UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Thiago Amaral Guarnieri

Uma proposta dirigida por dados para a melhoria do engajamento e da alocação de recursos em transmissões adaptativas ao vivo

Belo Horizonte
2023

Thiago Amaral Guarnieri

Uma proposta dirigida por dados para a melhoria do engajamento e da alocação de recursos em transmissões adaptativas ao vivo

Versão Final

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Ciência da Computação.

Orientadora: Jussara Marques de Almeida
Coorientador: Alex Borges Vieira

Belo Horizonte
2023

Guarnieri, Thiago Amaral.
G916p Uma proposta dirigida por dados para a melhoria do engajamento e da alocação de recursos em transmissões adaptativas ao vivo [recurso eletrônico] : / Thiago Amaral Guarnieri – 2023.
1 recurso online (167 f. il, color.) : pdf.

Orientadora: Jussara Marques de Almeida
Coorientador: Alex Borges Vieira

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f.154-165

1. Computação – Teses. 2. Redes sociais – Teses. 3. Vídeos na internet – Teses. I. Almeida, Jussara Marques de. II Vieira, Alex Borges. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. IV.Título.

CDU 519.6*22(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

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

FOLHA DE APROVAÇÃO

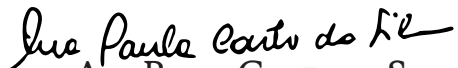
Uma proposta dirigida por dados para a melhoria do engajamento e da alocação de recursos em transmissões adaptativas ao vivo

THIAGO AMARAL GUARNIERI

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora
Departamento de Ciência da Computação - UFMG

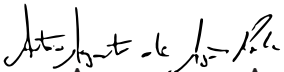
PROF. ALEX BORGES VIEIRA - Coorientador
Departamento de Ciência da Computação - UFJF



PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROF. DANIEL FERNANDES MACEDO
Departamento de Ciência da Computação - UFMG

PROF. DANIEL SADOC MENASCHE
Departamento de Ciência da Computação - UFRJ



PROF. ANTONIO AUGUSTO DE ARAGÃO ROCHA
Instituto de Computação - Universidade Federal Fluminense

Belo Horizonte, 28 de fevereiro de 2023.

Dedico este trabalho a Deus, minha esposa Tânia e minha família, e a todos que me ajudaram a chegar até aqui.

Agradecimentos

Em primeiro lugar, gostaria de agradecer a Deus por me dar força e me ajudar a superar os obstáculos que surgiram ao longo dos anos. Em seguida, agradeço também à minha orientadora, a Professora Jussara Almeida, e ao meu co-orientador, o Professor Alex Vieira, cujos conhecimentos foram cruciais para me auxiliar até o fim desta jornada. Desta relação de orientação fica a admiração e o respeito que vou carregar até o fim da vida. Agradeço também aos professores Ítalo Cunha e Idilio Drago pela colaboração e ajuda nos trabalhos publicados. Vale ressaltar também a importância da Universidade Federal de Minas Gerais (UFMG) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), cujo apoio financeiro e estrutura de pesquisa possibilitaram a dedicação necessária para produção dos resultados alcançados.

Gostaria também de agradecer a minha família, por cultivar em mim o desejo de aprender, a minha esposa Tânia, que foi o meu farol e tempos de tempestade e me estimulou a nunca desanimar. Por fim, agradeço também a todos os meus colegas do Laboratório de Computação Social (LOCUS) pela ajuda e compartilhamento de conhecimento.

"*Nosce te ipsum.*"
(Desconhecido)

Resumo

Transmissões de vídeo ao vivo na Internet têm alcançado grande popularidade. No entanto, à medida em que a audiência dessas transmissões aumenta, menos recursos se tornam disponíveis no provedor de conteúdo e na infraestrutura de rede para atender às demandas de desempenho de cada usuário. Ou seja, cresce o desafio de conciliar alto desempenho de transmissão geral e aumento de escala de distribuição.

Uma das formas de minimizar o impacto da escassez de recursos sobre o desempenho de transmissão é a utilização de mecanismos de adaptação, que ajustam a taxa de transmissão de acordo com a largura de banda do cliente. Estas adaptações, realizadas dinamicamente, minimizam a chance de congelamentos na reprodução, que são percebidos negativamente pelos usuários. Já o provedor de conteúdo, por outro lado, tem por objetivo admitir novos usuários, e recorre a planos de alocação de recursos, que reduzem, quando necessário, a taxa de transmissão de seus clientes correntes.

A redução