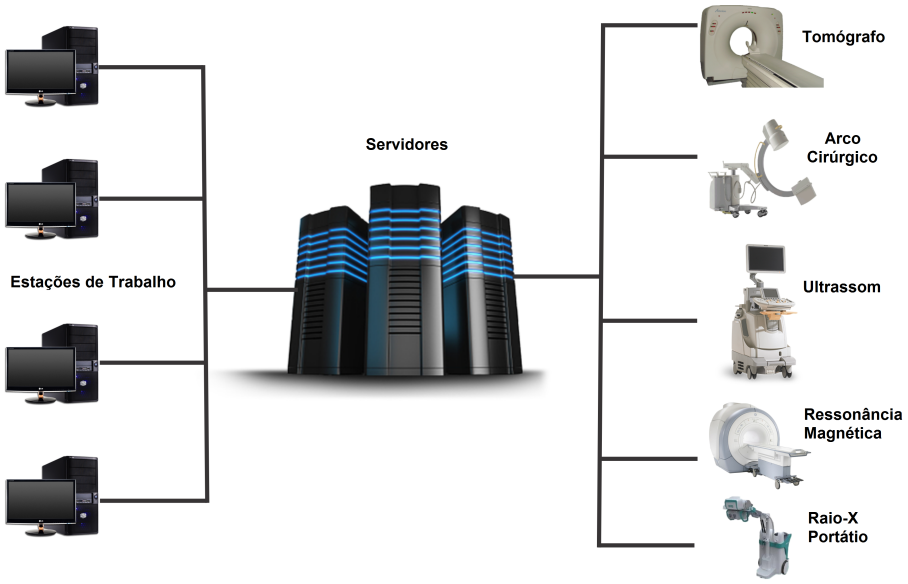

Figura 2.2: PACS Centralizado

o **RAID** (*Reduntant Array of Inexpensive Disks*). Estes dispositivos precisam prover alta disponibilidade e resiliência. Disponibilidade pode ser expressa pelo tempo que este dispositivo deve permanecer sem paradas; quanto mais alto o tempo, melhor. Resiliência é "a capacidade de retomada de uma tarefa após uma parada imprevista" (MARQUES; SALOMÃO, 2009).

Existem vários sistemas PACS disponíveis no mercado, havendo inclusive versões *open source* e softwares comerciais. Como exemplo de versão *open source* temos o DCM4CHEE que está disponível na internet para uso e é utilizado no ambiente de análise deste trabalho. Informações mais detalhas do funcionamento do PACS podem ser encontradas também em (KIM, 2015) e (FLOYD, 2015).

2.3 Transmissão de Imagens Médicas

Com o surgimento de várias modalidades médicas na década de 70, o ACR (American College of Radiology) (ACR, 2015) e o NEMA perceberam a necessidade de criação de um padrão para a transferência de imagens e informações entre aparelhos de diagnóstico por imagem fabricados por empresas diferentes. Para que os aparelhos seguissem uma padronização, o (ACR) e a (NEMA) formaram no ano de 1983 uma comissão mista, para criar um

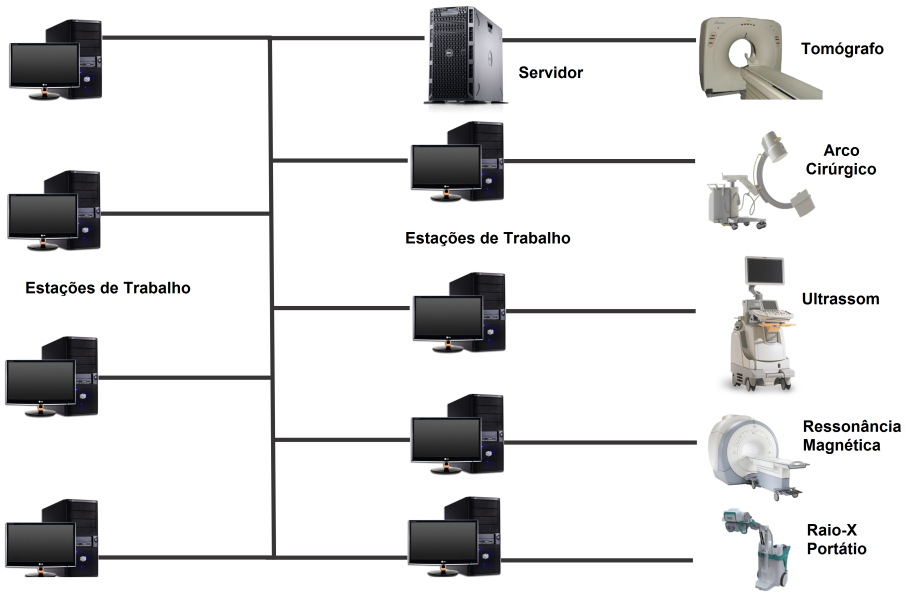

Figura 2.3: PACS Descentralizado

protocolo em que os aparelhos de diversos fabricantes pudessem se comunicar e trocar informações referentes aos exames médicos. Com os estudos dessa comissão, a ACR-NEMA publicou em 1985, a primeira versão do padrão DICOM 1.0.

No ano de 1988 foi publicada a versão 2.0 do padrão DICOM. Essa versão incluía toda versão 1.0 com suas revisões publicadas, introduzia um novo material que prestava apoio de comando para dispositivos com telas e colocava um regime novo de hierarquia para identificar uma imagem, adicionando novos elementos de dados para uma maior especificação no descritivo das imagens.

Atualmente, o protocolo é desenvolvido com colaboração de várias organizações. A versão atual encontrada é a 3.0 que foi publicada no ano de 1993, e vem se atualizando e definindo novas classes de serviços em todos esses anos com publicação de novos documentos, sempre respeitando as instruções da comissão criada pela ACR-NEMA (ASSOCIATION, 2015).

A conectividade deste padrão tem um custo benefício muito bom e a sua extrema flexibilidade possibilita a utilização de recursos tecnológicos já existentes. Estas características o levaram a ser adotado por diversas fabricantes em várias modalidades e o seu uso tem um alcance mundial.

O mesmo foi projetado para acompanhar a evolução das tecnologias de radiologia. Os

equipamentos e sistemas que utilizam esse padrão precisam sempre monitorar as mudanças que ocorrem para manterem a evolução da comunicação entre estes sistemas.

O DICOM é estruturado da seguinte forma:

- Classes de objetos que buscam, através de atributos, a padronização de formatos de dados de diagnóstico médico inerente ao mundo real. O uso das classes identifica os objetos encontrados em sistemas de diagnóstico por imagens e os empacotam para transmissão e armazenamento. Esses objetos no jargão da comunidade de radiologia é conhecido como "Estudo DICOM".

- Serviços DICOM utilizados para comunicação de objetos de informação dentro das modalidades e para execução de serviços para os objetos. Os serviços principais podem ser do tipo armazenamento C-STORY, busca de informações C-FIND e recuperação C-MOVE.

- Comunicação DICOM que utiliza padrões já existentes de comunicação, baseado no modelo **OSI** (*Open Systems Interconnection*) para transmissão de imagens médicas.

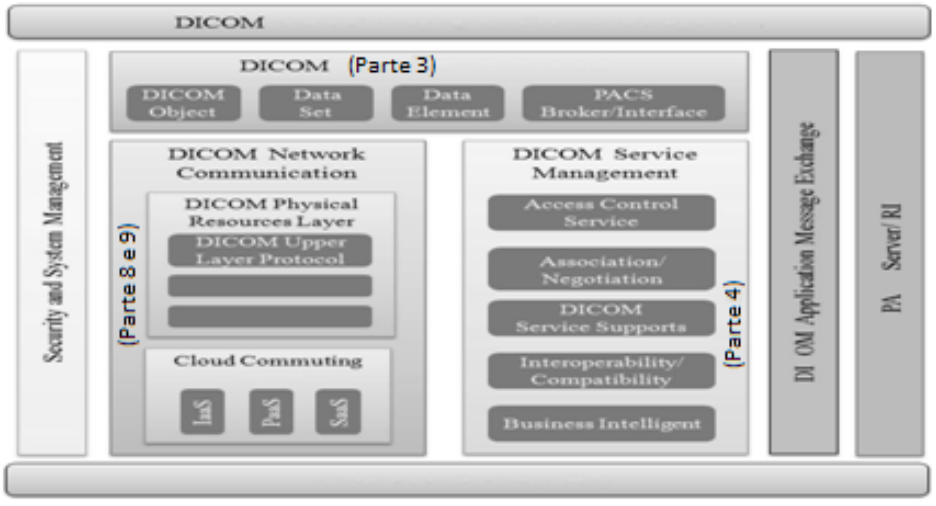

Figura 2.4: Arquitetura DICOM

Fonte: Patel, 2012

O DICOM é, atualmente, um padrão muito utilizado pela indústria de equipamentos médicos no mundo. Este padrão atualmente está dividido em 20 partes (ASSOCIATION, 2011). A documentação foi dividida em partes para facilitar as atualizações e flexibilizar a expansão

ao longo do tempo. A Figura 2.4 contém a arquitetura DICOM com algumas de suas partes (PATEL, 2012). Para um melhor entendimento, serão explanadas as partes mais importantes para este trabalho. Informações inerentes às demais partes podem ser encontradas em (ASSOCIATION, 2011).

A parte três, definição dos objetos de informação, apresentada na Figura 2.4 na parte superior central, procura produzir objetos da vida real, através das imagens, informações dos pacientes e informações correlacionadas, como, por exemplo, nível de radiação e relatórios médicos (ASSOCIATION, 2011). Estes objetos são encapsulados em arquivos com a extensão ".dcm". As formas de transmissão destes arquivos podem ser **SFD** (*Single-Frame DICOM*) ou **MSD** (*Multi-Series DICOM*) (ISMAIL; NING; PHILBIN, 2013). O SFD, mais tradicional, transmite os objetos de forma síncrona aguardando a confirmação de recebimento pelo receptor para envio do próximo objeto. O tempo de aguardo da confirmação resulta em tempo de transmissão mais lento para esse formato de dados. Já o MSD, envia vários objetos em uma única transmissão. Esse modo de transmissão assíncrona necessita de apenas uma confirmação para vários objetos, entretanto o tamanho dos objetos trafegados são bem maiores.

A parte quatro, como descrito por Sibarani (2012) especifica classes de serviço, associando um ou mais objetos de informações com um ou mais comandos a serem efetuados nestes objetos. Essas classes de serviços realizam o armazenamento, movimentação, recuperação e outros comandos para os objetos DICOM. Na Figura 2.4, ela é representada pelo retângulo nomeado de *"DICOM Service Management"*.

As partes oito e nove, suporte para troca de mensagens via rede, disponibilizam os elementos necessários para interface entre o protocolo DICOM, na camada de aplicação, e o protocolo **TCP** (*Transmission Control Protocol*) e **IP** (*Internet Protocol*), nas outras camadas. Esta parte da documentação descreve a forma de troca de informações entre as modalidades e as máquinas na rede através da pilha de protocolos TCP/IP e DICOM. A identificação desta parte na Figura 2.4 é "Comunicação de Rede DICOM".

A comunicação do protocolo DICOM acontece conforme a Figura 2.5. A modalidade em uso, neste caso um tomógrafo (TC), codifica todas as imagens em um objeto DICOM conforme item (a). Em seguida, a modalidade invoca uma série de serviços para mover o objeto até a camada física do modelo TCP/IP, item (b). O próximo passo é a estação de trabalho que utiliza uma série de serviços para receber o objeto através da camada física e

depois movê-lo para camadas de maior nível (c). E por último, em (d), a estação de trabalho decodifica o objeto DICOM.

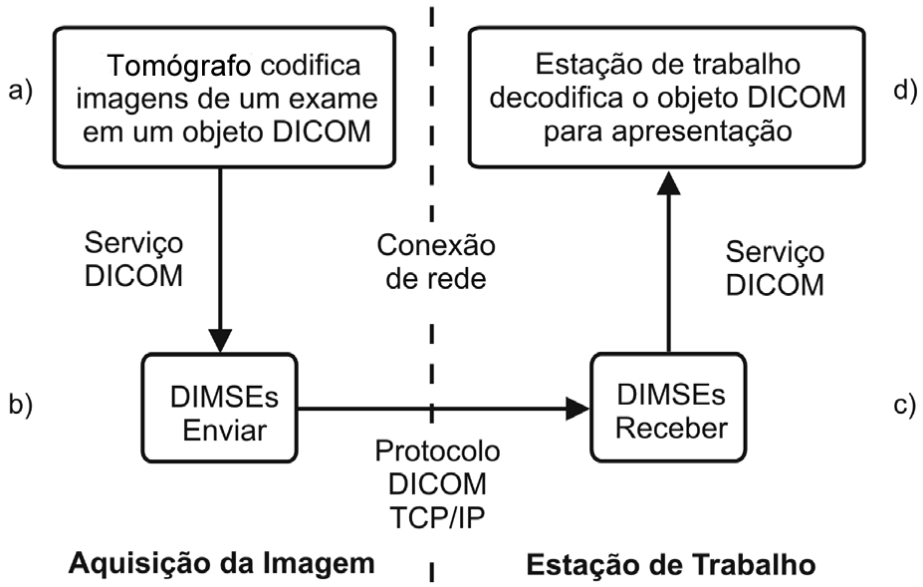

Figura 2.5: Envio de imagem de um TC para uma estação de trabalho com DICOM

Fonte: Marques e Salomão, 2009

2.4 Medições em Ambiente com PACS e DICOM

Nesta seção são apresentados trabalhos que proporcionaram informações para o entendimento do processos de medições dentro do contexto de imagens médicas. Ferramentas, metodologias de medição e técnicas de caracterização são algumas das informações obtidas aqui.

Em (SYSTEMS, 2008), foi realizado um estudo de medição de tráfego DICOM em rede WAN (Wide Area Network) em três tamanhos diferentes de hospitais com o uso de equipamentos Cisco. Os equipamentos testados tinham o objetivo de otimizar o desempenho das redes TCP em um ambiente de WAN. Os resultados obtidos expõem as características destes tipos de dados em diferentes tamanhos de hospitais e conclui que houve uma significativa redução de tráfego na rede WAN e uma melhora na produtividade do uso das modalidades na borda da rede com o uso dos equipamentos da Cisco no ambiente. Os valores apresentados

no trabalho informam que houve uma melhoria entre 33% e 36% no nível de desempenho da rede com o uso do equipamento Cisco sem a ativação de controle de cache. Já no teste com o equipamento e o uso de cache, o nível de eficiência e rendimento variaram entre 95% e 100% de melhoria.

O estudo de Perez e outros (2010) apresenta um experimento com o uso dos protocolos de segurança **TLS** (*Transport Layer Security*), **VPN** (*Virtual Private Network*) e **S-MIME** (*Secure Multipurpose Mail Extension*) em conjunto com as transmissões do protocolo DICOM. O artigo expõe a necessidade de segurança nas transmissões de dados médicos devido ao sigilo dos dados dos pacientes. Além disso, um software nomeado de DICUS foi criado para inserção de segurança por criptografia TLS. Nos ensaios foram enviados grupos de imagens médicas nos períodos da manhã, tarde e noite utilizando três formas de comunicação: DICOM Padrão, Security DICOM e Security **FTP** (*File Transfer Protocol*). Nos experimentos na Intranet, com velocidade de conexão de 100 Mbps, foram obtidos resultados que sugerem que as transmissões ficaram 65% mais lentas nas comunicações com os protocolos de segurança. Nos testes entre dois hospitais utilizando um link de acesso à Internet do tipo **ADSL** (*Asymmetric Digital Subscriber Line*), com velocidade de conexão de 1 Mbps em um dos lados e 300 kbps no outro, os resultados apontaram um aumento na lentidão da ordem de 5,86% na comunicação.

O artigo de Hasan (2012) inicia explicando os vários tipos de aplicações para a área de saúde e os vários protocolos existentes para comunicação. Os autores propõem um arcabouço chamado IHON que tem como objetivo facilitar a comunicação ou as transações entre as várias entidades como equipamentos, aplicativos, modalidades e etc, em rede hospitalar. O arcabouço trabalha inspecionando os pacotes que trafegam na rede, identificando com base em políticas de tráfegos correspondentes, pacotes DICOM que necessitam de prioridade e assim aplica políticas de rede e de QoS com base na correlação. O autor expõe um modelo do arcabouço explanando um exemplo de uso e as suas possíveis vantagens.

O artigo (ALVAREZ; VARGAS SOLIS, 2013) explica as mudanças realizadas em uma instituição de saúde para adequar o ambiente de rede para a utilização de PACS, **RIS** (*Radiology Information System*) e o protocolo DICOM. Os autores propõem algumas mudanças no cabeamento da rede para categoria 6, servidores redundantes para a instalação do PACS, além de *switch* gigabytes para o tráfego DICOM. Os softwares utilizados para implementar o PACS

neste estudo de caso foi o DCM4CHEE e um software clientes denominado de K-pacs. Para manter um bom funcionamento da rede, o trabalho recomenda o uso do simulador de tráfego DICOM, PACSPulse, software de análise de tráfego DICOM, DICOM Network Analyzer e outros softwares consagrados de monitoramento como Wireshark, NTop e etc. A pesquisa conclui que é possível montar uma rede para tráfego de imagens médicas com uma alta disponibilidade e rendimento com o uso dos softwares propostos e realizando as mudanças descritas.

Capítulo 3

Caracterização e Modelagem de Tráfego

Neste capítulo são apresentados conceitos de caracterização (Seção 3.1) e técnicas de modelagem de tráfego (Seção 3.2).

3.1 Caracterização de Tráfego

A expansão das redes de dados tanto corporativas como domésticas tem proporcionado o aumento do interesse da comunidade científica em relação ao estudo da caracterização de tráfego de dados em redes de computadores. Isto se deve ao fato de que com a descoberta das características de um tráfego, torna-se possível um melhor entendimento do comportamento dos fluxos existentes na rede, possibilitando com isso o uso de técnicas de planejamento, desenvolvimento, monitoramento, gerenciamento ou configuração dos fluxos (LEE; LEVANTI; KIM, 2014). O uso destas técnicas pode proporcionar um melhor aproveitamento dos *links* de dados e uma melhor justiça na utilização da largura da banda da rede e dos dispositivos existentes. A caracterização possibilita a disponibilização de informações relevantes para experimentos de avaliação de desempenho ou planejamento de capacidade, permitindo a reprodução do comportamento do tráfego na rede com outros volumes de dados ou com dispositivos diferentes. Esses experimentos podem guiar os operadores de rede para a tomada de uma melhor decisão de possíveis expansões da infraestrutura ou modificações em sistemas.

Além dos pontos citados acima, a detecção de anomalias no uso da rede também é um benefício da caracterização, visto que o uso de forma indevida de alguns serviços, no ambiente

corporativo, pode atrapalhar o tráfego de serviços essenciais (LEE; LEVANTI; KIM, 2014).

A caracterização de tráfego é realizada através dos processos de monitoramento, classificação e identificação por classes de dados.

3.1.1 Monitoramento

O processo de monitoramento procura identificar padrões e tendências no tráfego da rede (LEE; LEVANTI; KIM, 2014). De acordo com os resultados encontrados no monitoramento, os operadores da rede podem reconfigurá-la na busca de melhores resultados. Como exemplo deste processo, um operador pode encontrar durante o monitoramento do tráfego, uma grande quantidade de tráfego de vídeo que está causando perda de pacotes de outros tipos de dados na rede. Para resolver esse problema o operador da rede pode limitar a taxa deste tipo de tráfego. A não detecção de situações como essa podem levar a interrupções dos serviços e prejuízos financeiros. O monitoramento utiliza técnicas de medição e amostragem para obtenção dos dados a serem analisados.

Medição de Tráfego

O processo de medição pode ser realizado com tráfego real local ou a partir de tráfego sintético injetado na rede.

De acordo com Finamore et al. (2011), a abordagem ativa adota a injeção de tráfego na rede, com o intuito de induzir um efeito mensurável para medição. Ou seja, a abordagem ativa gera tráfego de pacotes sintéticos e específicos em um ponto da rede, permitido assim a alteração do estado da rede para, por exemplo, simular a perda de pacotes de forma artificial. O complemento deste processo é a medição do comportamento da rede em um certo período. Um dos exemplos mais adequados para o uso dessa técnica é a localização de pontos de congestionamento que possam ter impacto no funcionamento da rede.

Já a medição passiva, contrariamente à medição ativa, busca a utilização de técnicas que permitam a análise do tráfego real que passa em determinados pontos de observação na rede (MARTINS, 2013). Essa técnica não interfere no fluxo da rede, pois coleta os pacotes "escutando" o tráfego passante na rede. Um grande problema desta abordagem é quando há um alto volume de tráfego, visto que a "escuta" e coleta destes dados podem gerar *traces*

com grande volume de dados. Os traços de rede ou *traces* como são comumente conhecidos na literatura especializada, guardam informações sobre os pacotes coletados e o tempo em que cada pacote foi detectado.

A técnica híbrida, como o próprio nome diz, utiliza as duas formas de medição.

Técnicas de Amostragem

Devido à grande quantidade de informações geradas na medição das redes, técnicas estatísticas podem ser utilizadas para diminuir o tamanho da amostra de dados coletadas em cada medição (CALLADO, 2009). A análise de informações a partir do tráfego observado na rede, necessariamente passa por uma seleção de uma boa amostragem de dados. As principais técnicas de amostragem conhecidas são a convencional, adaptativa e multi-adaptativa.

Técnica Convencional

Técnicas convencionais têm por princípio a utilização de regras fixas para inicializar e paralisar a captura dos dados. Essas regras podem ser sistemáticas ou aleatórias. As regras sistemáticas utilizam funções determinísticas para definir a duração da seleção dos dados. As regras aleatórias, em contraponto, utilizam funções aleatórias para definir o intervalo da amostragem (ZSEBY; MOLINA; DUFFIELD, 2009).

A Técnica de amostragem convencional pode funcionar com diferentes abordagens:

A abordagem **Sistemática Baseada em Contadores** realiza a definição do ponto de início e o ponto final de seleção das amostras de acordo com o posicionamento de chegada dos pacotes, ou seja, cada pacote será selecionado de acordo com o seu número de chegada em relação ao último pacote selecionado, sendo essa diferença entre pacotes de acordo com o valor definido na regra de seleção baseada na contagem do número de pacotes.

Outra abordagem **Sistemática Baseada no Tempo** é análoga a baseada em contagem, mas considera o tempo de chegada. Neste caso, os pacotes são selecionados de acordo com o tempo de chegada, e a separação entre as amostras a serem selecionadas se dá pelo tempo definido na regra.

No caso da abordagem **Aleatória Estratificada**, são selecionados n pacotes em cada coleta, dentro de cada intervalo de N pacotes. Esta técnica mistura intervalos fixos temporal ou por posição de pacotes.

Na abordagem **Aleatória Probabilística**, a seleção de um pacote será de acordo com uma

probabilidade predefinida. Essa técnica pode utilizar abordagem probabilística uniforme ou não uniformes. Para maiores detalhes veja (ZSEBY; MOLINA; DUFFIELD, 2009).

Técnica Adaptativa

A técnica adaptativa varia a frequência de amostragem de acordo com o estado da rede. Este processo permite uma redução do *overhead* em comparação com as técnicas convencionais. A sua adaptação é regida de acordo com a variação de métricas específicas como atraso, *jitter* ou perda de pacote.

Técnica Multi-adaptativa

Na técnica adaptativa apenas os intervalos entre as amostras variam, mantendo-se o tamanho das amostras. A abordagem multi-adaptativa procura variar o intervalos entre as amostras e também o tamanho das mesmas. Neste processo, caso a atividade da rede aumente, a variação de amostras coletadas também aumenta, mas o tamanho da amostra diminui para evitar sobrecarga do sistema de medição. O inverso também ocorre na busca de um equilíbrio.

Para escolher a melhor técnica de amostragem é preciso conhecer quais os principais pontos de coletas ou gargalos da rede. Também é necessário conhecer o comportamento do tráfego.

Além do monitoramento, o estudo da caracterização de tráfego também é realizado através da classificação e a identificação por classes de dados.

3.1.2 Classificação

A classificação é normalmente realizada por duas técnicas principais: a primeira técnica é a classificação baseada no conteúdo dos pacotes ou dos fluxos (HEADQUARTERS, 2008), que procura classificar os dados por características estatísticas dos tráfegos. Normalmente são analisadas as informações na perspectiva das portas de origem/destino da camada de transporte. A segunda técnica é baseada na análise do *payload* dos pacotes (CALLADO, 2009), que busca características de classificação mediante identificação de padrões das aplicações ou protocolos no *payload*.

A Figura 3.1 exibe uma metodologia de classificação de pacotes. Neste exemplo, os pacotes capturados são classificados com base em características dos dados do cabeçalho do pacote ou de assinaturas dos aplicativos. Após essa etapa, os pacotes são agregados

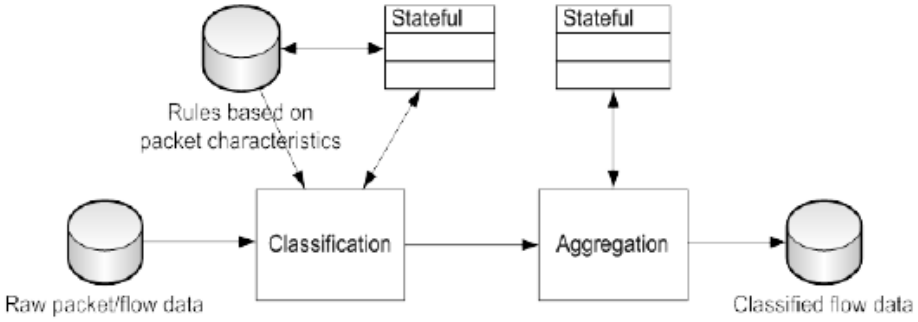

Figura 3.1: Metodologia de Classificação de Pacotes

Fonte: Callado et al., 2009

em fluxos, antes do armazenamento. As assinaturas de tráfego devem ser criadas antes da classificação e exigem o trabalho de um especialista. Elas também devem ser atualizadas com frequência para readaptá-las para novas aplicações.

3.1.3 Identificação de Tráfego

A identificação de tráfego é realizada por agrupamento de acordo com o comportamento dos fluxos de dados na rede. De acordo com Brownlee e Claffy (2002) os fluxos são classificados por classes de dados realizando uma analogia com as características de alguns animais. Podemos ter fluxos pequenos como "Ratos", fluxos grandes como "Elefantes", fluxos rápidos como "Libélulas" e fluxos lentos como "Tartarugas". Este tipo de classificação já é comum em estudos científicos da área de **TE** (*Traffic Engineering*). Essa analogia está presente no trabalho de (GUO; MATTA, 2001) e (PAPAGIANNAKI, 2002).

Para identificar um fluxos como elefante, é necessário que este ultrapasse a fase de início lento do TCP, portanto, o seu comportamento, incluindo a forma como ele interage com outras sessões TCP, é controlado por um retorno dos algoritmos de congestionamento do TCP. Já os fluxos ratos não ultrapassam a fase do início lento do TCP.

Uma outra forma de identificação dos fluxos elefantes é através da sua caracterização qualitativa. O trabalho de Mori e outros (2004) realizou o reconhecimento deste tipo de fluxo comparando o volume total de pacotes com o número de pacotes de um fluxo. Na definição de Mori e outros (2004) um fluxo será considerado elefante quando o número de pacotes de um fluxo for maior que 0,1% do total de pacotes da amostra.

Já os fluxos libélulas e tartarugas têm sua caracterização baseada no tempo dos fluxos. Para um fluxo ser considerado libélula, o seu tempo de vida não pode ultrapassar 2 segundos. Os fluxos entre 2 segundos e 15 minutos são considerados curtos e não possuem uma nomenclatura específica. No caso do fluxo tartarugas, seus tempos de vida deve durar mais do que 15 minutos.

A identificação destes tipos de tráfego pode ser usada para configuração e definição de rotas ou larguras de banda específicas para cada tipo de tráfego. Também é possível inserção de regras que eliminem tráfegos indesejados.

O processo de melhoria das redes de dados, através de caracterização, passa por um ciclo de trabalho. Após o monitoramento, classificação e identificação de tráfego, dependendo dos resultados encontrados, um redesenho da infraestrutura e das configurações da rede pode ser necessário. Esse redesenho pode ser implantado e um novo processo de monitoramento deve ser realizado para verificação dos resultados das mudanças e, caso seja necessário, realizar novos redesenhos. Este é um ciclo contínuo presente nas operações de gerenciamento da rede.

3.1.4 Metodologias de caracterização

Para uma caracterização eficiente do tráfego, faz-se necessário o uso de alguma metodologia de medição e análise de dados. Nos estudos realizados com essa finalidade, a variação do tráfego, o nível de fluxo, o nível de pacotes e as formas de distribuição dos dados são alguns pontos importantes a serem mensurados para caracterização.

O primeiro ponto, variação do tráfego, procura entender, dentro de certos períodos de tempo, qual o padrão de variação dos dados. Em (THOMPSON; MILLER; WILDER, 1997), um dos primeiros estudos em tráfego comercial, foi utilizada a escala de tempo de 24 horas e 7 dias em termos de volume de tráfego, volume de fluxo, duração do fluxo, tamanho dos pacotes e composição de tráfego por protocolos ou aplicações. Esses parâmetros de variação do tráfego buscam encontrar padrões de comportamento dos dados dentro da variação do tempo.

Outro ponto a ser mensurado é o nível de fluxo de uma rede. Para uma correta mensuração do nível de fluxo, faz-se necessário entender a definição de fluxo. Como definição de fluxo, neste trabalho é adotada a terminologia proposta por Brownlee e Claffy (2002),

onde um fluxo é um conjunto de pacotes trafegando em qualquer direção entre dois pares de *hosts*. Os mesmos autores basearam-se na **RTFM** (*Realtime Traffic Flows Measurement*) (HANDELMAN, 1999) que consiste de uma arquitetura de medição de tráfego por fluxo no qual a medição de um fluxo é composta por quatro atributos:

- Atributos de endereço de origem;
- Atributos de endereço de destino;
- Atributos de hora de início do fluxo;
- Atributos de hora de fim do fluxo.

Esses atributos em conjunto com a medição bidirecional definem um fluxo, e a caracterização do nível do fluxo dá-se pelo tempo que um conjunto de pacotes consome para trafegar entre dois pontos na rede em um certo período dentro de um fluxo. Em (BROWNLEE; CLAFFY, 2002) foi utilizado esse tipo de mensuração para analisar o comportamento do fluxo em relação ao tempo.

A caracterização a nível de pacote busca identificar a quantidade de pacotes trafegando na rede ou o tamanho dos pacotes de acordo com uma classificação. A caracterização em (DAINOTTI; PESCAPÉ; VENTRE, 2006) utiliza essa técnica para identificar os pacotes grandes trafegando na rede e sua influência em relação aos demais pacotes.

3.2 Modelagem de Tráfego

A ideia central da modelagem é analisar e representar os padrões de um tráfego em um ponto de vista específico. Através da representação destes padrões é possível criar uma versão simplificada dos dados, capaz de representar o comportamento do tráfego em determinadas situações. Com esses modelos é possível avaliar analiticamente, por meio de simulações e/ou medições, o comportamento de sistemas existentes ou que serão construídos no futuro, procurando assim, melhorar os resultados do desempenho, planejamento e das escolhas de investimentos nas mudanças da rede.

Modelos probabilísticos têm sido usados em telecomunicações desde o trabalho de Erlang (1910). Devido às chamadas telefônicas comportarem-se de forma semelhante ao pro-

cesso de Poisson, Erlang propôs um modelo de fila simples para representar o tráfego telefônico (SANTOS, 2009). Para as primeiras modelagens de tráfego, modelos de telecomunicação foram utilizados como base (CASTRO, 2010). O modelo de Poisson também foi proposto para a análise de filas em redes de pacotes. O modelo de Poisson pode ser definido como uma regra matemática que atribui probabilidade ao número de ocorrências de um evento (SANTOS, 2009). Poisson tem um único parâmetro, a taxa média de chegada λ. Assim, o intervalo entre chegadas de eventos tem uma distribuição exponencial com média $\frac{1}{\lambda}$ e são variáveis independentes. Modelos de Poisson são largamente utilizados na teoria das filas.

Outro modelo bastante utilizado é o auto-similar. Segundo alguns autores, o modelo de Poisson se mostrou incompleto para representar tráfego em rajadas, sendo considerados mais adequados o modelo de auto-similaridade (JAIN, 1991). Em 1990 um grupo de pesquisados analisou um grande volume de dados concluindo que a distribuição do intervalos entre chegadas dos pacotes possuía fortes evidências de auto-similaridade (WILLINGER, 2000). A partir dai, vários modelos evoluíram para tráfego auto-similar. Auto-similaridade refere-se a distribuições que apresentam as mesmas características em várias escalas no tempo.

Como visto anteriormente, modelos de tráfego são utilizado há bastante tempo. Para definição de um modelo de Tráfego são necessários 3 passos, sendo eles: problema da seleção do modelo, problema da estimação de parâmetros e validação do modelo.

A seleção do modelo busca um ou mais modelos que possam proporcionar uma boa descrição do tipo de tráfego.

Estimativa de parâmetros é baseada em um conjunto de medidas estatísticas, por exemplo, média, variância, função densidade de probabilidade ou função autocovariância, que são calculadas sobre os dados observados. O conjunto estatístico de medidas a ser utilizado no processo de inferência dos dados depende do impacto que podem ter nas principais métricas de desempenho (NOGUEIRA, 2003).

A validação do modelo usa testes estatísticos para avaliar se o modelo considerado é adequado para descrever o tipo de tráfego em análise. Alguns teste usados com essa finalidade são os teste de Kolmogorov-Smirnov, Anderson-Darling e Chi-Quadrado. O teste de Kolmogorov-Smirnov é usado para determinar se duas distribuições de probabilidade diferem uma da outra (MASSEY JR, 1951), sendo baseado na função de distribuição acumulativa. O teste Anderson-Darling (ANDERSON; DARLING, 1954) é uma modificação do teste de

Kolmogorov-Smirnov com mais ênfase nos valores da cauda. Já o teste de Chi-Quadrado, compara a frequência do intervalo da amostra com o valor que ela deveria ter se seguisse uma determinada distribuição de probabilidade, ou seja, ele nos diz com quanta certeza os valores observados podem ser aceitos como regidos pela distribuição em questão.

É importante salientar que modelos não replicam o comportamento do tráfego de forma idêntica, sendo eles apenas aproximações do comportamento dos dados reais (THYAGO ANTONELLO; CUNHA et al., 2008).

Distribuições de Probabilidade

Existem várias distribuições de probabilidade que são utilizadas comumente na literatura de modelagem de tráfego para representar como as variáveis aleatórias se comportam. Ao longo deste trabalho são citadas algumas delas, notadamente a distribuição Log-normal, Exponencial, Weibull, Gamma, Pareto e Dagum.

Distribuição Log-normal aparece naturalmente como um produto de várias variáveis independentes, sempre positivas. Existe uma relação entre as distribuições Log-normal e Normal. O logaritmo de uma variável que segue distribuição Log-normal com parâmetros μ e σ tem distribuição normal com média μ e desvio-padrão σ. Essa relação significa que dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, se considerarmos o logaritmo dos dados ao invés dos valores originais. A função densidade de probabilidade da distribuição log-normal é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi}\ t\sigma}\ \exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right], \qquad t\ >\ 0,$$

sendo $-\infty\ <\ \mu\ <\ \infty$ e $\sigma\ >\ 0$.

Na Figura 3.2 pode ser visto um exemplo do gráfico pdf da distribuição Log-normal com parâmetros $\mu = 1$ e $\sigma = 1$.

A distribuição Exponencial é geralmente aplicada a dados com forte assimetria como aqueles cujo histograma siga a forma da pdf como o exemplo exibido com $\lambda = 1$, na Figura 3.3. As características desta distribuição são: não simetria dos valores; a variável aleatória x assume somente valores positivos; é definida por um único parâmetro λ e é um caso especial da distribuição gama com o parâmetro $\lambda = 1$. Sua função densidade de

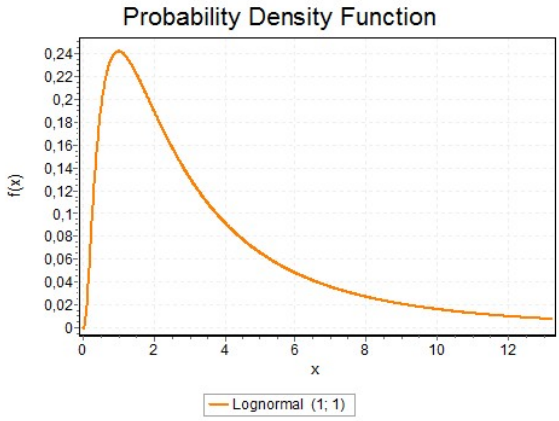

Figura 3.2: Função de Densidade de Probabilidade (pdf) - Lognormal

probabilidade é definida por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ se } x \geq 0 \\ 0 \text{ se } x < 0 \end{cases}$$

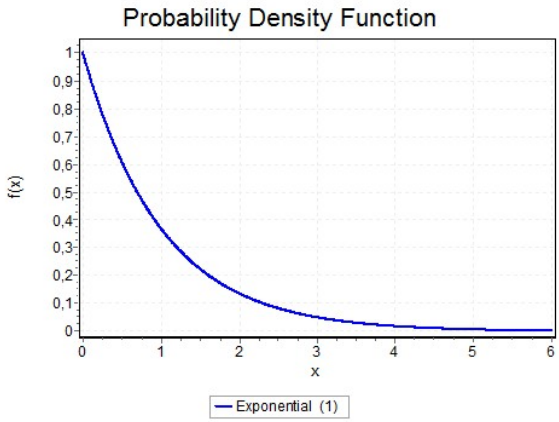

Figura 3.3: Função de Densidade de Probabilidade (pdf) - Exponencial

A distribuição de Weibull é nomeada assim devido a Waloddi Weibull que em 1951 lançou um artigo descrevendo a distribuição em detalhes e propondo diversas aplicações. O sucesso da distribuição se justifica pela sua capacidade de fazer previsões de acurácia bem razoáveis mesmo quando a quantidade de dados disponível é baixa. Podemos definir essa distribuição como uma variável aleatória x que tem sua função densidade de probabilidade igual a:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

O parâmetro λ está definido de 0 a $+\infty$ e é medido na mesma unidade que x. Graficamente a pdf da Weibull pode ser visto na Figura 3.4, com parâmetro $\lambda = 1$ e $k = 1,5$.

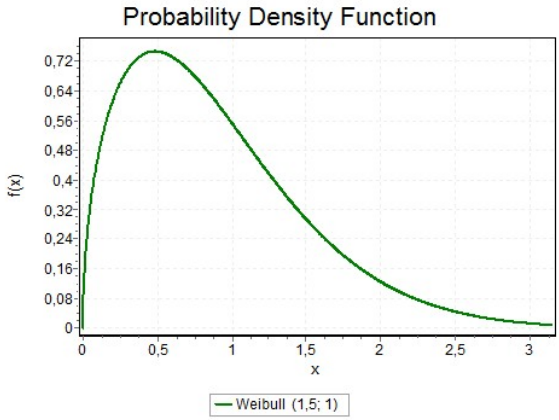

Figura 3.4: Função de Densidade de Probabilidade (pdf) - Weibull

A distribuição Gama é uma das distribuições mais gerais, pois várias distribuições são derivadas dela, por exemplo, exponencial e a qui-quadrado. Na Figura 3.5 pode-se ver a sua forma e a mesma pode ser definida como uma variável aleatória x que tem com parâmetros $\alpha > 0$, parâmetro de forma, e $\beta > 0$, parâmetro de taxa, denotando-se $x \sim$ Gama(α, β), se sua função densidade for dada por: $f(x) = \begin{cases} \dfrac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \text{ se } x \geq 0 \\ 0, \text{ caso contrário} \end{cases}$

Para o exemplo de pdf da Figura 3.5, $\alpha = 2$ e $\beta = 1$.

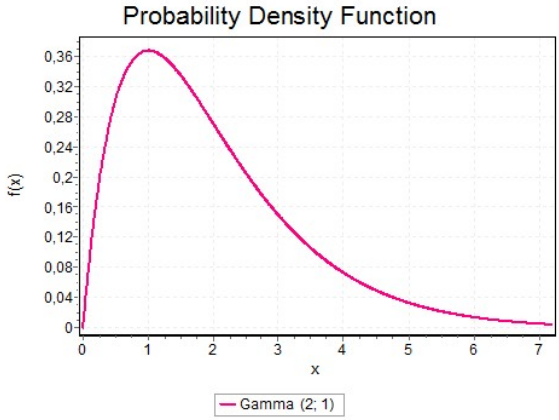

Figura 3.5: Função de Densidade de Probabilidade (pdf) - Gamma

A distribuição Pareto, em homenagem ao engenheiro civil italiano Vilfredo Pareto, é uma distribuição de probabilidade que é usada na descrição científica, geofísica e muitas

outras áreas do conhecimento para modelar vários tipos de fenômenos observáveis. Seja X a variável aleatória que segue uma lei e parâmetros de Pareto, com x um número real positivo, então essa distribuição caracteriza-se por: $\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geq x_\mathrm{m}, \\ 1 & x < x_\mathrm{m}. \end{cases}$

onde x_m é o valor mínimo possível de X, sempre positivo, e α é um parâmetro positivo. Um exemplo de gráfico da pdf de Pareto com parâmetros $\alpha = 5$ e $\beta = 1$ encontra-se na Figura 3.6.

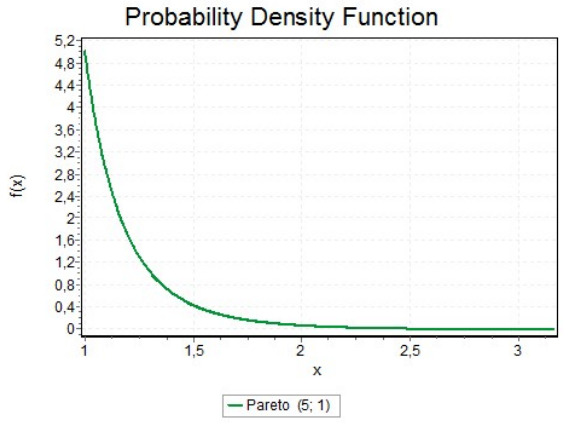

Figura 3.6: Função de Densidade de Probabilidade (pdf) - Pareto

Por último a distribuição de probabilidade contínua definida sobre todos os números reais positivos nomeada de Dagum. Tem esse nome devido a Camilo Dagum, que a propôs em uma série de artigos na década de 1970. A distribuição de Dagum é especificada de duas formas: um modelo com três parâmetros e outro com quatro parâmetros. A função de densidade de probabilidade desta distribuição para o modelo com três parâmetros α, β e k é definida por:

$$f(x) = \frac{ak\left(\frac{x}{\beta}\right)^{ak-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{k+1}}$$

A Figura 3.7 apresenta graficamente um exemplo de função de densidade de probabilidade da distribuição de Dagum com parâmetros $\alpha = 2$, $\beta = 1$ e $k = 3$.

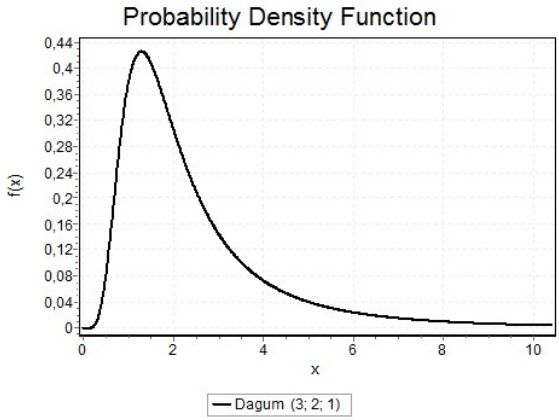

Figura 3.7: Função de Densidade de Probabilidade (pdf) - Dagum

Análise Estatística

O uso de análise estatística busca definir uma distribuição estatística teórica que represente o conjunto de dados coletados. O método utilizado para isso é a técnica de *fitting*. *Fitting* é o procedimento de seleção de uma distribuição estatística teórica que melhor se adapta a um conjunto de dados gerado por algum processo aleatório. Em outras palavras, se você tiver alguns dados aleatórios disponíveis, e quiser saber qual distribuição em particular pode ser usada para descrever estes dados, *fitting* é o método que você está procurando.

Para definição de uma distribuição adequada através de *fitting* são necessários 3 passos, similares aos de modelagem de tráfego: problema da seleção do modelo, problema da estimação de parâmetros e validação do modelo

O problema da seleção do modelo baseia-se na escolha da distribuição com as melhores características a serem modeladas. Na escolha da distribuição deve-se levar em consideração a natureza dos dados, visto que dependendo de suas características, distribuição probabilística discreta ou contínua podem ser utilizadas. Uma distribuição é considerada discreta se assumir um número enumerável de valores inteiros como, por exemplo, os valores de um dado ou a quantidade de pacotes trafegados. Já uma distribuição é considerada contínua se assumir uma quantidade não enumerável de valores, por exemplo, peso, voltagem, tamanho de arquivo e etc. A seleção de uma distribuição deve ser feita através da escolha de um modelo de distribuição que mais se assemelha com os dados a serem analisados e que tenha o menor desvio entre o modelo escolhido e os dados em análise. Como exemplo, no caso da análise do tamanho de arquivos, não faz sentido pensar em distribuições que aceitem nú-

meros negativos entre seus valores. Outro ponto importante, é a realização do emprego de métodos estatísticos para encontrar quais distribuições possuem maior semelhança entre os dados empíricos e os teóricos. Os testes mais utilizados para a realização deste tarefa são Kolmogorov-Smirnov, Anderson-Darling e Chi-Quadrado. Neste ponto, as distribuições que obtiverem os melhores resultados nos testes serão analisadas na próxima etapa.

O próximo passo, problema da estimação de parâmetros, tem como objetivo escolher os parâmetros que deixam a distribuição teórica mais parecida com a distribuição empírica. A parametrização de uma distribuição é um processo de abstração, que possibilita a flexibilização de uma distribuição. Uma distribuição parametrizada constitui na transformação de uma distribuição padrão em uma distribuição nova e individual, amparada pelos valores particulares do conjunto de parâmetros. A realização desta tarefa é normalmente realizada com o método da Máxima Verossimilhança do inglês MLE (*Maximum Likelihood Estimation*). De acordo com Law e Kelton (1982), o MLE de uma distribuição são os parâmetros dessa função que maximizam a semelhança da distribuição com o conjunto de dados observados.

Como terceiro passo temos a validação do modelo. Nesta etapa é analisado se a semelhança entre os dados empíricos e teóricos é boa o suficiente para validar a distribuição definida. Os valores obtidos com os testes estatísticos no primeiro passo serão utilizados para rejeitar ou validar a distribuição baseado em um grau de significância escolhido. Um valor de grau de significância normalmente usado é de 0,05. Uma distribuição pesquisada poderá ser rejeitada caso o resultado do valor estatístico dos testes esteja com valores fora da margem de significância definida. O valor estatístico dos testes é normalmente definido através do p-value. O p-value é um indicador do valor limiar do nível de significância no sentido de que a hipótese nula (H_0) será aceita ou rejeitada.

3.3 Trabalhos Relacionados

3.3.1 Caracterização

Muito antes do surgimento do serviço web na internet, o comportamento do tráfego das redes vem sendo estudado por vários pesquisadores. Na área médica, ainda são poucos os estudos sobre o tráfego deste tipo de dados. Entretanto, algumas pesquisas têm sido produzidas com

o intuito de conhecer o comportamento deste tipo de tráfego. Nos próximos parágrafos são apresentados trabalhos que realizaram classificação, identificação de tráfego, metodologias de caracterização e modelagem de tráfego.

Classificação de Tráfego

Nesta seção são apresentados estudos que realizaram a classificação de tráfego em vários tipos de dados, no ensejo de conhecer essa forma de caracterização.

Sobre caracterização, Thompson e outros (1997) apresentam padrões e características do tráfego na Internet, coletados com uma ferramenta de monitoramento de tráfego para troncos OC3. Neste trabalho, os autores coletaram mais de 240.000 fluxos em um *backbone* comercial pelo período de 24 horas durante 7 dias. Foram analisadas as características do tráfego em relação a várias métricas e em relação à composição dos protocolos e aplicações. Os resultados expostos indicaram que o protocolo TCP era o protocolo dominante nos *links* medidos. Em um dia o TCP correspondeu a 95% dos bytes, 90% dos pacotes e 75% dos fluxos. No caso das aplicações, constatou-se que o tráfego Web era dominante com 75% dos bytes e fluxos e 70% dos pacotes. Além destas métricas já informadas, o estudo faz análise de várias outras métricas, tais como, largura de banda, protocolos mais utilizados, média do tamanho de pacotes, entre outros.

Dainotti e outros (2006) expõem os resultados da caracterização a nível de pacotes com o objetivo de encontrar padrões espaciais e temporais de aplicações baseadas em tráfego TCP. O trabalho desenvolveu uma metodologia para construção, a nível de pacote, das características estatísticas do tráfego de uma rede. Os testes foram realizados sobre o tráfego HTTP e SMTP. Os resultados mostraram invariância temporal e espacial nos ambientes medidos. Os autores também modelaram o tráfego para uso em plataformas de simulação. Na caracterização realizada, o padrão encontrado possuía semelhança com distribuição Lognormal quando analisados sobre os valores de tempo entre pacotes, Já para o tempo dos fluxos, a distribuição Weibull foi a mais adequada.

Em (PLOUMIDIS; PAPADOPOULI; KARAGIANNIS, 2007), procurou-se caracterizar o tráfego de dados de uma rede *wireless* de uma universidade da Carolina do Norte nos Estados Unidos. O estudo foi realizado sobre 382 pontos de acesso (*AP - Access Point*) do campus universitário e tinha como objetivo descobrir as principais aplicações em uso na rede e a

composição do tráfego. O estudo revelou que 35,6% do tráfego da rede era Web e 30,04% era de P2P, ou seja, mais de 65% do tráfego era dominado por apenas dois tipos de aplicações. Já em relação ao número de pacotes para esses dois tipos de aplicação, esse número é ainda maior chegando a 46,8% Web e 34,46% P2P, representando juntos mais de 80% dos pacotes.

As análises de Finamore e outros (2011) apresentam um estudo de caracterização de tráfego utilizando a ferramenta TStat. Neste estudo, realizado em alguns ISPs, os resultados obtidos demonstram que diferentemente de outros trabalhos analisados pelos autores, a utilização de P2P que estava em queda voltou a aumentar. Devido a essa mudança de perfil do tráfego P2P apresentado, o estudo ressalta a importância da manutenção do monitoramento do tráfego de dados. Outro dado contestado pelo pesquisa é a crença de que tráfego UDP é insignificante. Os autores informam que este cenário mudou e o tráfego UDP aumentou consideravelmente chegando a aproximadamente 30% do tráfego em alguns casos.

O trabalho (CHIMMANEE; PATPITUCK, 2013) apresenta a característica de tráfego DICOM nas redes com fios e sem fios com base nas medidas do intervalo de tempo dos pacotes, na distribuição de tamanho de pacote, em bytes, e variação da perda de pacotes. Todos os resultados foram obtidos sobre tráfego real. Os valores observados nas medições dos intervalos de tempo dos pacotes na rede com fio foram bem diferentes dos encontrados nos intervalos de tempo do ambiente da rede sem fio, enquanto que na distribuição de tamanho de pacote DICOM, também na rede com fio em relação a sem fio foram quase os mesmos, concentrando a maioria dos pacotes nas faixas de 54 a 64 e 1024 a 1518 bytes. No estudo de perda de pacotes, a média de perda com o uso da ferramenta ping apresentou um melhor resultado na rede com fio do que na sem fio, com diferença de quase 3 vezes nos valores observados. Em alguns casos, na rede sem fio, os valores de perda de pacotes chegaram a 8,95% enquanto que em testes similares, na rede com fio, esses valores não ultrapassaram a 1,7%.

Identificação de Tráfego

Essa subseção apresenta os trabalhos de caracterização por identificação por tipo de tráfego. Este tipo de estudo analisou o tráfego nas redes avaliando a velocidade e o volume dos fluxos de dados.

Guo e Matta (2001) analisaram a influência do tamanho dos dados trafegados na rede sobre o seu desempenho. Os autores discorrem sobre o fenômeno dos "Ratos" e "Elefantes" relatando que a maioria do tráfego na Internet são considerados pequenos, "Ratos", enquanto que uma pequena fração de conexões trafegam dados grandes, "Elefantes". O artigo expõe que sem a existência de uma política de gerenciamento ativo de filas que privilegie conexões pequenas em detrimento às grandes, as primeiras tendem a perder a competição por banda de rede em favor das segundas. Nas simulações realizadas, os resultados obtidos com os algoritmos de gerenciamento de filas padrão e configurações baseadas na arquitetura **Diffserv** (*Differentiated Services*) não foram satisfatórios, levando os autores a propor um novo algoritmo de política de fila chamado de RIO (RED *with In and Out*) que, segundo eles, ajuda a melhorar o desempenho de fluxos de dados pequenos e balanceia de forma mais justa o tráfego de dados.

Brownlee e Claffy (2002) descrevem um método de medição do tamanho dos dados de *streams* de vídeos trafegados na Internet e o tempo de vida deles. Com esse método foi realizada a caracterização da distribuição deste tipo de tráfego em dois sites diferentes. A pesquisa indica que 45% do tráfego dos *streams* medidos levam menos de 2 segundos de tempo de vida, chamados neste trabalho de tráfegos "Libélulas", enquanto que 1,5% do tráfego restante é considerado tráfego lento, chamados de tráfegos "Tartarugas" com mais de 15 segundos de duração. O percentual restante são dados não *streams*. Complementando o estudo, também são analisados os tamanhos dos pacotes trafegados. O artigo indica que 95% do tamanho do tráfego possuía uma média de 15 Kbytes.

A pesquisa de Papagiannaki e outros (2002) apresenta um esquema de classificação de fluxos elefantes baseado em um ponto que os fluxos devem atingir para serem considerados elefantes. Esse trabalho indica uma forma de classificação que incorpora características temporais que, segundo os autores, é mais bem sucedida na identificação de tráfegos elefantes. Eles propõem uma técnica para identificação de fluxos elefantes considerando-o nesta categoria quando este se encontra na cauda da distribuição dos fluxos e com um tempo de fluxo maior que um valor definido pelos autores. Essa forma de definição de fluxo leva em conta a natureza da distribuição dos fluxos estudados que tem formato de cauda pesada, segundo os autores. Nos dois *links* analisados, com amostragem de 24 horas foram encontrados 500 e 600 fluxos elefantes respectivamente.

Na investigação de Megyesi e Molnár (2013), os autores buscam estudar a relação entre as características dos fluxos e o perfil dos usuários da Internet. O autor usa o termo "Usuário Elefante" proposto em outro trabalho para classificar o perfil dos usuários. O perfil é calculado usando o coeficiente de GINI para encontrar o coeficiente de distribuição do número de bytes gerados pelos usuários. Os resultados comparando o volume de tráfego dos usuários elefantes com os fluxos elefantes existentes na rede, mostram que há características semelhantes. Entretanto, apenas 10% a 30% dos fluxos elefantes encontrados foram gerados por usuários elefantes.

Mori e outros (2004) estudaram a identificação de fluxo elefantes baseado no teorema de Bayes. O objetivo do trabalho foi desenvolver um arcabouço que identificasse fluxos elefantes em amostras periódicas de pacotes para uso em links com alta velocidade de transmissão. Segundo os autores, o arcabouço proposto por eles é bem genérico e realiza de forma apropriada o reconhecimento de falso positivo e falso negativo na análise dos fluxos. Na investigação, identificou-se que fluxos elefantes normalmente possuem alto volume de tráfego em baixa quantidade de fluxos. Para o processo de identificação de fluxo pelo arcabouço desenvolvido, foi definido que seriam considerados fluxos elefantes todo fluxo que correspondessem em sua composição de pacotes a mais do que 0,1% do total de pacotes da amostra. Os resultados obtidos com um dos *traces* analisados mostraram que 0,02% de todos os fluxos deste *trace* correspondiam a 59,3% do volume total de tráfego.

Um outro trabalho que estudou a identificação de tráfego é o de Lan e Heidemann (2006). Nele são estudadas as correlações entre os vários tipos identificados de fluxos: por tamanho, "Rato" e "Elefante"; por duração, "Tartaruga" e "Libélula"; por rajada, "Chita" e "Porco-espinho". As explorações das correlações entre os vários tipos de fluxos existentes mostraram que existem uma forte relação entre o tamanho dos dados e a taxa de fluxo. De acordo com os autores, os resultados indicaram que os fluxos elefante são de longa duração, mas não são rápidos e não realizam rajadas. Os tráfegos tartaruga são considerados lentos e também não realizam rajadas. Os fluxos chita são tipicamente pequenos, e realizam rajadas. Finalmente, os fluxos porco-espinho são ponderados como fluxos grandes e rápidos. O artigo conclui que fluxo por tamanho e duração devem ser tratados de forma diferenciada e independente.

3.3.2 Métodos de Caracterização

As metodologias permitem organizar e padronizar a forma de caracterização, possibilitando a sua utilização em outros estudos. As informações expostas aqui serviram de inspiração para elaboração de uma proposta metodológica mais específica para caracterização de tráfego de imagens médicas.

A metodologia utilizada por Thompson e outros (1997) para caracterização de tráfego fez-se, inicialmente, através de coletas de dados em duas escalas de tempo, sendo elas 24 horas e 7 dias da semana, medindo o volume de tráfegos, volume e duração de fluxos, composição de tráfego dos protocolos IP, TCP e UDP, composição de tráfego dos protocolos da camada de aplicação e tamanho dos pacotes. Os dados foram coletados em dois *backbones* comerciais da internet para efeito de comparação. Dois pontos de monitoramento foram colocados nos nós dos links do *backbone* entre o roteador do core e o *switch* para realização das coletas. As primeiras análises foram realizadas sobre os dados coletados, mensurando o volume do tráfego global em termos de bytes, pacotes e fluxos. Os autores demostraram a variação média diária do tamanho dos pacotes e a distribuição do tamanho dos pacotes. Concluindo a análise, são explanadas as composições do tráfego em relação ao tempo sobre o protocolo IP, TCP, UDP e os protocolos de aplicação.

Em Dainotti e outros (2006) a metodologia de caracterização foi baseada na decomposição do tráfego de rede em conversações, sendo essas conversações definidas como o intervalo de tempo durante o qual dois *hosts* diferentes trocam pacotes pertencentes a uma associação a nível dos protocolos de aplicação, separados por um período fixo de tempo de silêncio e do tipo endereço de origem e endereço de destino. Na proposta deles, ficou definido que pertencerá à mesma conversa, todos os pacotes de origem e de destino à porta TCP 80 **HTTP** (*Hypertext Transfer Protocol*) e porta TCP 25 **SMTP** (*Simple Mail Transfer Protocol*), viajando entre os dois *hosts*, com um tempo limite de inatividade de 15 minutos. A abordagem utilizada considera todo o tráfego que aconteceu durante a conversa como um fluxo único bi-direcional de dados, que é dividido em *upstream*, que é o tráfego do cliente para o servidor e *downstream*, o tráfego do servidor para o cliente. O trabalho estudou separadamente *upstream* e *downstream* estimando o tamanho e distribuição dos pacotes em um intervalo de tempo. Um aspecto importante da metodologia é que não foram levado em consideração pacotes com *payload* vazios e todo o tráfego específico do TCP tais como pacotes de

estabelecimento e confirmação de conexão. O período de coleta das amostras utilizadas pelo estudo compreenderam 5 dias da semana. Os dados coletados possuem uma resolução de 10 ms entre os pacotes. Além disso, foram comparados os resultados das medições realizadas com distribuições teórica selecionadas.

Outra metodologia analisada, Ploumidis e outros (2007), realizou-se caracterização de uma rede sem fio na perspectiva dos Clientes, Rede e Pontos de Acesso. O método proposto realizou coletas em um roteador de acesso da universidade e também coletou logs de 488 pontos de acesso (*AP - Access Points*). Os logs dos APs foram usados para obtenção do endereço MAC e informações de dados a nível de fluxo dos APs. No cruzamento dos dados coletados através de *traces* com os endereços MAC dos logs, foram detectados 9.125 endereços IP internos mapeados para 3.241 endereços MACs, isto foi necessário para identificar a quantidade de clientes e APs geradores de tráfego na rede. Os dados foram coletados durante 7,5 dias. Também foi realizada a comparação entre este estudo em rede sem fio com outros estudos anteriores em ambiente de rede com fio e sem fio. Os autores utilizaram como forma de classificação, na perspectiva de rede, os tipos de aplicação por número de bytes, fluxos e popularidade entre os usuários da rede. Na perspectiva dos APs, foi analisada a distribuição de fluxo por APs. Já na perspectiva do Cliente, utilizou-se os dois critérios, tipos de aplicação e distribuição de fluxo.

Finamore e outros (2011) basearam-se na classificação do volume de tráfego por tipos de aplicações da pilha de protocolo TCP/IP para sua análise. No estudo do protocolo TCP, o monitoramento dos dados foi focado em dois tipos de aplicação, sendo elas HTTP e P2P. Já no tipo de protocolo UDP, o foco foi nos tipos de dados de aplicação P2P. Essa escolha dos autores se deve a estudos anteriores que foram realizados com este mesmo foco e que servem de comparativo com o trabalho atual. Além dessa forma de análise, os autores utilizaram coletas mensais totalizando 10 meses em 5 **ISP** (*Internet Service Provider*) distintos.

Chimmanee e outros (2013) utilizaram como metodologia a inserção de fluxos de dados de imagens médicas, com acréscimo linear no número de fluxos, na rede de um hospital de grande porte. O processo utilizou coletas contínuas de 1 a 6 fluxos, mensurando a distribuição do tamanho, perdas e intervalo de tempo entre os pacotes, tudo isso sobre dois tipos de ambientes de rede, com fio e sem fio. Foram realizados 3 experimentos, sendo o primeiro um comparativo do intervalo de tempo entre os pacotes, o segundo a distribuição do tamanho

dos pacotes em 6 intervalos que variam entre 54 bytes e 1518 bytes e por último o efeito de variação dos pacotes perdidos com o uso da ferramenta ping em conjunto com a carga de tráfego de imagens médicas.

3.3.3 Modelagem de Tráfego

Os vários estudos sobre a modelagem de tráfego ampliaram as possibilidades de melhoria na gestão do tráfego. As linhas seguintes expõem alguns trabalhos sobre esta área de pesquisa.

Em (DOWNEY, 2001) foi descrita a modelagem baseada no tamanho dos arquivos. Downey expõe que há vários artigos sobre modelagem de tamanhos de arquivo, tanto local quanto web, que consideram que a distribuição de tamanho de arquivos pode ser modelada através de distribuições com cauda longa, constantemente utilizando a distribuição de Pareto como modelo. O trabalho dele se opõe a esse modelo e defende que a distribuição do tamanho de arquivos, na grande maioria dos sistemas, se adapta à distribuição Lognormal. Para realizar a comparação entre os dados empíricos e as distribuições propostas como modelo, o processo de *fitting* foi usado. Para avaliação do modelo no processo de *fitting*, foi utilizado o teste estatístico de Kolmogorov-Smirnov. Os resultados encontraram fortes indícios que a distribuição Lognormal é mais apropriada para representar o tamanho dos arquivos que a distribuição de Pareto. Também não foram encontradas evidências que a distribuição de tamanho de arquivos possuem necessariamente característica de cauda longa.

Castro e outros (2010) modelaram a distribuição de tamanho de pacotes do tráfego na rede. Os experimentos realizados contaram com vários tipos de pacotes de dados incluindo os da internet, vídeos, download de arquivos com P2P, FTP entre outros. Os resultados obtidos nos experimentos foram similares aos resultados obtidos por Rastin (2009) e Tafvelin (2007). Cerca de 90% dos pacotes UDP foram considerados pequenos e com tamanho abaixo de 500 bytes. No caso dos pacotes TCP, 40% dos pacotes tinham por volta de 44 bytes e outros 40% com tamanho próximo a 1500 bytes. Diante dos dados foi proposto um modelo matemático, criado pelos autores, para estimar a distribuição de tamanho dos pacotes.

Acompanhando os novos serviços que surgem na internet, Salvador e outros (2014) estudaram o serviço de P2P-TV pela internet. O trabalho apresenta como resultados a caracterização deste tipo de tráfego e o modelo do tráfego de dados recebido e enviado dos canais CNN, ESPN e Sky, realizando esta modelagem com o uso da técnica de *fitting*. A

validação do processo foi realizada com o teste de Kolmogorov-Smirnov comparando as distribuições das taxas de transmissões enviadas e recebidas, com distribuições teóricas. Os modelos escolhidos para representar o tráfego recebido dos 3 canais foram Laplace, Pareto e Extreme-Value, respectivamente. Já para as tráfegos de envio, as distribuições definidas foram Gamma para CNN e ESPN e Pareto para Sky.

Experimentos e Simulações com PACS e DICOM

Os trabalhos descritos nessa subseção explanam formas de experimentos e simulações com PACS e tráfego DICOM. As informações colhidas nestes artigos permitem a escolha de possíveis técnicas para realização de experimento ou simulações com esses tipos de dados.

Arney e outros (2012) buscam quantificar a QoS (*Quality of Service*) necessária para o tráfego de dados médicos nas larguras de banda de 100 Mbps, 500 Mbps e 1 Gbps utilizando o protocolo DICOM. O autores simularam o tráfego DICOM junto com outros tráfegos da rede de um hospital e concluiu propondo o isolamento do tráfego DICOM devido ao grande fluxo de imagens, pois o mesmo pode degradar os outros serviços na rede. Os resultados encontrados indicam um aumento significativo no tempo de resposta da rede com o tráfego DICOM, saindo de 0,002 a 0,003 segundos nos experimentos sem DICOM para 12,434 a 50,191 nos experimentos com DICOM.

Chimmanee (2013) procurou estender a metodologia Smetric, que foi projetada especificamente para suportar **VoIP** (*Voice over Internet Protocol*) e Aplicações Telnet com SSH (*Secure Shell*), para uso com tráfego de PACS. O mecanismo de qualificação da metodologia Smetric foi modificado para um métrica de qualificação por fluxo, QFlows (Qualificação por Fluxo), por ser mais adequada para o uso com tráfego de imagens médicas do que a qualificação por pacotes QPKTS (Pacotes Qualificados). Ele usa técnicas matemáticas de regressão para quantificar alterações no congestionamento. O autor realizou vários experimentos e encontrou indícios de alta precisão nos resultado da nova métrica proposta. A precisão para regressão linear foi de 93,26%. Nas regressões quadrada e cúbica os valores obtidos foram 89,56% e 86,81%, respectivamente.

Challita e outros (2013) demostram os tipos de QoS para **WIMAX** (*Worldwide Interoperability for Microwave Access*) e algoritmos de escalonamento para alocação de banda, propondo o uso de uma intra-classe baseada na emergência dos dados médicos. O artigo

expõe a necessidade de uma classe específica para priorização de dados médicos, visto que alguns dados, de mesma classe, precisam ter prioridade no fluxo mais que outros. Nas simulações realizadas pelo autor conclui-se que o **WRR** (*Weighted Round Robin*), algoritmo de escalonamento proposto pelo autor para a priorização de tráfego com o WIMAX, pode respeitar os graus de importância e emergências dos dados médicos.

Singh et al. (2014) discorrem sobre o estudo de QoS sobre dados de saúde para melhoria da qualidade das redes. Nele foi simulado o ambiente de um hospital na Índia e medida a quantidade de pacotes adequados para cada tipo de tráfego médico nas classes de QoS. O artigo explica a arquitetura da rede de saúde e apresenta uma tabela com a classificação de diferentes tipos de dados na área de saúde por classe de QoS. Nas simulações realizadas as métricas utilizadas foram *Jitter* e *Throughput*, calculadas através de medições fim a fim. Nos testes feitos em rede TCP, os resultados mostram que pacotes com até 500 bytes são mais adequados para tipos de dados classificados na tabela nas classes de 0 a 2, que são compostas, por exemplo, por aplicações que trafegam dados com informações do tipo consulta médica através de teleconsulta, tráfego DICOM e etc. Já nos tipos de dados classificados nas classes de 3 a 5, dados do tipo documentos clínico, banco de dados em tempo real e etc, o tamanho mais indicado é de 2500 bytes. No caso de redes **UDP** (*User Datagram Protocol*) os resultados foram iguais aos do TCP, exceto nos tipos de dados das classes 3 e 5 onde o tamanho mais indicado é a partir de 1500 bytes.

Pode ser visto nos estudos analisados, que a caracterização de dados médicos que trafegam nas redes são escassos e quando existem são realizados através de simulações ou em ambiente com homogeneidade de equipamentos. Diante deste cenário, a caracterização e modelagem deste tipo de dados com informações de ambiente real pode facilitar as tomadas de decisão sobre implantações futuras de configurações ou novas infraestruturas. Com relação à identificação de tráfego, não foram encontrados estudos direcionados ao tráfego de imagens ou dados médicos, indicando neste caso uma grande lacuna de conhecimento sobre esses dados. Algoritmos de gestão de tráfego, como o reencaminhamento ou balanceamento de carga, podem explorar os dados identificados através desta técnica de caracterização para assim propor tratamento para estes fluxos, buscando com isso possíveis melhorias na rede.

Capítulo 4

Metodologia Proposta

De acordo com Lee e Kim (2014) as operações de gerenciamento das redes de computadores envolvem um ciclo contínuo de monitoramento, projeto e implantação de melhorias. Essa pesquisa pretende com os seus resultados, propor uma forma de análise de tráfego que ajude no gerenciamento, projeção e simulação de tráfegos hospitalares. Neste trabalho, técnicas de monitoramento são utilizadas para realizar a caracterização do tráfego de redes hospitalares em conjunto com técnicas de modelagem para definição de um modelo de entrada de dados, do tráfego DICOM. Mais especificamente, o foco deste trabalho é a análise do comportamento de tráfego DICOM. Com base nessas análises, uma metodologia para caracterização de tráfego e modelagem de fonte de dados foi proposta para auxiliar na obtenção de informações que permitam o diagnóstico de problemas e o aperfeiçoamento da qualidade dos serviços da rede.

A metodologia proposta nesta dissertação caracteriza o tráfego de um hospital em conjunto com o tráfego DICOM. Ademais, um modelo da fonte de dados da modalidade de ultrassom foi indicado. A caracterização do tráfego de imagens médicas é realizada através da medição, classificação e identificação do tráfego de dados. Já o processo de modelagem da fonte de dados, utiliza o método de *fittting* entre as distribuições de dados coletadas, tamanho dos arquivos, e as distribuições probabilísticas para sugestão de um modelo de fonte de dados. Finalizando, um planejamento de simulação foi definido para a realização de estudo de caso para o uso do modelo de tráfego. Toda essa metodologia é resumida no fluxograma da Figura 4.1.

Como pode ser visto na Figura 4.1, a metodologia é definida em 2 etapas e 2 passos:

caracterização de tráfego DICOM e modelagem da fonte de dados DICOM mais definição da política de transferência dos dados e avaliação de desempenho.

A definição de uma política de transferência deve seguir os processos definidos pelo ambiente em estudo. Essa política de transferência influencia no volume de tráfego que estará presente na rede e pode alterar o processo de coleta para caracterização e modelagem do tráfego.

O processo de caracterização se inicia com a coleta do tráfego necessário para análise do comportamento do tipo de tráfego que se deseja entender. A coleta é realizada através de medições do tráfego utilizando técnicas de medição, esse procedimento está descrito na Subseção 4.2.1. O próximo passo é a análise dos traces utilizando métodos de filtragem e classificação. A filtragem e classificação, consiste no processamento estatístico dos traces para obtenção do resultado do comportamento do tráfego. Maiores detalhes são disponíveis na Subseção 4.2.2.

O processo de modelagem busca definir uma distribuição de probabilidade que possibilite a geração de tráfego sintético para simulações. Na Figura 4.1 no lado esquerdo pode ser vista a sequência definida. O primeiro passo é a coleta dos dados diretamente na modalidade. Esse processo pode ser feito de forma manual ou automática. Logo em seguida, com os dados coletados, a modelagem deve ser realizada definindo uma distribuição para representar a fonte de dados de uma modalidade. Maiores detalhes sobre a forma de modelagem podem ser visto na Seção 4.3.

Por último, são realizados experimentos de simulação através da geração de tráfego sintético para avaliação de desempenho e demonstração do funcionamento do modelo.

4.1 Política de Transferências DICOM

O método de transferências dos dados DICOM e a configuração das modalidades envolvidas devem ser analisados para um correto processo de caracterização.

O processo de transferência dos dados devem ser realizados seguindo algumas políticas do ambiente a ser monitorado. A primeira análise a ser feita é a forma como o sistema PACS está configurado para transferência de dados DICOM, pois essa forma influencia no volume de tráfego na rede. No caso do hospital onde foram realizadas as coletas, as transferências

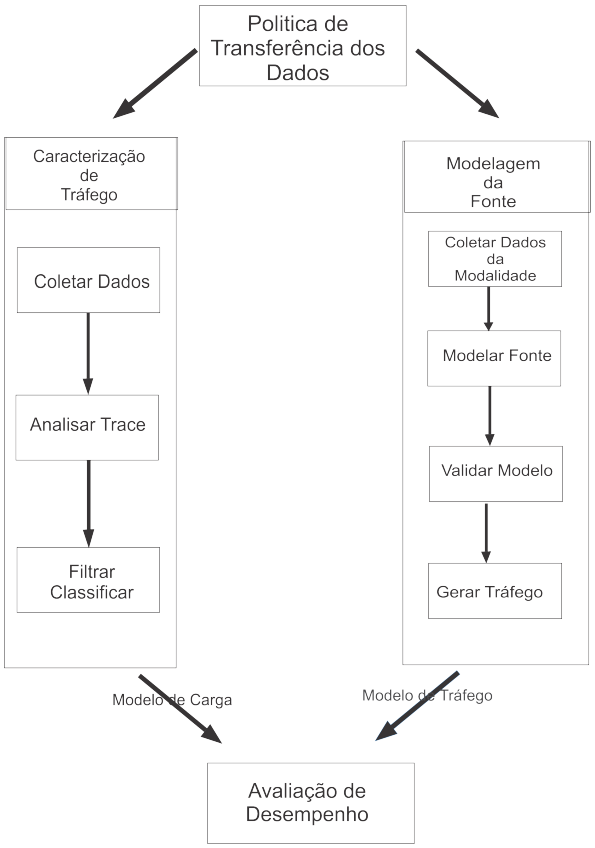

Figura 4.1: Metodologia de Caracterização e Modelagem de Tráfego

das imagens médicas geradas pela modalidade não são feitas de forma automática. O entendimento da política de transferência e das definições dos parâmetros configurados como, por exemplo, o tamanho da MTU definida no equipamento e a método de transferência dos exames, automático ou manual alteram as características do tráfego.

No caso em estudo, a pedido dos médicos, os dados gerados pelos equipamentos devem ficar armazenados na própria modalidade por um período mínimo de 1 mês. Após esse período, os dados podem ser transferidos de forma manual para o PACS. Nesta situação, as coletas foram realizadas somente no momento das transferências. Os dados foram transferidos para o PACS em lotes de exames. Esta forma de trabalho amplia a quantidade de tráfego na rede. Em ambientes no qual a política de transferência dos dados for feita de forma automática, o processo de coleta deve ser realizado de forma contínua.

4.2 Caracterização de Tráfego

O processo de caracterização de tráfego inicia-se com a definição da frequência das coletas a serem realizadas e a escolha da técnica de coleta das amostras a serem medidas. A frequência e a quantidade de amostras dar-se-ão de acordo com características do ambiente a ser medido. Como exemplo, podemos dizer que um ambiente com transmissão constante dos dados a serem analisados, provavelmente terá que ter medições contínuas durante 24 horas e vários dias da semana. Já a técnica de coletas das amostras devem ser definidas conforme informações expostas na Capítulo 3. A escolha da técnica correta pode facilitar o obtenção dos resultados desejados. A técnica de amostragem depende da disponibilidade de recursos de armazenamento.

Neste trabalho foram escolhidos, como periodicidade de coletas dos dados, horários específicos de transmissão das imagens médicas iguais aos horários normalmente utilizados pelo hospital para realizar suas transmissões, realizando-os em período noturno com coletas em vários dias da semana para obtenção de uma amostragem mais real. Como forma de amostragem, foi utilizada a técnica convencional para coletas das informações com regras fixas para coletas dos dados, tanto no tempo quanto na frequência. No caso desta metodologia foram utilizadas coletas de 1 hora por dia em 5 dias da semana.

As medições do tráfego no ambiente a ser analisado podem ser realizadas com software de captura de rede *open source* ou software proprietário. O coletor de tráfego deve coletar todos os pacotes que trafegam na rede ou apenas o cabeçalho do pacote. Nos casos em que a análise dos dados serão feitas pelo *payload*, a coleta dos pacotes deve ser completa. Em outros casos, fazem-se necessários apenas os cabeçalhos dos pacotes. Para o estudo do tráfego das imagens médicas, apenas os cabeçalhos dos pacotes são suficientes. Essa forma de coleta preserva a confidencialidade dos dados dos pacientes evitando a sua exposição de forma desnecessária.

O método de coleta dos dados é realizado com a interligação de um *host* contendo um coletor de tráfego conectado a um *switch* core da rede. A escolha do ponto de coleta levou em consideração o local da rede onde existam as maiores concentrações de fluxos e pacotes trafegando. Além disso, é necessária a configuração deste *switch* do core para realizar o espelhamento de suas portas com a porta onde está interligado o *host* contendo o coletor de

rede. Esse procedimento faz-se necessário para que o software de coleta de dados possa ter acesso a todos os pacotes que trafegam na rede. No caso deste trabalho, o software utilizado para coleta dos dados foi o TCPDump. O TCPDump é um software de coleta de tráfego de dados da rede com licença *open source* e utilizado por vários trabalhos científicos para estudo de tráfego.

4.2.1 Forma de Coleta de Dados

Neste estudo foram realizadas 5 coletas de dados com aproximadamente um hora de duração. Em cada uma destas horas coletadas foram feitas quatro transferências correspondente a quatro dias de exames realizados no mês anterior. Essas quantidades de dados escolhidas buscou replicar o modo como os dados são transferidos de forma cotidiana no hospital em estudo. Além disso, as movimentação de dados foram realizadas em dias da semana para uma maior diversidade dos dados.

4.2.2 Filtragem e Classificação

Com os dados coletados, o próximo passo é o processamento estatístico dos pacotes coletados e a gravação destes dados processados em *traces* para análise e plotagem dos resultados. Para o processamento estatístico dos pacotes ficou definida a resolução entre as amostras de dados de 1 segundo. Os recursos utilizados para geração das informações a serem analisadas foram os softwares TCPStat e Wireshark. Esses dois softwares são também *open source*. Para análise estatística foram utilizados os softwares R (FOUNDATION, 2014) e Easyfit (TECHNOLOGIES, 2015).

4.3 Modelagem da Fonte de Dados

As modalidades são as principais fontes de dados no sistema em estudo. A partir do estudo prévio de caracterização, observou-se que o tráfego gerado pela modalidade ultrassom demonstra comportamento de elefante. Como os dados DICOM são transmitidos agrupados, em conexão TCP, verificou-se que o tamanho do bloco de dados influencia diretamente o tráfego, devido ao controle de congestionamento do TCP. Diante disso, a modelagem da fonte

destes dados foi estabelecida. A modelagem da fonte de dados das imagens médicas se inicia com a realização de coletas de dados direto da modalidade de ultrassom. Como as imagens são transmitidas em blocos, não é possível identificar cada estudo DICOM individualmente ou em um fluxo único. Essas coletas foram realizadas durante 4 meses e os dados coletados compreenderam o período de 1 ano de exames realizados. Nesta pesquisa, todo o processo de coletas destes dados foram feitas de forma manual, diretamente no aparelho de ultrassom, devido ao bloqueio inserido pelo fabricante do equipamento no acesso a base de dados dos exames direto no aparelho.

A partir das coletas, utilizando análise estatística dos dados, deve ser criado um modelo de fonte de dados de imagens médicas. O processo de análise da fonte de dados dar-se-á por comparação entre a fonte de dados e a distribuições probabilísticas (SALVADOR, 2014). Esse processo de comparação é conhecido como curva de ajuste. Os resultados deste processo de *fitting* fornecem uma distribuição que melhor se encaixa com os dados coletados na fonte de dados, podendo essa distribuição ser utilizada como parâmetro para simulação. Todo o processo de cálculo e testes estatísticos, para comparação das distribuições, foram realizados com software *open source* R e o software de análise estatística Easyfit.

O próximo passo desta metodologia foi a realização da escolha das distribuições "candidatas" a representar os dados empíricos. Os dados a serem modelados foram os tamanhos dos arquivos dos exames que são transmitidos pelo equipamento de ultrassom. A escolha das possíveis distribuições a serem analisadas para representar os dados devem seguir a natureza dos dados. Nesta caso, primeiramente verifica-se os valores enumeráveis são do tipo inteiros ou reais. Caso os dados analisados sejam inteiros, devem ser selecionadas distribuições discretas. Já no caso dos valores serem números reais, as distribuições contínuas devem ser utilizadas. Para modelagem do tamanho dos arquivos da fonte de dados, distribuições contínuas foram utilizadas.

Após a escolha das possíveis distribuições, métodos estatísticos devem ser utilizados para encontrar quais distribuições maximizam a semelhança entre os dados coletados e a distribuições probabilísticas. Nesta metodologia foram usados os testes estatísticos Kolmogorov-Smirnov, Anderson-Darling e Chi-Quadrado. A utilização destes testes permitem a seleção das distribuições que mais se aproximam dos dados.

Outro passo na modelagem por distribuição probabilística é a estimação de parâmetros

das distribuições. Essa estimação tem como objetivo deixar a distribuição selecionada o mais parecida com a distribuição real. Para estimação foi utilizada a máxima verossimilhança como método. Esse processo ocorre modificando parâmetros das distribuições até que os mesmos fiquem o mais semelhantes possíveis com a distribuição que se deseja modelar.

Finalmente para validação do modelo, utilizou-se os resultados dos testes estatísticos realizados para aceitação da distribuição que obteve melhor aproximação com os dados reais. Para isso, o nível de significância dos testes foram ajustados para 0,05. Esse nível de significância tem como objetivo obtermos um modelo que se comporte em 95% dos casos de forma parecida com a distribuição analisada.

Após a realização da modelagem, testes por intermédio de simulador foram realizados para demonstrar a usabilidade do modelo. Os resultados da caracterização e das medições realizadas com o tráfegos de imagens médicas foram usados para auxiliar no processo de configuração das simulações, inclusive no tráfego de *backgroud*. A avaliação de desempenho será feita com geração de tráfego sintético com o software IPerf (NLANR/DAST, 2015).

4.4 Planejamento da Avaliação de desempenho

Para a realização de estudos de desempenho com o modelo de fontes definido, cenários de rede experimental foram construídos a partir de emuladores como Mininet (LANTZ; HELLER; MCKEOWN, 2010). O ambiente a ser utilizado para experimentos deve aproximar as condições da rede operacional. Estudos de avaliação de desempenho podem ser executados com o objetivo de avaliar projetos de expansão da rede, avaliação da inclusão de novas modalidades e ampliação do atendimento ou planejamento de capacidades do uso das modalidades.

Um cenário de rede operacional comum em sistemas hospitalares é composto por uma sub-rede com servidores, identificada como zona de servidores, interligada a sub-rede com estações de trabalho e outra sub-rede interconectando as modalidades radiológicas.

Os serviços disponibilizados nesses sistema incluem o envio de estudos DICOM das modalidades para o servidor PACS, localizado na zona de servidores, o envio de estudos DICOM do servidor PACS para as estações localizadas na zona de estações de trabalho, a transmissão de dados de outras aplicações entre estações de trabalho e servidores.

Medidas de desempenho como a vazão, perdas e atraso podem ser utilizadas para avaliar

pontos de gargalo representados pelo compartilhamento da conexão dos servidores com a zona de estações de trabalho e a zona de modalidades.

Neste trabalho, alguns parâmetros foram definidos para a execução de estudos de caso:

1. número de modalidades;
2. tamanho dos arquivos com blocos de estudos DICOM;
3. intervalo entre transmissão de estudos DICOM;
4. intervalo entre pacotes de dados de outras aplicações.

Alguns fatores causam maior impacto nas medidas de desempenho. O número de modalidades que podem transmitir dados simultaneamente é um fator que deve ser avaliado com mais atenção. Outro fator importante é o tamanho do arquivo DICOM, pois esses arquivos podem ultrapassar a fase de início lento do TCP e se comportarem como elefante.

O processo de geração de tráfego foi realizado com o software Iperf. Parâmetros de tamanho do bloco de dados do Iperf foram utilizados para representar o tamanho do bloco de dados DICOM. O Iperf é uma ferramenta de medição *open source* para a máxima largura de banda disponível na rede. Essa ferramenta utiliza tanto protocolo TCP quanto UDP para injeção de tráfego na rede. Para os experimentos realizados neste trabalho, o TCP foi utilizado como protocolo de transporte tanto para a transmissão de dados DICOM assim como para as demais aplicações representadas pelo tráfego de *background*.

A geração de carga sintética foi realizada a partir de dois *scripts* que representavam a geração de tráfego *background* e a transmissão de dados DICOM a partir das modalidades. No primeiro caso, para o tráfego de *backgroud*, os valores do tamanho médio de pacotes e o intervalo entre pacotes, calculados durante a caracterização, foram utilizados como parâmetros de um modelo de tráfego de Poisson. Já para a transmissão de dados DICOM, a distribuição de Dagum foi utilizada para representar o tamanho de arquivos DICOM. O intervalo entre transmissões de pacotes foi definido como um parâmetro informado pelo operador. O intervalo entre transmissões de arquivos DICOM deve ser definido de acordo com o experimento, pois ele tem influencia diretamente no total de tráfego agregado.

Para o processo de medição foram utilizados o TCPDump, TCPStat e Wireshark.

Capítulo 5

Caracterização e Modelagem de Tráfego DICOM

5.1 Caracterização Realizada

Como explanado anteriormente, as pesquisas sobre caracterização de tráfego são importantes para a gestão das redes. Neste trabalho foi realizada a caracterização do tráfego de imagens médicas em conjunto com outros tráfegos existentes na rede, com o intuito de disponibilizar informações sobre o comportamento deste tipo de tráfego.

Para realizar a caracterização, foram utilizadas as principais técnicas de medição, classificação e identificação de tráfego da metodologia proposta. No caso da classificação, realizou-se a análise por pacotes e bytes por segundo. No âmbito da identificação, a análise do comportamento dos fluxos de dados das imagens médicas e dos outros fluxos da rede foram realizadas de acordo com a metodologia proposta no Capítulo 4.

5.1.1 Estudos Preliminares

Projetos de expansão do número de modalidades, com a aquisição de novos equipamentos, motivaram a análise do tráfego DICOM. Os primeiros estudos realizados sobre o tráfego de imagens médicas, neste trabalho, mostraram grandes picos de tráfego dos dados DICOM em relação ao restante do tráfego da rede. Como pode ser observado na Figura 5.2, o tráfego DICOM possui picos de tráfego nos tempos de 17:15, 18:45, 19:50 e 20:45. Estes picos

possuem vazão alta em um curto período de tempo. diante destas primeiras informações foram verificado o tamanho dos pacotes DICOM em relação a outros tráfegos. A Figura 5.1 mostra um gráfico com o tamanho dos pacotes em relação ao tempo. Podemos ver que o tamanho dos pacotes DICOM são relativamente grandes em relação ao tamanho dos outros pacotes no ambiente analisado.

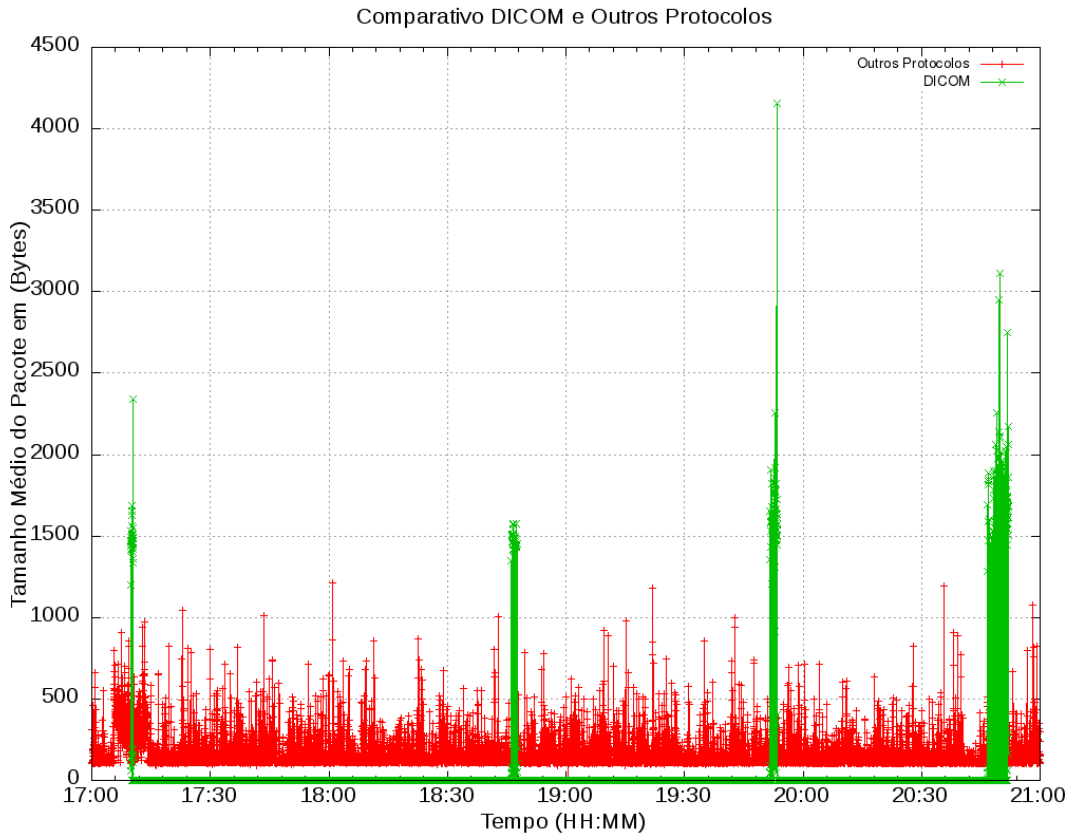

Figura 5.1: Tamanho médio dos pacotes comparando 4 amostras DICOM com outros protocolos

Diante destes primeiros gráficos, um estudo mais aprofundado sobre as características deste tipo de tráfego nas redes hospitalares tornou-se necessário. O objetivo do estudo proposto era avaliar o impacto do tráfego DICOM ao compartilhar a rede com outros tipos de tráfego, assim como com outras fontes oriundas de novos equipamentos.

Em Brownlee e Claffy (2002) foi estudado a identificação de tráfego da internet com foco principalmente no protocolo HTTP, devido ao grande número de aplicações que utilizam serviços web. Seguindo a mesma linha de raciocínio Brownlee e Claffy (2002), neste trabalho foram identificados como principais tipos de protocolos o DICOM e HTTP.

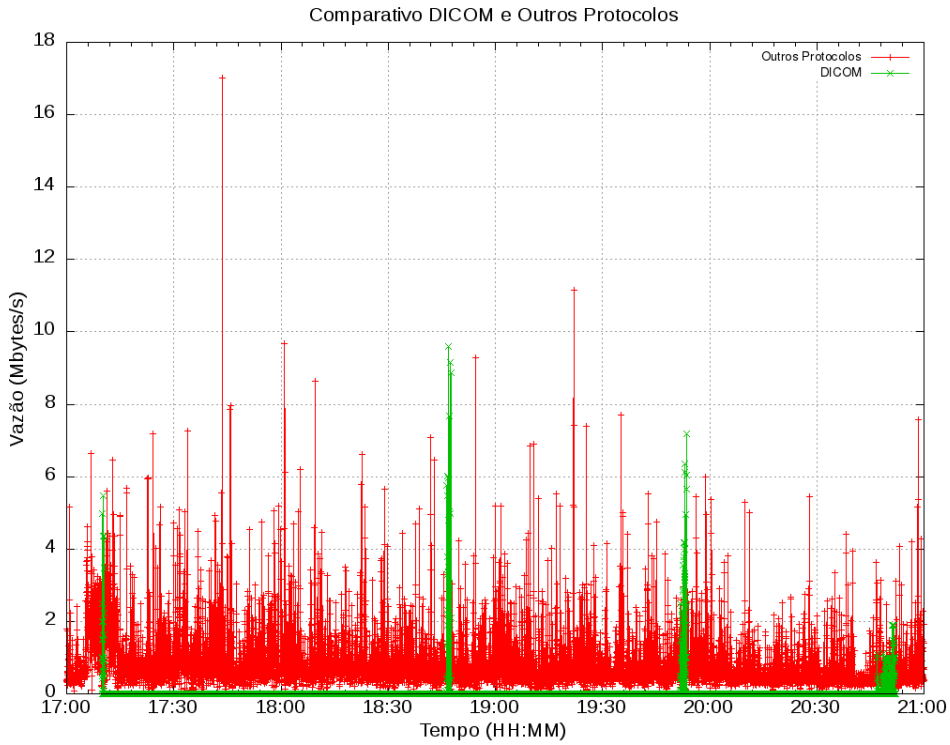

Figura 5.2: Vazão do tráfego da rede comparando 4 amostras DICOM com outros protocolos

A análise inicial dos dados apresentados na Tabela 5.1, indicaram que 98,6% do tráfego DICOM são libélulas. Já os resultados obtidos com protocolo HTTP, os resultados indicavam 73,54% do tráfego possuíam mais de 15 min de duração, ou seja, tráfego tartaruga. Esses percentuais foram bem diferente do trabalho Brownlee e Claffy (2002), visto que os valores obtidos por eles correspondentes ao tráfego HTTP, identificados como tartaruga, era de apenas 1,5% do tráfego. Esse fato ocorreu como esperado, visto que uma grande evolução dos serviços web aconteceu nos últimos anos e por consequência aumentou o tráfego HTTP. Essas primeiras coletas foram realizadas no dia 13 de agosto de 2014.

Após essas primeiras análises, estudos mais detalhados sobre o tráfego de imagens médicas foram realizados.

5.1.2 Estudo de Caso

A caracterização aqui apresentada é proveniente da aplicação da metodologia de caracterização a uma rede hospitalar real. O hospital possui uma modalidade de Ultrassom e um servidor PACS para armazenamento das imagens médicas geradas e está em processo de

Tabela 5.1: Identificação de Tráfego por Fluxos Preliminar

Protocolo	Total	Até 2 seg.	Entre 2 seg. a 15 min.	Mais de 15 min.
HTTP	35,68%	4,74%	21,72%	73,54%
DICOM	0,06%	73,43%	25,87%	0,7%
Outros	64,36%	93,37%	3,73%	2,09%

aquisição de um Tomógrafo Computadorizado e um Raio-X Digital. Devido a essas novas aquisições, a caracterização deste dados foram propostas aqui para ajudar no entendimento do comportamento destes tráfegos na rede e assim auxiliar os operadores da infraestrutura de rede nas decisões de possíveis adequações necessárias para inclusão destes novos equipamentos no ambiente. Entendendo as características deste tráfego, experimentos podem ser realizados a partir da adoção de modelos de tráfego com essas mesmas características.

As transmissões das imagens médicas neste ambiente hospitalar, possuem algumas particularidades como já descritas anteriormente. Elas são realizadas de forma manual e em blocos com vários exames de um dia, aumentando com isso a quantidade de tráfego transferido de forma simultânea. Em outras instituições hospitalares este processo pode ser diferente, realizando as transferências exame por exame ao término da realização deles de acordo com a política da instituição. Essa outra forma de transferência não foi analisada neste estudo.

O ambiente medido é formado por uma infraestrutura similar ao da Figura 5.3. Neste ambiente foram coletados dados de transferências realizadas entre a zona de equipamentos médicos e a zona dos servidores, além de outros dados que trafegam na rede. Essas transferências de dados seguiram as seguintes direções conforme diagrama da Figura 5.3: no caso do tráfego geral da rede, foram coletados tráfegos da zona de equipamentos médicos para a zona dos servidores, da zona das estações de trabalho para a zona dos servidores e da zona dos servidores para a zona das estações de trabalho. No caso do tráfego DICOM, foram coletados dados apenas da zona dos equipamentos médicos para a zonas dos servidores, visto que devido à política de trabalho do hospital, não há ainda transferência de dados entre a zonas dos servidores, o servidor PACS, e a zona das estações de trabalho. A escolha destas direções tem como base a forma como a rede foi concebida e a forma como as informações médicas trafegam. Na primeira direção de tráfego, os dados gerais da rede e os dados de imagens médicas trafegam entre os servidores e os equipamentos médicos com vários tipos de

protocolo. Além disso, as estações de trabalho consomem vários serviços na rede vindo dos servidores e da Internet. Na segunda direção de tráfego, os equipamentos médicos enviam as imagens médicas geradas na modalidade para o servidor através do protocolo DICOM. Essa forma de infraestrutura do PACS foi descrita na Seção 2.2 como arquitetura centralizada.

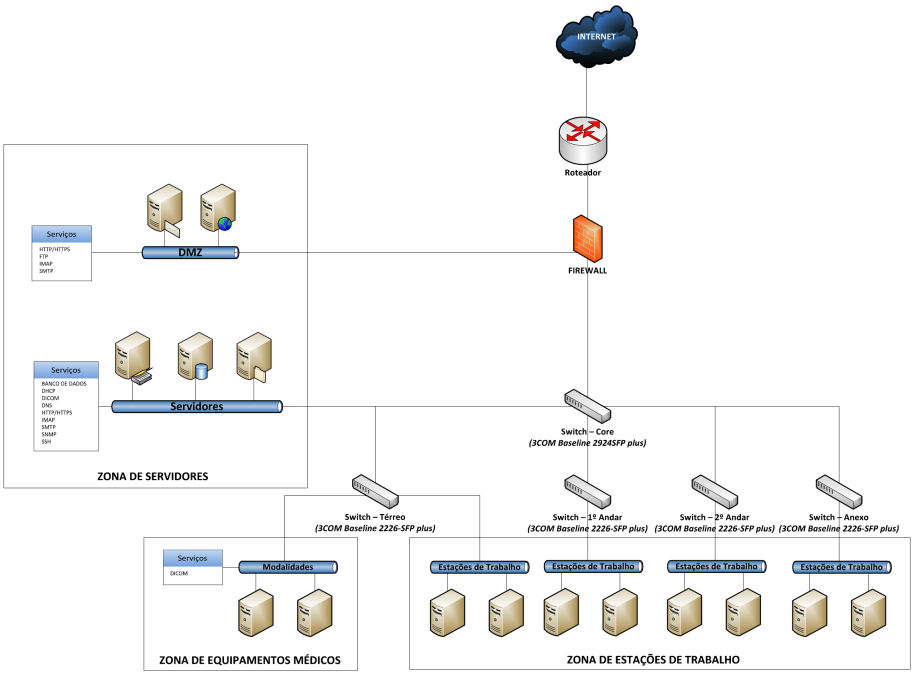

Figura 5.3: Ambiente de Rede do Estudo

O ambiente de rede existente é composto de um enlace *Ethernet* variando de 1 a 100 Mbps entre a zona das estações de trabalho, a zona das modalidades e os *switches* de borda. O restante do ambiente é composto de um enlace *Ethernet* variando de 1 a 1000 Mbps entre os *switches* de borda, o *switch* do core e os servidores.

Recursos do Sistema

Os equipamentos do *core* são formados por 5 servidores físicos com 2 processadores quad-core de 2,8 GHz, 8 Gb de memória RAM e sistema operacional VMWare ESXi, 2 servidores físicos com 2 processadores octa-core de 3,2 GHz, 128 Gb de memória RAM e sistema operacional VMWare Sphere 5, além de 17 máquinas virtuais para provimento de serviços aos usuários, sendo uma delas o servidor de armazenamento de imagens médicas com sistema operacional Linux e o software DCM4CHEE realizando o papel de PACS. Também há 2 servidores de banco de dados com 2 processadores quad-core de 2,8 GHz, 8 Gb de memória RAM e sistema operacional Linux Red Hat 4 e 2 servidores de banco de dados

com 2 processadores octa-core de 3,2 GHz, 128 Gb de memória RAM e sistema operacional Linux Oracle.

O Switch Core é um 3COM layer 3 com 24 portas 100/1000 Mbps e os *switchs* de borda são 3COM layer 2 com 24 portas 10/100 Mbps e 2 portas 100/1000 Mbps, estando estas portas interligadas ao *switch Core*.

Além dos equipamentos do *core* existem 106 estações de trabalho com Intel core i5, 4 Gb de memória RAM e sistema operacional Windows 7 ou Windows XP.

Os detalhes dos recursos do servidor PACS e da modalidade utilizados nas medições deste trabalho são os seguintes:

- Sistemas Operacionais:
  - PACS, DCMLinux Ubuntu Server 11.10;
  - Modalidade Ultrassom, Windows XP.
- Hardware:
  - Servidor virtual PACS, processador dual-core 3,2 GHz, 2 GB RAM, HD 320GB;
  - Modalidade Ultrassom, Processador Intel 3,0 GHz, HD 160 GB.

Recursos do Monitoramento

Para a coleta dos dados foi utilizado como hardware um notebook com processador core i7 e 8 GB de memória RAM com sistema operacional Ubuntu 14.04. Neste notebook o software TCPDUMP foi utilizado para coleta dos dados e o software TCPStat para filtragem dos *traces*. Os resultados foram armazenados em um arquivo com a extensão .pcap gerado com o TCPDUMP. Os dados foram coletados através da conexão do notebook no *Switch* Core. O *Switch Core* foi configurado para realizar espelhamento das portas conectadas nos servidores e aos outros *Switches*, com a porta conectada no notebook.

5.1.3 Coletas de tráfego

As medições de tráfego foram realizadas no período entre os dias 03 e 09 de fevereiro de 2015 no horário das 20:57 horas às 22:04 horas. A carga de trabalho da medição foi gerada pelos usuários da rede e a modalidade de Ultrassom. As coletas foram realizadas de forma

sistemática e seguindo a metodologia proposta neste trabalho, buscando obter dados que contemplassem uma melhor amostragem do ambiente real. As coletas foram feitas conforme Tabela 5.2 e metodologia definida no Capítulo 4.

Como exposto na Tabela 5.2, o tamanho das coletas variam de acordo com o fluxo de dados de cada dia. O período de coleta foi definido para iniciar próximo do final das 20 horas e o término para o início das 22 horas. A exceção a este horário se faz no dia 07/02/2015, um sábado, onde as coletas foram realizadas entre o final das 17 horas e o início das 19 horas. O tempo de cada coleta ficou em média de 1 hora e 4 minutos.

Tabela 5.2: Coletas de Tráfego Realizadas para Classificação

Data da Coleta	Tamanho em Gb	Horário de Inicio	Horário de Fim	Tempo da Coleta
03/02/2015	3,25	20:57:53	22:03:46	01:05:52
05/02/2015	2,43	20:58:07	22:02:30	01:04:22
06/02/2015	3,93	20:58:53	22:04:42	01:05:49
07/02/2015	1,87	17:58:05	19:05:03	01:06:57
09/02/2015	1,74	21:00:29	22:03:01	01:02:32

A forma como as coletas estão dispostas tem a função de contemplar amostras de dias da semana diferentes, proporcionando dados mais reais do ambiente. O horário das coletas e das transmissões seguiram a política definida pela instituição de saúde, realizando transmissões de blocos completos de dias de exames dentro de uma hora por dia. Em cada dia há 4 transmissões de imagens médicas, sendo que cada transmissão possui uma dia inteiro de exames.

As informações com os dados de imagens médicas que estão nas coletas se encontram na Tabela 5.3. Como pode ser verificar no dia 03 de fevereiro de 2015, foram transmitidos dados referente a bloco de exames dos dias 12, 13, 14 e 15 de janeiro de 2015. Esses conjuntos de exames possuem para cada dia os tamanhos de 91,48; 49,92; 81,09 e 110,2 MB respectivamente, perfazendo um total de 332,69 MB de dados transmitidos do Ultrassom naquela data. Nos outros dias de coletas, as transmissões seguiram esses mesmos padrões. Essa organização das coletas permitiu que as amostras obtidas contivessem dados de exames transmitidos correspondentes à totalidade dos exames de janeiro de 2015 divididos em grupos de dados, contemplando vários dias das semanas, completando assim um ciclo de

Tabela 5.3: Data e Tamanho dos Exames X Dias das Coletas

Data da Coleta	**Data dos Exames / Tamanho dos Lotes de Exames Enviados (MB)**			
03/02/2015	12/01/2015	13/01/2015	14/01/2015	15/01/2015
	91,48	49,92	81,09	110,2
05/02/2015	19/01/2015	20/01/2015	21/01/2015	22/01/2015
	159,4	61,59	73,87	111,3
06/02/2015	26/01/2015	27/01/2015	28/01/2015	29/01/2015
	85,3	140,9	84,79	146,2
07/02/2015	02/02/2015	03/02/2015	04/02/2015	05/02/2015
	153,4	47,63	83,58	104,8
09/02/2015	02/01/2015	05/01/2015	07/01/2015	08/01/2015
	32,1	97,81	94,86	78,43

transmissões referentes a um mês.

Tabela 5.4: Estatística Geral das Coletas

Data da Coleta	**Total pkts**	**Média pkt/s**	**Média Tam. pkts (Bytes)**	**Média Mbit/s**	Tempo entre pkts (ms)
03/02/2015	8.247.765	2.086,849	407,923	6,810	2,086849
05/02/2015	3.666.425	949,241	698,429	5,304	0,949241
06/02/2015	9.679.893	2.451,145	420,663	8,249	2,451145
07/02/2015	5.875.712	1.462,689	326,258	3,818	1,462689
09/02/2015	3.428.078	913,701	531,282	3,883	0,913701

Na Tabela 5.4 temos informações estatísticas gerais sobre as coletas de tráfego. Os dados coletados possuem uma variação na quantidade total de pacotes entre 3.428.078 e 9.679.893. A variação na quantidade de pacotes influencia diretamente a média de pacotes por segundo e o tempo entre pacotes. Nas amostras com maior quantidade de pacotes, como a apresentada no dia 06/02/2015, a média de pacotes por segundo aumenta de forma proporcional à quantidade de pacotes. Já o tamanho médio dos pacotes e a média de megabits por segundo não variam na mesma proporção que a quantidade de pacotes das amostras. Em relação ao tempo entre pacotes, os valores aumentam quando a quantidade de pacotes também aumenta e o contrário também é verdadeiro. O tamanho médio dos pacotes variaram de 326,258 a

698,429 e a média de megabits por segundo de 3,818 a 8,249. A variação destes dados são influenciadas pela largura de banda da rede e o controle de congestionamento do TCP.

5.1.4 Classificação dos Dados

Como exposto no Capítulo 3, existem várias formas de classificação dos dados coletados. Neste trabalho foram utilizadas a classificação por porta, pacotes e por fluxos. A classificação por porta utilizada nos pacotes coletados segue as definições da IANA (*Internet Assigned Numbers Authority*) (ASSIGNED NAMES; NUMBERS, 2014). Foram selecionados os tráfegos de dados que utilizavam a porta do protocolo DICOM. Já as formas de classificação por pacotes e fluxos foram realizadas de acordo com Thompson e outros (1997), Brownlee e Claffy (2002) , Dainotti e outros (2006) e Arney e outros (2012).

Na classificação por porta, foi classificada a quantidade de pacotes e *bytes* para o tráfego DICOM e para todo o tráfego da rede conforme dados apresentados na Tabela 5.5. A quantidade de pacotes DICOM nas amostras variaram de 200.006 na coleta do dia 09/02/2015 a 287.776 na coleta do dia 06/02/2015. Pode-se verificar que a quantidade de pacotes não passou de 7,117% do total geral de pacotes no maior valor. Já a quantidade de bytes DICOM existentes nas amostras obtiveram valores entre 344.558.942 e 510.534.501 também nos dias 09/02/2015 e 06/02/2015. Pode ser visto que o percentual de bytes trafegado chega a 22,982% em uma das amostras. Comparando o percentual de bytes com a quantidade de pacotes, podemos verificar que o volume de tráfego DICOM possui um valor considerável em relação ao total de tráfego na rede, mesmo utilizando um número pequeno de pacotes em comparação ao todo.

Tabela 5.5: Estatística DICOM das Coletas

Data da Coleta	Total pkts	DICOM pkts	Total Bytes	Bytes DICOM	Total Mbit/s	Mbit/s DICOM
03/02/2015	8.247.765	214.054	3.364.449.510	379.276.601	6,810	1,131
05/02/2015	3.666.425	260.928	2.560.737.238	457.696.261	5,304	1,315
06/02/2015	9.679.893	287.776	4.071.976.809	510.534.501	8,249	1,334
07/02/2015	5.875.712	250.499	1.916.998.644	440.564.322	3,818	1,261
09/02/2015	3.428.078	200.006	1.821.276.892	344.558.942	3,883	1,017

Outro dado relevante é a quantidade média de megabits por segundo do protocolo DICOM. Os valores obtidos para DICOM comparados com a média de megabits por segundo

do total de dados chegou, em alguns casos, a 33,027% de participação no tráfego total. Devemos ressaltar que não há transmissões contínuas do protocolo DICOM durante todo o tempo das amostras.

Realizando a classificação por pacotes, foi dividida a quantidade de pacotes em grupos formados por faixas de tamanho por bytes conforme trabalho de (CHIMMANEE; PATPITUCK, 2013). Na Tabela A.1, no Apêndice, são encontrados esses dados divididos neste tipo de agrupamento. Pode-se também observar os resultados através do histogramas na Figura 5.4. Nota-se que em todas as amostras coletadas mais de 60% dos pacotes tem tamanho de até 250 bytes.

Em relação aos pacotes DICOM, mostrados nos histogramas da Figura 5.5, em todas as coletas realizadas a grande maioria dos pacotes está incluída em duas faixas. A primeira faixa está entre 40 e 79 bytes correspondendo a cerca de 60% dos pacotes. A segunda, corresponde à faixa 2.560 e 5.119 bytes representando mais de 25% dos pacotes. Assim sendo, acima de 80% dos pacotes estão nestas duas faixas de tamanho. A quantidade de dados detalhada por faixas pode ser visto na Tabela A.2 também no apêndice. É importante observar que o comportamento do tráfego DICOM, apresenta semelhança com o modelo de distribuição de pacotes de Castro e outros (2010). Os autores apresentaram histogramas semelhantes quando analisaram o tráfego oriundo de transmissões FTP e de aplicações industriais.

A classificação também pode ser utilizada para verificar a quantidade de pacotes ou de bytes que trafegam na rede por protocolo (THOMPSON; MILLER; WILDER, 1997). No contexto deste trabalho, uma das formas de classificação utilizada foi a quantidade de dados trafegando por protocolos da camada 3, da pilha de protocolo TCP/IP. A Tabela 5.6 exibe a quantidade de pacotes classificados pelos protocolos TCP, UDP e Outros. Pode-se verificar que em todas as amostras coletadas na rede, mais de 94% dos pacotes são TCP.

Analisando na perspectiva dos bytes, Tabela 5.7, a quantidade de bytes TCP chega a passar de 98% dos dados. Esses números demonstram que o tráfego predominante na rede hospitalar é TCP, sendo portanto dados em concorrência direta por largura de banda com o tráfego DICOM.

Tabela 5.6: Pacotes por Protocolo

Data da Coleta	Total	TCP	UDP	Outros
03/02/2015	8.247.765	8.056.396	44.949	146.420
	100%	97,68%	0,54%	1,78%
05/02/2015	3.566.425	3.462.576	54.492	49.357
	100%	97,08%	1,49%	1,43%
06/02/2015	9.679.893	9.510.283	44.682	124.928
	100%	98,25%	0,46%	1,29%
07/02/2015	5.875.712	5.671.910	64.818	138.984
	100%	96,53%	1,1%	2,37%
09/02/2015	3.428.078	3.246.082	42.137	139.859
	100%	94,69%	1,23%	4,08%

5.1.5 Identificação do Tráfego

Na parte de análise dos pacotes por fluxo de dados foi realizada caracterização para identificação dos tráfegos de acordo com o exposto anteriormente no Capítulo 3. Várias pesquisas neste assunto vêm propondo soluções no âmbito de outros tipos de fluxos de dados, como por exemplo HTTP (*Hypertext Transfer Protocol*). Todavia, não foram encontrados estudos sobre a identificação do tráfego por fluxos com dados de imagens médicas, especificamente. As investigações destes resultados foram divididas em duas partes, sendo a primeira uma análise por tempo de fluxo e a segunda por quantidade de pacotes e bytes. Como exposto anteriormente na Seção 3.1, a decomposição por tempo de fluxo busca identificar os fluxos libélulas, rápidos, ou tartarugas, lentos. Acerca dos fluxos por quantidade de pacotes e bytes, a identificação se dá por fluxos grandes, elefantes ou pequenos, ratos.

A primeira parte da análise sobre os dados coletados nas datas entre 03 e 09 de fevereiro de 2015 foi obtida utilizando uma técnica similar à usada por Brownlee e Claffy (2002) . Nesta técnica, são quantificados os dados por tempo de fluxo. Os resultados encontrados e descritos na Tabela 5.8 expõem que a grande maioria dos fluxos de dados da rede, entre 77% e 98%, possui o tempo de duração do fluxo abaixo de 2 segundos, indicando que estes fluxos na sua grande maioria são classificados como fluxos libélulas. Em relação aos fluxos

Tabela 5.7: Bytes por Protocolo

Data da Coleta	Total	TCP	UDP	Outros
03/02/2015	3.364.449.510	3.347.937.467	5.574.953	10.937.090
	100%	99,51%	0,17%	0,32%
05/02/2015	2.560.737.238	2.542.653.608	6.982.066	11.101.564
	100%	99,29%	0,27%	0,44%
06/02/2015	4.071.976.809	4.056.711.254	5.661.252	9.604.303
	100%	99,63%	0,14%	0,23%
07/02/2015	1.916.998.644	1.897.019.840	9.225.797	10.753.007
	100%	98,96%	0,48%	0,56%
09/02/2015	1.821.276.892	1.805.049.396	5.313.147	10.914.349
	100%	99,11%	0,29%	0,60%

tartarugas, menos de 0,2% de todo o tráfego se encaixariam nessa categoria com mais de 15 minutos.

No caso dos fluxos DICOM para as mesmas amostras coletadas em fevereiro, ver Tabela 5.9, os resultados apresentados indicam indícios ainda maiores de que os fluxos das imagens são do tipo libélula, visto que acima de 98% do tráfego analisado tiveram tempos de fluxos abaixo de 2 segundos. Além disso, nenhuma das amostras coletadas foram consideradas tartarugas.

Na segunda parte da análise por fluxo, está de acordo com Mori e outros (2004). Verificando os fluxos que continham a quantidade de pacotes maior que 0,1% do total de pacotes, encontrou-se entre 55 e 114 fluxos classificáveis como elefante. Pode-se verificar na Tabela 5.10 que esses fluxos correspondem em bytes a percentuais que variam entre 52,20% e 84,81% do volume de tráfego. Esses resultados indicam que pequenas quantidade de fluxos podem compor a grande maioria do tráfego da rede. Essa pequena quantidade de fluxo é classificada como fluxo elefantes.

Realizando essa mesma análise para o tráfego DICOM, os resultados foram ainda mais expressivos. Na Tabela 5.11 pode-se constatar que cerca de 30% dos fluxos apreciados se ajustam a identificação como tráfego elefante. Essa porcentagem de fluxo corresponde em bytes a mais de 99% do volume de tráfego DICOM na rede. Apenas um percentual ínfimo

Tabela 5.8: Quantidade de Fluxo Geral por Tempo

Data da Coleta	Total	Até 2 seg.	Entre 2 seg. a 15 min.	Mais de 15 min.
03/02/2015	235154	227913	7058	182
	100%	96,92%	3,00%	0,08%
05/02/2015	31826	24703	6904	219
	100%	77,61%	21,69%	0,70%
06/02/2015	282993	277367	5485	141
	100%	98,01%	1,93%	0,06%
07/02/2015	174639	168856	5609	174
	100%	96,68%	3,21%	0,11%
09/02/2015	31484	24927	6385	172
	100%	79,17%	20,28%	0,55%

do tráfego se enquadraria com a nomenclatura de ratos. Verificando esse pequeno volume de tráfego não elefante, concluí-se que se trata de pacotes para estabelecimento de conexão e manutenção desta conexão entre o servidor PACS e a modalidade de Ultrassom.

5.2 Modelagem de Tráfego DICOM

Analisando trabalhos de modelagem realizados anteriormente, observa-se que a criação de modelos de tráfego auxiliam projetistas e operadores de rede na concepção de novas infraestruturas. Neste trabalho foi realizada a modelagem da fonte de dados de um equipamento de ultrassom, baseando-se no tamanho dos arquivos representando um 1 ano de exames.

5.2.1 Coleta de Dados da Modalidade

As coletas de dados da modalidade ultrassom foram realizadas no período de novembro de 2013 a outubro de 2014 em um hospital de pequeno porte. Estas coletas ocorreram no período noturno e nos finais de semana, visto que o equipamento de ultrassom normalmente está em uso durante o dia. Além da impossibilidade de coleta dos dados durante o dia, interrupções para realização de exames de emergência também eram constantes. Eventualmente

Tabela 5.9: Quantidade de Fluxo DICOM por Tempo

Data da Coleta	Total	Até 2 seg.	Entre 2 seg. a 15 min.	Mais de 15 min.
03/02/2015	165	162	3	0
	100%	98,18%	1,82%	0,00%
05/02/2015	219	215	4	0
	100%	98,17%	1,83%	0,00%
06/02/2015	197	191	6	0
	100%	96,95%	3,05%	0,00%
07/02/2015	234	233	1	0
	100%	99,57%	0,43%	0,00%
09/02/2015	174	172	2	0
	100%	98,85%	1,15%	0,00%

Tabela 5.10: Identificação de Tráfego Geral (Fluxos Elefantes)

Data Coleta	Total Fluxos	Fluxos Elefantes	Total Bytes	Bytes Elefantes	% Bytes Elefantes
03/02/2015	235.155	57	3.364.449.510	2.384.452.564	70,87%
05/02/2015	31.827	114	2.560.737.238	2.171.801.542	84,81%
06/02/2015	282.994	55	4.071.976.809	2.895.862.571	71,12%
07/02/2015	176.640	95	1.916.998.644	1.000.591.579	52,20%
09/02/2015	31.485	104	1.821.276.892	1.493.147.616	81,98%

essas interrupções geraram retardo na coleta dos dados, em virtude da realização de exames ter maior prioridade.

As coletas foram feitas de forma manual, anotando as informações disponibilizadas diretamente na tela do equipamento. A escolha dos dados anotados foi baseada em duas premissas importantes: a manutenção do anonimato dos dados dos pacientes e as informações disponíveis de forma visual na tela do equipamento. No caso do anonimato, a legislação existente veta a exposição dos dados dos pacientes. Quanto às restrições de acesso á base de dados do equipamento, estas foram definidas por política de acesso do fabricante que impede o seu acesso direto. Para coleta de dados na base de dados interna do equipamento, é necessário um certificado digital que está disponível apenas para os técnicos de empresas associadas ao fabricante dos equipamentos.

Tabela 5.11: Identificação de Tráfego DICOM (Fluxos Elefantes)

Data Coleta	Total Fluxos	Fluxos Elefantes	Total Bytes	Bytes Elefantes	% Bytes Elefantes
03/02/2015	166	55	379.276.601	378.878.885	99,90%
05/02/2015	220	73	457.696.261	457.204.032	99,89%
06/02/2015	198	66	510.534.501	510.060.325	99,91%
07/02/2015	235	78	440.564.322	440.049.362	99,88%
09/02/2015	175	58	344.558.942	344.161.630	99,88%

As informações colhidas, referente a carga de entrada das transmissões em blocos por dia, possuem os seguintes dados: data de realização, horário de início, horário de fim dos exames, quantidade de exames por dia, imagens por dia, MB total dos exames por dia, média de MB por exame e média de MB por imagem.

5.2.2 Caracterização da Fonte

O armazenamento dos dados no ultrassom é feito com o uso de objetos DICOM. Cada objeto é conhecido como estudo e pode conter imagens, vídeos e informações inerentes aos pacientes. Esses objetos DICOM são transmitidos através da rede com o uso do protocolo DICOM. A transmissão destes objetos pode ser feita por estudo ou por grupo de estudo. No caso de agrupamento, estes podem ser realizados por dia, por tipo de exame ou por médico executante do exame. Neste trabalho, foram feitas transmissões agrupadas por dia, contendo vários estudos nestes agrupamentos. A Tabela 5.12 contém um compilação geral dos dados coletados no equipamento de ultrassom correspondente ao período informado na Seção 5.2.1. Podemos observar que na coluna "Média de Dia de Exames" temos uma média de 3,2 dias de exames durante 7 dias da semana. Este fato ocorre devido à dificuldade de manutenção de médicos plantonistas em toda e escala semanal e o não atendimento contínuo nos finais de semana. Diante deste fato, em uma escala de plantão completa, o volume de dados médio por dia poderia ser muito maior. Podemos ressaltar também que dentro de um mês a quantidade de dias de exames equivale a 43% dos dias do mês gerado em torno de 2 gigabytes de dados.

Tabela 5.12: Características da fonte de dados Ultrassom (US) por período

Período	Média de Dias de Exames	Média de Exames	Média de Imagens	MB Médio	MB/Exames (Média)	MB/Imagem (Média)
Dia	1	10,9894	153,4149	163,48857	15,0874879	1,72195755
Semana	3,20689655	35,4828	496,1034	526,14193	12,0566786	0,92731186
Mês	13,4285714	147,571	2060,143	2181,9894	14,6618642	1,05593326

5.2.3 Modelagem da Distribuição

O tipo de rede em estudo e as características de tráfego influenciam na escolha das distribuições probabilísticas a serem utilizadas para modelagem de tráfego. Modelos de tráfego que não podem capturar, ou descrever as características estatísticas do tráfego real da rede, devem ser evitados, uma vez que a escolha de tais modelos resultará em subestimação ou superestimação. Não há um modelo único que pode ser usado de forma eficaz para todos os tipos de tráfegos de rede, mas a escolha de distribuições com grande verossimilhança com os dados reais podem resultar em bons modelos de tráfego.

No processo de modelagem deste trabalho, a medida de interesse utilizada foi o tamanho dos arquivos de imagens médicas de um equipamento de ultrassom. Como forma de modelagem, a técnica de *fitting* foi utilizada para selecionar uma distribuição probabilística semelhante com os dados do tamanho dos arquivos.

O ato de modelar tem início com a escolha das distribuições que tenham características semelhantes com os dados estudados. Primeiramente foi analisado se os valores da distribuição coletada são compatíveis com distribuições probabilísticas discretas ou contínuas. Os valores do tamanho dos arquivos de imagens médicas podem assumir valores reais não enumeráveis, ou seja, não há uma precisão limitada do tamanho. Partindo desta informação, apenas distribuições contínuas foram analisadas. Entre as distribuições contínuas temos distribuições que podem assumir apenas valores positivos e outras que podem assumir ambos os valores, negativos e positivos. Como os valores dos tamanhos de arquivos nunca contém magnitude negativa, as distribuições que pertencem à classe com valores não positivos foram descartadas.

O resultado deste processo de seleção gerou 6 distribuições contínuas, não negativas. Entre as distribuições selecionadas, 5 delas são comumente referenciadas na literatura de modelagem de tráfego: Lognormal, Exponencial, Weibull, Gamma e Pareto.

Prosseguindo no processo de modelagem, foram ajustados os parâmetros de todas dis-

tribuições probabilísticas em relação à distribuição real através do método da máxima verosimilhança. Como resultado da parametrização, 3 distribuições probabilísticas foram selecionadas. Analisando o gráfico na Figura 5.6 e verificando o ajuste de curva entre as 3 distribuições probabilísticas selecionadas e a distribuição a ser modelada, pode-se verificar que a distribuição Dagum conseguiu um ajuste mais fino em relação ao histograma dos dados do tamanho dos arquivos. Os valores finais dos parâmetros ajustados podem ser vistos na Tabela 5.13.

Tabela 5.13: Parâmetros Utilizados nas Distribuições Probabilísticas

Distribuições	**Parâmetros**
Dagum	κ=0,4774 α=2,7321 β=120,56
Weibull	α=1,2233 β=109,91
Lognormal	α=1,004 μ=4,2564
Exponencial	λ=0,00802
Pareto	α=0,33388 β=3,53
Gamma	α=0,25318 β=492,43

Para finalizar a modelagem da distribuição, após o ajuste dos parâmetros, uma análise com 3 testes estatísticos foi realizada para validar qual distribuição possui a melhor aderência com os dados coletados. Observando a Tabela 5.14, pode-se verificar a ordem das distribuições probabilísticas, que nos 3 testes, obtiveram as maiores aproximações com os dados reais. Os resultados dos testes mostram que a distribuição Dagum, entre as distribuições testadas, obteve a melhor aderência aos dados reais. Na outra ponta da tabela, pode-se verificar que a distribuição Gamma obteve a menor aderência entre todas as distribuições analisadas.

A validação do modelo pode ser feita comparando os valores obtidos com os teste estatísticos em relação ao nível de significância definido. Quanto menor o valor obtido no teste, melhor é a aderência do modelo aos dados reais. Normalmente se usa um nível de significância de 0,05 para validação dos resultados científicos. A distribuição Dagum, no teste de Kolmogorov-Smirnov, obteve o valor de 0,07732 como valor de p-value. Este resultado

Tabela 5.14: Qualidade do Ajuste entre as Distribuições Real e Probabilística

Distribuições	Kolmogorov-Smirnov	Anderson-Darling	Chi-Squared
Dagum	0,07732	1,7169	21,223
Weibull	0,11291	3,9903	24,319
Lognormal	0,12533	4,4114	52,384
Exponencial	0,17494	6,9924	44,506
Pareto	0,34056	39631	311,8
Gamma	0,38099	44,772	277,12

apesar de estar fora do nível de significância padrão, ainda pode ser considerado como um valor aceitável por se tratar de dados com um alto nível de variação. Esse resultado significa que em 92,26% dos casos de uso do modelo, os resultados terão o mesmo comportamento da distribuição real.

Graficamente, pode ser visto na CDF (*Função de Distribuição Cumulativa*), Figura 5.7, que a distribuição selecionada como modelo para os dados reais se aproxima de forma estreita por toda a sua formação, havendo apenas um leve desvio em alguns pontos.

No caso do gráfico pdf (*Probability Density Function*), ver Figura 5.8, a distribuição Dagum foi ajustada para acompanhar o histograma dos dados reais. No gráfico nota-se que o modelo da distribuição Dagum sintetiza o formato dos dados reais.

A Figura 5.9 exibe um gráfico P-P Plot entre as duas distribuições. Observa-se uma boa aproximação dos dados da distribuição Dagum com dados reais.

5.3 Avaliação de Desempenho

Para demonstrar a utilidade do modelo de distribuição definido, experimentos de simulações foram realizados. Os ensaios efetuados simularam o aumento do número de modalidades de ultrassom transmitindo na rede. O modelo de distribuição definido corresponde ao tamanho de arquivos de apenas um modalidade.

Os software utilizados na simulação foram o Iperf e Mininet. Já para análise dos resultados os softwares TCPDump, TCPStat e Wireshark foram utilizados.

Estudo de desempenho foi realizada de acordo com o método apresentado na seção 4.4. O cenário de rede foi construído a partir de uma rede experimental com MininetHiFi. A infraestrutura de rede experimental é composta por:

- Uma máquina virtual para representar o PACS;
- Dois switches virtuais interconectando a sub-rede das modalidades, representando a zona de equipamentos médicos com a sub-rede do PACS, representando a zona dos servidores. A sub-rede das modalidades está limitada a 100 Mbps, enquanto que a sub-rede do PACS está limitada a 1 Gbps;
- Uma máquina virtual para geração de tráfego sintético de *background*, para representar o tráfego da zona de estações de trabalho;
- 10 máquinas virtuais para executar a geração de tráfego sintético DICOM de acordo com modelo de tráfego definido. Durante as medições, a ativação de uma maquina virtual representa a aquisição de uma nova modalidade.

O tráfego sintético foi gerado a partir da ferramenta Iperf com o uso de scripts para geração de valores aleatórios de acordo com modelo definido. Dois servidores Iperf foram configurados na máquina virtual PACS, para receber tráfego TCP em duas portas diferentes. As máquinas virtuais que representam ultrassom executam clientes TCP direcionando tráfego para o servidor PACS virtual. Já o tráfego de *background* representa o conjunto de tráfego de outros serviços também executando clientes TCP. Este ambiente foi montado baseado no ambiente real usado no processo de caracterização.

A coleta dos dados gerados pela simulação foi feita com o uso do coletor TCPDUMP, realizando as coletas no servidor PACS, que está conectado na sub-rede do PACS, na maquina virtual, que está gerando o tráfego de background, e nas sub-rede das modalidades.

No processo de envio dos dados dos clientes, que representam o tráfego DICOM, para o servidor PACS foi definido uma intervalo entre envios de 30 segundos entre as transmissões dos arquivos. Os valores dos tamanho de arquivos na transmissão foram gerados aleatoriamente de acordo com a distribuição de probabilidade Dagum. A quantidade de estudos DICOM representados em um arquivo DICOM seguiu a política de transferência definida pelo hospital, no qual os tamanhos correspondem a grupos de exames equivalentes a um dia.

Tabela 5.15: Parâmetros da Simulação

Parâmetros	Valores
Janela TCP	85,3 KBytes
Buffer	128 KBytes
MTU	40 Bytes
Tráfego na Dagum	TCP
Tráfego na background	TCP

O tráfego de *background* foi gerado de acordo com o modelo de Poisson. Para o envio dos dados, a distribuição Exponencial foi utilizada para definição do intervalo entre envio dos pacotes. Os valores dos intervalos entre chegadas foram gerados aleatoriamente, utilizando uma taxa média de 349,6503 pps como parâmetro.

O tráfego gerado por todos os clientes do sistema foi praticado de forma unidirecional, da fonte de tráfego em direção ao PACS.

A definição dos parâmetros do experimento de medição foi realizada com valores exibidos na Tabela 5.15.

O resultado da vazão é apresentado na Figura 5.10 e na Tabela 5.16. A Figura 5.10, contém 10 gráficos com os resultados de simulações do uso de 1 a 10 máquina virtual representando modalidade de ultrassom cliente, gerando tráfegos DICOM simultâneos. A presença de picos no gráfico 1 da Figura 5.10 indica que o tráfego TCP está ultrapassando a fase de início lento do TCP e consequentemente aumentando a janela de congestionamento. Este comportamento é similar ao apresentado na Figura 5.1, no processo de caracterização de tráfego. Pode-se identificar uma grande variação no fluxo de tráfego nos primeiros 5 gráficos que representam as transmissões de até 5 modalidade simultâneas. Já nos 5 gráficos seguintes pode-se observar uma certa estabilidade do tráfego, principalmente no gráfico com 9 e 10 modalidades simultâneas. Considerando-se a justiça do TCP, a vazão de 10 modalidades simultâneas tende a um valor em torno de 4,64 Mbps para cada modalidade. Essa redução na vazão com relação à vazão conseguida com apenas uma modalidade indica uma possível diminuição na QoS. Essa diminuição na vazão pode influenciar também na diminuição da QoS de serviços de consulta dos usuários no servidor PACS.

Em relação as médias de vazão apresentadas na Tabela 5.16 constata-se que o tráfego

DICOM variou entre valores próximo a 31 Mbps e 59 Mbps não prejudicando o tráfego de *background*, em relação a média de vazão, que variou de 9 Mbps a 22 Mbps. Percebe-se que mesmo com o aumento do número de modalidades transmitindo na rede, não houve uma redução expressiva do tráfego de *background* como observado para o caso de apenas uma modalidade transmitindo.

A medida que a carga aumenta, a vazão também aumenta, mas devido aos mecanismos de controle de congestionamento, quedas rápidas da vazão ocorrem constantemente.

Analisando o tráfego *background* os maiores picos encontrados aconteceram nos momentos de redução da vazão do tráfego das modalidades. A vazão do tráfego de *background* conseguiu atingir no seu maior pico valores próximo a 60 Mbps em alguns momentos. Esse fato ocorrido é esperado visto que os mecanismos de retransmissão e retransmissão rápida do TCP aumentam a vazão do tráfego *background* quando não compete com o tráfego de alguma modalidade.

Tabela 5.16: Vazão média na simulação com até 10 modalidades

Núm. Modalidades	Background (Mbps)	Dagum (Mbps)	Geral (Mbps)
1	13,137	31,802	41,891
2	21,259	51,732	69,793
3	12,458	55,431	67,210
4	9,740	59,070	67,404
5	13,285	42,936	55,617
6	15,789	44,728	59,844
7	15,225	48,618	63,075
8	14,039	51,496	64,752
9	22,046	53,461	72,203
10	17,879	46,411	63,360

Verificando a métrica de pacotes perdidos e atrasos no tráfego, os resultados da análise destas métricas podem ser vistos na Tabela 5.17. Pode-se verificar que o número de pacotes perdidos aumenta na mesma proporção que a quantidade de pacotes trafegando na rede aumentam. Esses resultados indicam que o aumento do número de modalidades na rede não altera de forma significativa o percentual do volume de perda de pacotes. Entretanto, a inser-

Tabela 5.17: Pacotes Perdidos e Atraso na simulação com até 10 modalidades

Núm. Modalidades	**Pkts Perdidos**	**% Pkts Perdidos**	**Total Pkts**	**Maior Atraso (seg)**	Média Atraso (Seg)
1	71.237	19,28%	369.547	32,095147	0,12667
2	146.966	21,23%	692.148	24,047065	0,01553
3	253.525	20,12%	1.259.959	48,127760	0,01330
4	311.046	20,97%	1.483.417	48,093160	0,03138
5	223.869	18,72%	1.196.139	48,128614	0,01489
6	242.747	19,13%	1.269.114	16,028247	0,00372
7	258.861	19,65%	1.317.177	32,062769	0,01238
8	271.825	19,94%	1.363.205	20,382996	0,00732
9	333.820	20,16%	1.655.632	4,011858	0,00197
10	237.540	18,61%	1.276.108	3,157095	0,00210

ção de modalidades, mesmo que seja apenas uma, amplia o percentual de perda de pacotes. Com a inserção do tráfego DICOM, o índice de perda variou entre 18,61% e 21,23%.

Já a métrica atraso, pode ser vista na coluna 5 e 6 da Tabela 5.17. Os picos de atrasos no tráfego variaram bastante. Os valores ficaram entre 3,157095 e 48,128614 segundos. Nos primeiros experimento com até 5 modalidades, os picos de atraso crescem com o acréscimo da quantidade de modalidades, mas a partir de 6 modalidades esses valores diminuem de forma significativa. O mesmo acontece com a média da variação de atraso. Isso ocorre devido a atuação do mecanismo de controle de congestionamento do protocolo TCP.

Segundo (SYSTEMS, 2008), os hospitais de grande porte costumam ter em seu parque uma média de até 8 modalidades de ultrassom. Diante do resultado expostos anteriormente, pode-se concluir que existem indícios que o uso de até 10 modalidades de ultrassom com esse limitante de 100 Mbps da sub-rede, pode influenciar o tráfego das demais aplicações quando em ambiente compartilhado, aumentando a perda de pacote e o atraso em alguns momentos, produzindo com isso a redução da qualidade dos serviços já existentes. Pode-se propor a partir dos resultados obtidos a inserção de mecanismos de garantia de qualidade dos serviços para manutenção da qualidade dos serviços já presentes na rede.

Estudos de QoS podem ser propostos e o modelo de fonte com distribuição de Dagum pode ser utilizado para representar fluxos elefante que ultrapassam a fase de início lento do TCP.

Devido à variação no comportamento no início e final das transmissões DICOM, traba-

lhos futuros podem considerar o uso de tráfego UDP para representar tráfegos sem controle de congestionamento.

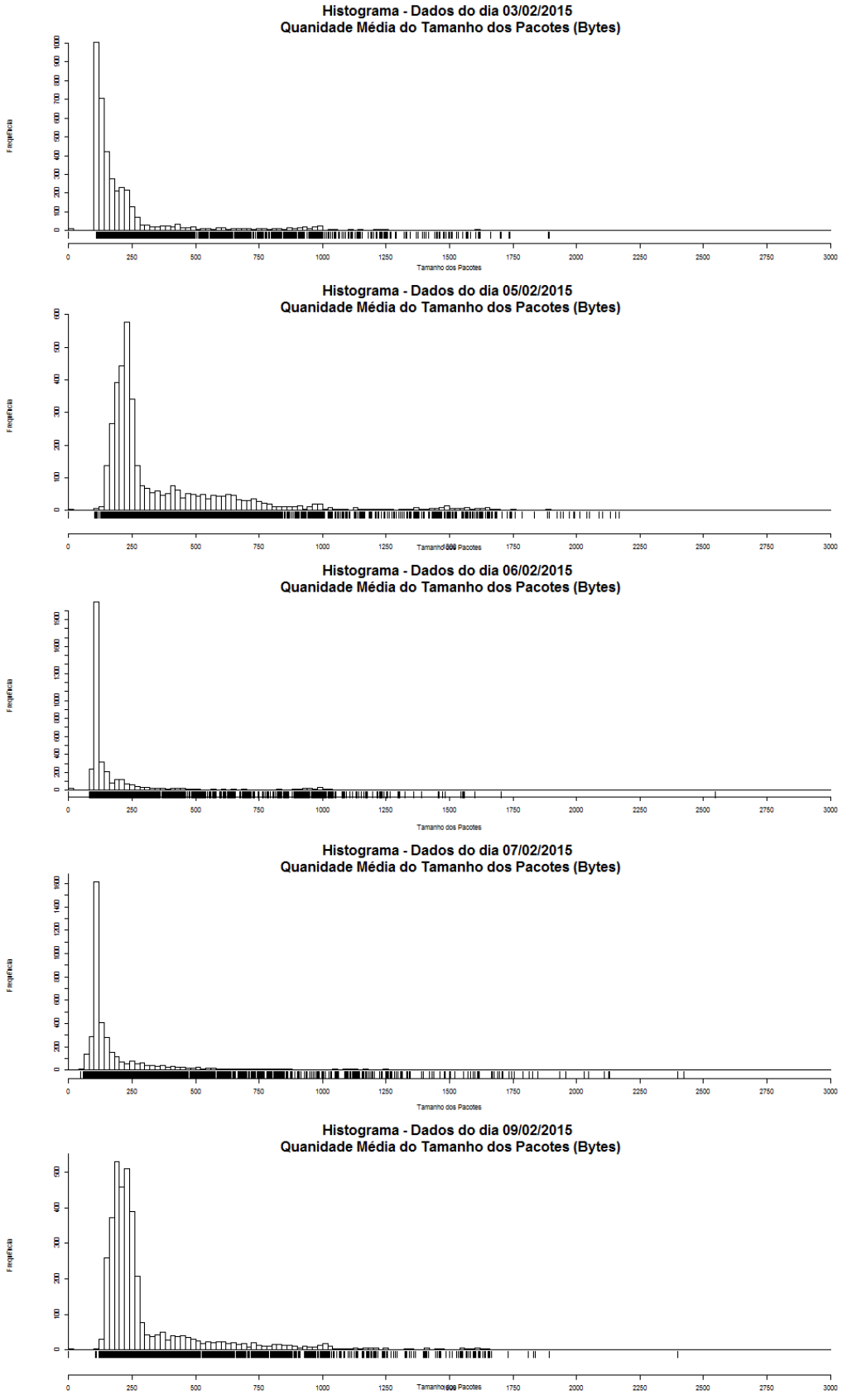

Figura 5.4: Histograma de distribuição da Quantidade de Pacotes por Tamanho

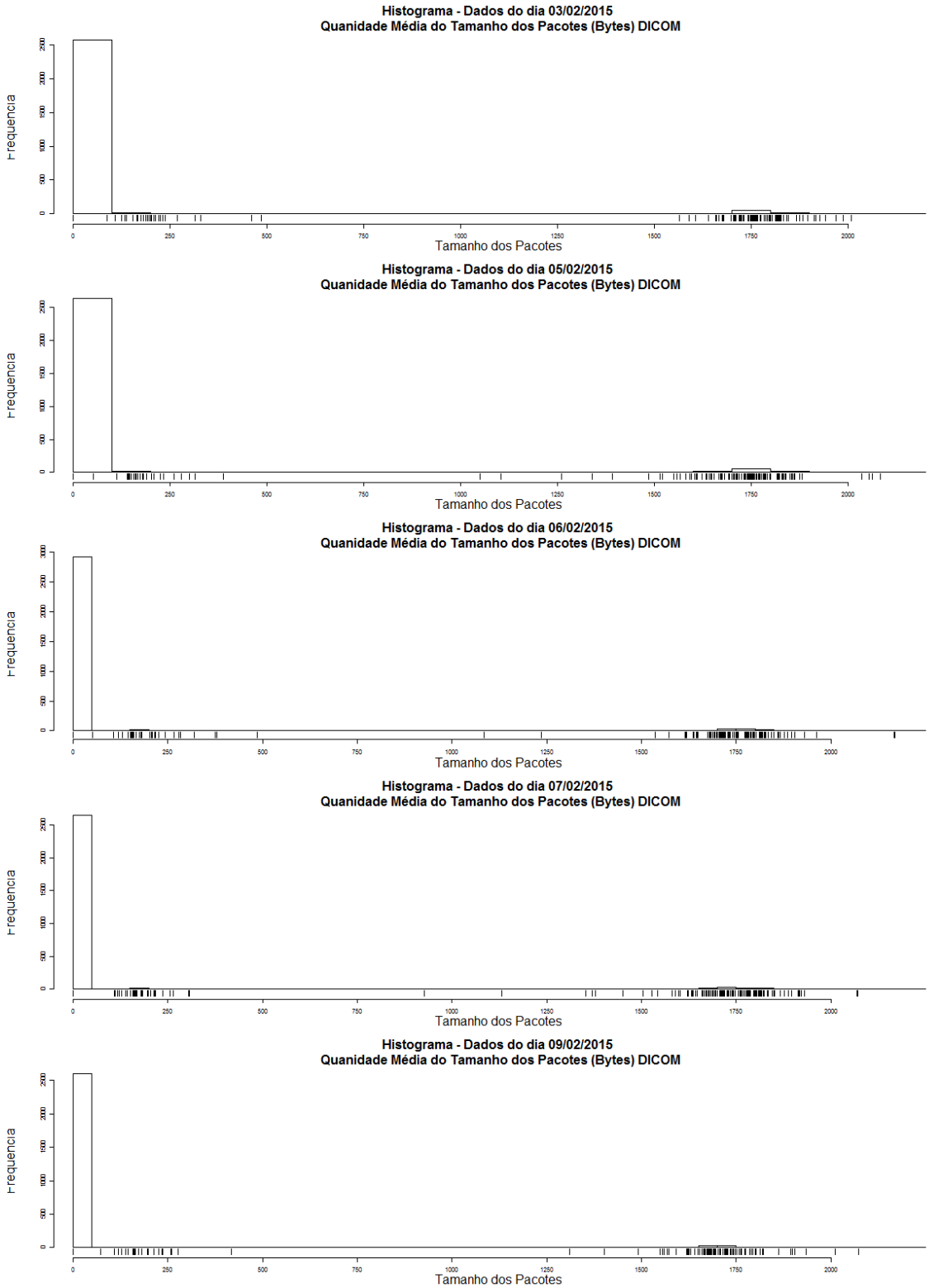

Figura 5.5: Histograma de distribuição da Quantidade de Pacotes DICOM por Tamanho

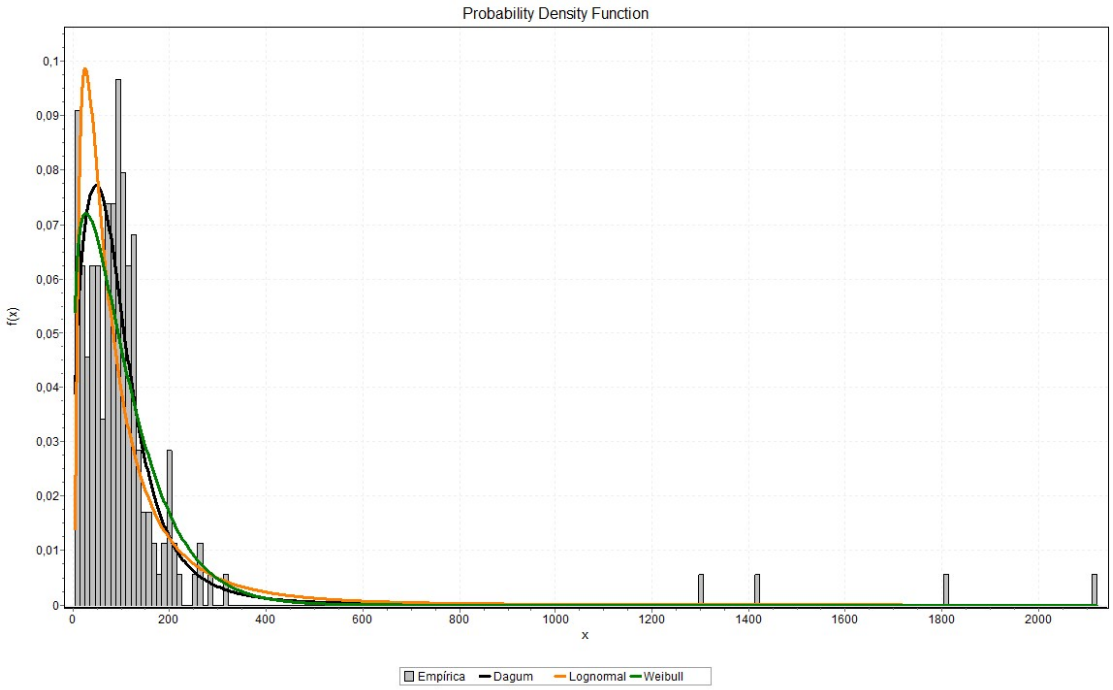

Figura 5.6: Comparação entre Distribuição Real e Probabilísticas após Ajuste dos Parâmetros

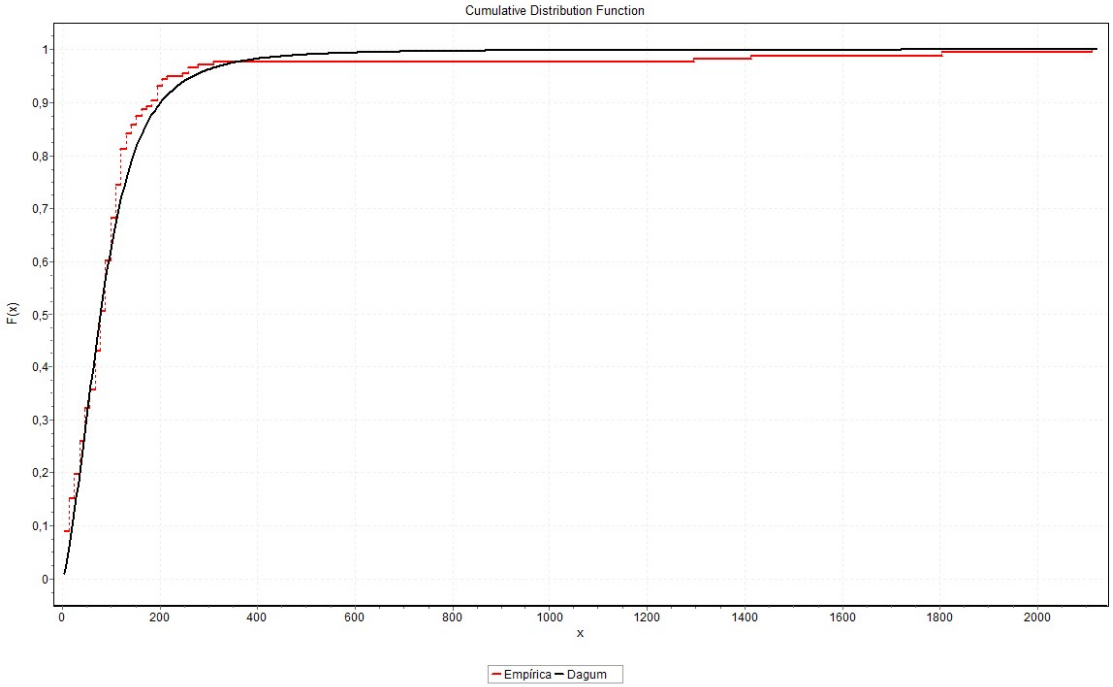

Figura 5.7: CDF de Comparação entre Distribuição Real e Probabilística

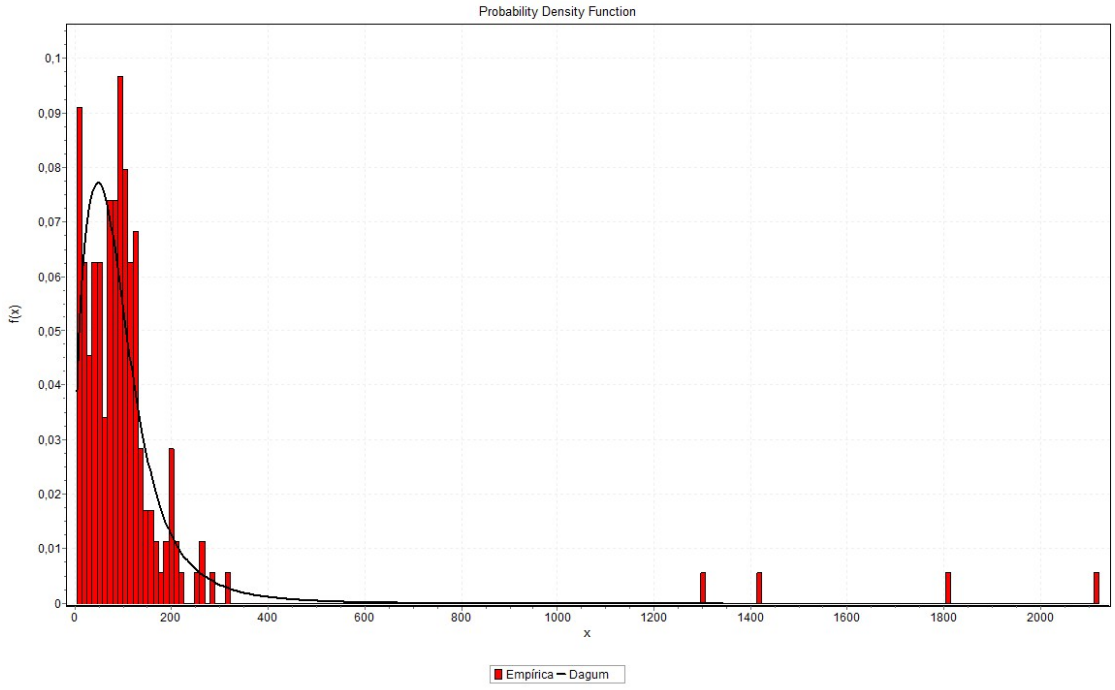

Figura 5.8: pdf de Comparação entre Distribuição Real e Probabilística

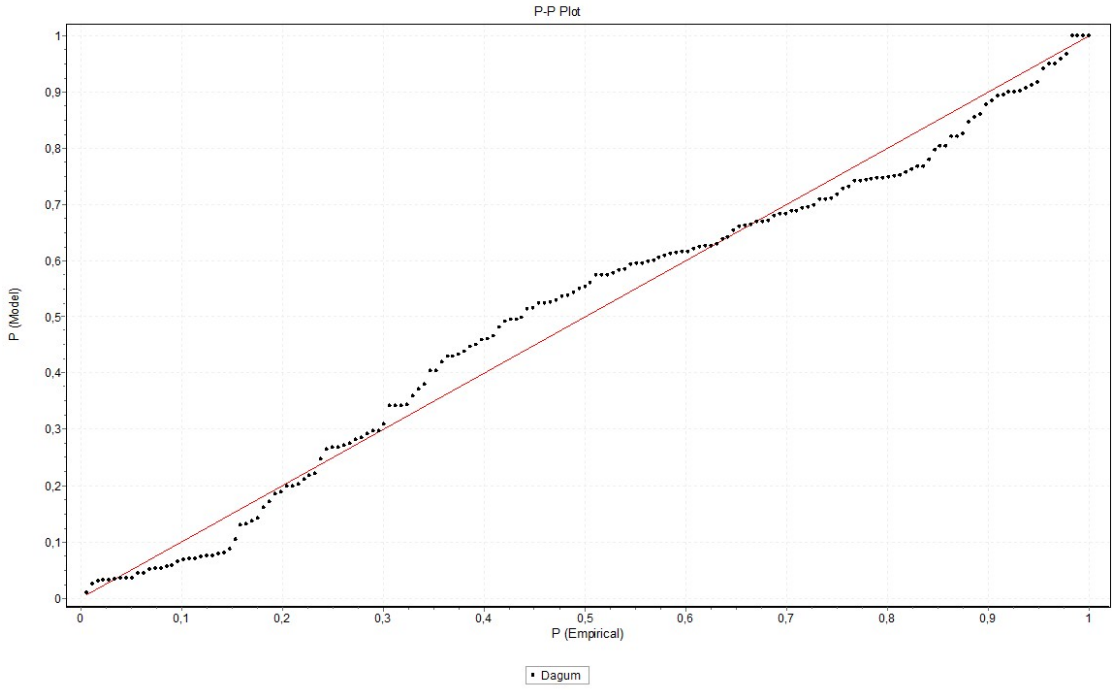

Figura 5.9: P-P Plot de Comparação entre Distribuição Real e Probabilística

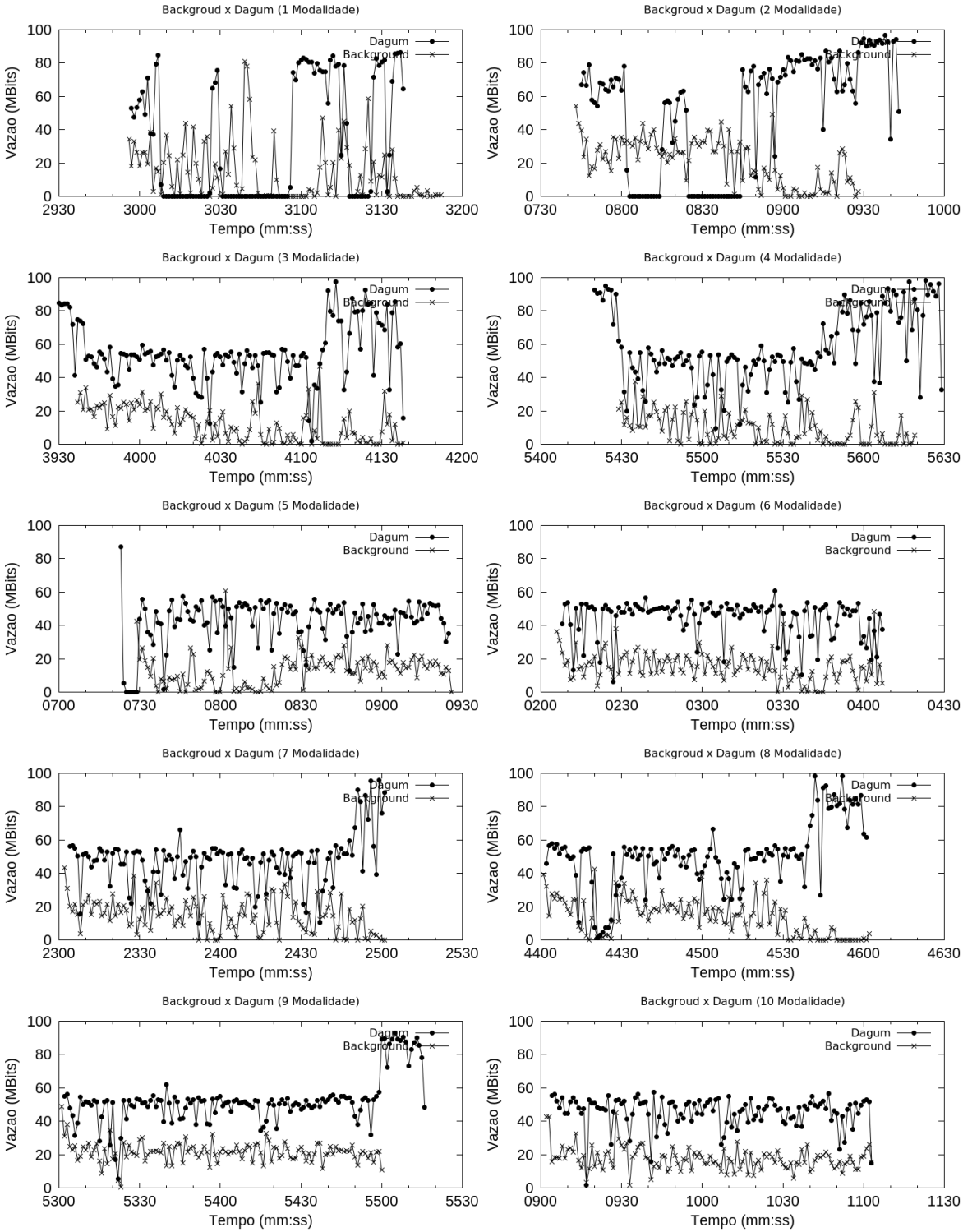

Figura 5.10: Simulação do Distribuição Dagum com Transmissão de até 10 Modalidade

Capítulo 6

Conclusão

Esta dissertação apresenta uma metodologia de caracterização e modelagem de tráfego para aplicações de imagens médicas radiológicas. A metodologia foi aplicada em uma rede hospitalar com tráfego DICOM gerado por modalidade radiológica de ultrassom. Dados obtidos a partir do monitoramento da rede foram utilizados para caracterização e classificação do tráfego da rede real.

No processo de caracterização de tráfego foi possível obter informações sobre as características do tráfego de imagens médicas, objetivando a realização do planejamento e expansão do parque de modalidades do hospital em estudo.

A caracterização foi realizada através da avaliação de medições realizadas em um hospital de pequeno porte. O produto desta caracterização demonstra que no ambiente de estudo o tráfego da rede existente é composto, em sua maioria, por tráfego do protocolo TCP, cuja participação corresponde a mais de 90% do volume de dados coletados. Outro ponto importante foi o índice de participação do protocolo DICOM, que obteve com apenas uma modalidade cerca de 20% do total de tráfego, atingindo esse valor com uma baixa quantidade de pacotes. Essa característica permitiu a identificação deste tipo de tráfego como elefante.

Para a definição do modelo, foi realizada a caracterização de uma fonte de tráfego, representada pela modalidade Ultrassom. No processo de modelagem foi realizada a comparação entre as distribuições probabilísticas e os valores reais do tamanho dos arquivos de imagens médicas. Com essa comparação constatou-se que a distribuição probabilística Dagum possui os valores que melhor aproximam aos valores do tamanho dos arquivos de imagens médicas. Esses resultados foram verificados, quanto à sua aderência, através testes estatísticos de

Kolmogorov-Smirnov, Anderson-Darling e Chi-Squared, gerando como consequência a não rejeição do modelo.

Após a modelagem, um processo de avaliação de desempenho foi realizado para avaliar a consequência da ampliação da carga gerada com a inclusão de novas modalidade radiológicas na rede. Os resultados apresentados demonstram indícios de que a ampliação do número de modalidades comprometem o desempenho da rede com o aumento das perdas o tráfego da rede no ambiente analisado.

Observa-se que os resultados encontrados neste trabalho podem ajudar projetos de expansão e de gestão das redes na área da saúde. É possível, com as informações aqui contidas, simular o tráfego de imagens médicas e propor formas de melhorias no desenvolvimento, configuração e ampliação tanto das redes existentes quanto das futuras.

6.1 Contribuições

As principais contribuições deste trabalho são o conhecimento sobre as características do tráfego de imagens médicas, a metodologia de modelagem de tráfego e a seleção de uma modelo de distribuição probabilística da fonte de dados do protocolo DICOM.

Os dados caracterizados contribuem com a exposição do volume e tipos de tráfego existentes na rede para auxiliar na gestão da rede. Já a metodologia cria um caminho para modelagem de tráfego de vários tipos de dados.

A modelagem de tráfego colabora com a comparação entre a distribuição do tamanho do arquivo e as distribuições Dagum, Lognormal, Exponencial, Weibull, Gamma e Pareto, propondo como modelo de representação da fonte de dados de imagens médicas a distribuição Dagum.

6.2 Trabalhos Futuros

Propomos as seguintes sugestões para trabalhos futuros:

- Caracterizar o tráfego de dados com uma maior quantidade e variedades de modalidades radiológicas;

- Modelar o tráfego de fonte de dados de outras modalidades;
- Avaliar cenários com tráfego UDP.

REFERÊNCIAS

ACR. American college of radiology. http://www.acr.org/. Acessado em 23/01/2015. 2015.

ALVAREZ, L.; VARGAS SOLIS, R. DICOM RIS/PACS telemedicine network implementation using free open source software. *Latin America Transactions, IEEE (Revista IEEE America Latina)*, v. 11, n. 1, p. 168–171, 2013. ISSN 1548-0992. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6502797>.

ANDERSON, T. W.; DARLING, D. A. A test of goodness of fit. *Journal of the American Statistical Association*, v. 49, n. 268, p. 765–769, 1954. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1954.10501232>.

ARNEY, D. et al. Simulation of medical device network performance and requirements for an integrated clinical environment. *Biomedical Instrumentation and Technology*, v. 46, n. 4, p. 308, 2012.

ASSIGNED NAMES, I. C. for; NUMBERS. Internet assigned numbers authority. http://www.iana.org/. Acessado em 23/10/2015. 2014.

ASSOCIATION, N. E. M. *Digital Imaging and Communications in Medicine (DICOM)*. [S.l.], 2011. Disponível em: <http://medical.nema.org/standard.html>.

ASSOCIATION, N. E. M. Digital imaging and communications in medicine. `http://dicom.nema.org/`. Acessado em 13/01/2015. 2015.

BIDGOOD, W. D. et al. Understanding and using DICOM, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, BMJ Publishing Group Ltd, v. 4, n. 3, p. 199–212, 1997. Disponível em: <http://jamia.bmj.com/content/4/3/199.full.pdf+html>.

BROWNLEE, N.; CLAFFY, K. Understanding internet traffic streams: dragonflies and tortoises. *Communications Magazine, IEEE*, v. 40, n. 10, p. 110–117, Oct 2002. ISSN 0163-6804.

CALLADO, A. et al. A survey on internet traffic identification. *Communications Surveys Tutorials, IEEE*, v. 11, n. 3, p. 37–52, rd 2009. ISSN 1553-877X.

CASTRO, E. et al. A packet distribution traffic model for computer networks. In: *The International Telecommunications Symposium, ITS*. [S.l.: s.n.], 2010.

CHALLITA, N.; ABDALLAH, R.; TAHER, N. C. Analysis and enhancement of wimax scheduling for telemedicine support. In: *Advances in Biomedical Engineering (ICABME), 2013 2nd International Conference on*. [S.l.: s.n.], 2013. p. 61–64.

CHIMMANEE, S. PACS metric based on regression for evaluating end-to-end qos capability over the internet for telemedicine. In: *Information Networking (ICOIN), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 359–364. ISSN 1976-7684.

CHIMMANEE, S.; PATPITUCK, P. Picture archiving and communication system (PACS) characteristic on wired-line and wireless network for traffic simulation. In: *Information Networking (ICOIN), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 589–594. ISSN 1976-7684.

DAINOTTI, A.; PESCAPÉ, A.; VENTRE, G. A packet-level characterization of network traffic. In: IEEE. *Computer-Aided Modeling, Analysis and Design of Communication Links and Networks, 2006 11th International Workshop on*. [S.l.: s.n.], 2006. p. 38–45.

DOWNEY, A. The structural cause of file size distributions. In: *Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2001. Proceedings. Ninth International Symposium on*. [S.l.: s.n.], 2001. p. 361–370. ISSN 1526-7639. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=948888>.

FINAMORE, A. et al. Experiences of internet traffic monitoring with tstat. *Network, IEEE*, v. 25, n. 3, p. 8–14, May 2011. ISSN 0890-8044.

FLOYD, D. et al. Study of radiologic technologists' perceptions of picture archiving and communication system (pacs) competence and educational issues in western australia. *Journal of Digital Imaging*, Springer US, p. 1–8, 2015. ISSN 0897-1889. Disponível em: <http://dx.doi.org/10.1007/s10278-014-9765-1>.

FOUNDATION, T. R. The r project for statistical computing. http://cran.r-project.org/mirrors.html. Acessado em 03/08/2014. 2014.

GULATI, N. Big data opportunities in the us medical imaging market. In: *Frost & Sullivan*. [S.l.: s.n.], 2015. Disponível em: <http://ww2.frost.com/>.

GUO, L.; MATTA, I. The war between mice and elephants. In: *Network Protocols, 2001. Ninth International Conference on*. [S.l.: s.n.], 2001. p. 180–188.

HANDELMAN, S. et al. *RTFM: New Attributes for Traffic Flow Measurement*. [S.l.], October 1999. Disponível em: <http://tools.ietf.org/pdf/rfc2724.pdf>.

HASAN, M. Intelligent healthcare computing and networking. In: *e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on*. [S.l.: s.n.], 2012. p. 481–485.

HEADQUARTERS, A. *WAN and Application Optimization Solution Guide Cisco Validated Design*. [S.l.]: Citeseer, 2008.

ISMAIL, M.; NING, Y.; PHILBIN, J. Transmission of DICOM studies using multi-series DICOM format. *Proc. SPIE*, v. 8674, p. 86740E–86740E–6, 2013. Disponível em: <http://dx.doi.org/10.1117/12.2007590>.

JAIN, R. *Art of Computer Systems Performance Analysis Techniques For Experimental DesignMeasurements Simulation And Modeling*. [S.l.]: 05/01/91, 1991.

KIM, T. et al. Medical image exchange and sharing between heterogeneous picture archiving and communication systems based upon international standard: pilot implementation. In: *15th International HL7 Interoperability Conference (IHIC 2015) I*. [S.l.: s.n.], 2015. p. 37. Disponível em: <http://ihic2015.hl7cr.eu/Proceedings-web.pdf#page=38>.

LAN, K.-c.; HEIDEMANN, J. A measurement study of correlations of internet flow characteristics. *Computer Networks*, Elsevier, v. 50, n. 1, p. 46–62, 2006.

LANTZ, B.; HELLER, B.; MCKEOWN, N. A network in a laptop: rapid prototyping for software-defined networks. In: ACM. *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. [S.l.: s.n.], 2010. p. 19.

LAW, A. M.; KELTON, W. D. *Simulation Modeling and Analysis*. 2nd ed. [S.l.]: McGraw-Hill Higher Education, 1997. ISBN 0070366985.

LEE, S.; LEVANTI, K.; KIM, H. S. Network monitoring: Present and future. *Computer Networks*, v. 65, n. 0, p. 84 – 98, 2014. ISSN 1389-1286. Disponível em: <http://www.sciencedirect.com/science/article/pii/S138912861400111X>.

MARQUES, P. M. d. A.; SALOMÃO, S. C. PACS: sistemas de arquivamento e distribuição de imagens. *Revista Brasileira de Física Médica*, v. 3, n. 1, p. 131–9, 2009. Disponível em: <http://www.abfm.org.br/rbfm/publicado/rbfm_v3n1_131-9.pdf>.

MARTINS, D. E. M. *Impacto da utilização de técnicas de amostragem na caracterização de fluxos de tráfego*. Dissertação (Mestrado) — Universidade do Minho, July 2013.

MASSEY JR, F. J. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, Taylor & Francis Group, v. 46, n. 253, p. 68–78, 1951.

MEGYESI, P.; MOLNÁR, S. Analysis of elephant users in broadband network traffic. In: BAUSCHERT, T. (Ed.). *Advances in Communication Networking*. [S.l.]: Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 8115). p. 37–45. ISBN 978-3-642-40551-8. Disponível em: <http://dx.doi.org/10.1007/978-3-642-40552-5_4>.

MORI, T. et al. Identifying elephant flows through periodically sampled packets. In: *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*. New York, NY, USA: ACM, 2004. (IMC '04), p. 115–120. ISBN 1-58113-821-0. Disponível em: <http://doi.acm.org/10.1145/1028788.1028803>.

NLANR/DAST. Iperf. https://iperf.fr/. Acessado em 03/03/2015. 2015.

NOGUEIRA, A. et al. Modeling network traffic with multifractal behavior. *Telecommunication Systems*, Kluwer Academic Publishers, v. 24, n. 2-4, p. 339–362, 2003. ISSN 1018-4864. Disponível em: <http://dx.doi.org/10.1023/A

PAPAGIANNAKI, K. et al. A pragmatic definition of elephants in internet backbone traffic. In: *Proceedings of the 2Nd ACM SIGCOMM Workshop on Internet Measurment*. New York, NY, USA: ACM, 2002. (IMW '02), p. 175–176. ISBN 1-58113-603-X. Disponível em: <http://doi.acm.org/10.1145/637201.637227>.

PATEL, G. Dicom medical image management the challenges and solutions: cloud as a service (caas). In: IEEE. *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on*. [S.l.: s.n.], 2012. p. 1–5.

PÉREZ, J. L. et al. Efficiency in the transmission of information through digital imaging and communications in medicine using security mechanisms: Tests with discus. *TELEMEDICINE and e-HEALTH*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 16, n. 5, p. 620–626, 2010. Disponível em: <http://online.liebertpub.com.ez20.periodicos.capes.gov.br/doi/pdf/10.1089/tmj.2009.0168>.

PLOUMIDIS, M.; PAPADOPOULI, M.; KARAGIANNIS, T. Multi-level application-based traffic characterization in a large-scale wireless network. In: *World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a*. [S.l.: s.n.], 2007. p. 1–9.

ROSTROM, T.; TENG, C.-C. Secure communications for PACS in a cloud environment. In: IEEE. *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. [S.l.: s.n.], 2011. p. 8219–8222. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6092027>.

RSNA. Radiology society of north america. http://www.rsna.org/. Acessado em 03/02/2015. 2015.

SALVADOR, P. et al. Statistical characterization of P2P-TV services and users. *Telecommunication Systems*, Springer, v. 55, n. 3, p. 363–376, 2014.

SANTOS, C. B. *Modelagem de TrÃ¡fego em Redes PLC (Powerline Communications) Utilizando Cadeias de Markov*. Dissertação (Mestrado) — Universidade Federal de GoiÃ¡s, 2009. Disponível em: <http://repositorio.bc.ufg.br/tede/handle/tde/995>.

SIBARANI, E. M. Simulating an integration systems: Hospital information system, radiology information system and picture archiving and communication system. In: IEEE. *Uncertainty Reasoning and Knowledge Engineering (URKE), 2012 2nd International Conference on*. [S.l.: s.n.], 2012. p. 62–66.

SINGH, E. V. J.; KUMAR, E. V.; BANSAL, K. L. Research on application of perceived qos guarantee through infrastructure specific traffic parameter optimization. *Computer Network and Information Security*, v. 3, p. 59–65, 2014.

SYSTEMS, C. DICOM traffic performance and WAAS application deployment guide. *Corporate Headquarters*, v. 1, p. 1–26, 2008.

TECHNOLOGIES, M. Easyfit - distribution fitting made easy. http://www.mathwave.com/easyfit-distribution-fitting.html. Acessado em 19/02/2015. 2015.

THOMPSON, K.; MILLER, G.; WILDER, R. Wide-area internet traffic patterns and characteristics. *Network, IEEE*, v. 11, n. 6, p. 10–23, Nov 1997. ISSN 0890-8044.

THYAGO ANTONELLO, R.; CUNHA, R. F. et al. *Análise e modelagem de tráfego do mundo virtual second life*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, March 2008.

WAMSER, F. et al. Traffic characterization of a residential wireless internet access. *Telecommunication Systems*, Springer US, v. 48, n. 1-2, p. 5–17, 2011. ISSN 1018-4864. Disponível em: <http://dx.doi.org/10.1007/s11235-010-9324-0>.

WILLINGER, W. The discovery of self-similar traffic. In: HARING, G.; LINDEMANN, C.; REISER, M. (Ed.). *Performance Evaluation: Origins and Directions*. [S.l.]: Springer Berlin Heidelberg, 2000, (Lecture Notes in Computer Science, v. 1769). p. 513–527. ISBN 978-3-540-67193-0. Disponível em: <http://dx.doi.org/10.1007/3-540-46506-5_24>.

ZSEBY, T.; MOLINA, M.; DUFFIELD, N. *Sampling and filtering techniques for IP packet selection*. [S.l.], March 2009. 1-46 p. Disponível em: <http://www.rfc-editor.org/rfc/rfc5475.txt>.

Apêndice A

Resultados Complementares

Amostras	40-79	80-159	160-319	320-639	640-1279	1280-2559	2560-5119	>5120
03/02/2015	4.984.947	691.560	786.704	492.001	112.911	1.000.453	139.576	39.613
	60,44%	8,38%	9,54%	5,97%	1,37%	12,13%	1,69%	0,48%
05/02/2015	1.599.805	682.726	350.842	65.336	100.392	660.477	153.687	53.160
	43,63%	18,62%	9,57%	1,78%	2,74%	18,01%	4,19%	1,45%
06/02/2015	5.660.200	844.282	866.704	592.369	45.862	1.521.594	110.658	38.224
	58,47%	8,72%	8,95%	6,12%	0,47%	15,72%	1,14%	0,39%
07/02/2015	3.313.415	1.053.000	786.218	382.726	56.293	103.293	135.658	45.109
	56,39%	17,92%	13,38%	6,51%	0,96%	1,76%	2,31%	0,77%
09/02/2015	1.418.391	818.558	436.314	81.310	104.299	442.597	90.428	36.181
	41,38%	23,88%	12,73%	2,37%	3,04%	12,91%	2,64%	1,06%

Tabela A.1: Distribuição Total da Quantidade de Pacotes por Tamanho

Amostras	40-79	80-159	160-319	320-639	640-1279	1280-2559	2560-5119	>5120
03/02/2015	126.166	0	1.668	218	1.283	11.497	57.393	15.829
	58,94%	0,00%	0,78%	0,10%	0,60%	5,37%	26,81%	7,39%
05/02/2015	153.156	221	1.939	291	1.539	15.433	68.408	19.941
	58,70%	0,08%	0,74%	0,11%	0,59%	5,91%	26,22%	7,64%
06/02/2015	169.960	0	2.014	255	1.734	16.568	75.394	21.851
	59,06%	0,00%	0,70%	0,09%	0,60%	5,76%	26,20%	7,59%
07/02/2015	147.599	1	2.020	282	1.516	14.330	65.192	19.559
	58,92%	0,00%	0,81%	0,11%	0,61%	5,72%	26,02%	7,81%
09/02/2015	114.462	0	1.603	218	1.252	12.983	55.483	14.005
	57,23	0,00%	0,80%	0,11%	0,63%	6,49%	27,74%	7,00%

Tabela A.2: Distribuição DICOM da Quantidade de Pacotes por Tamanho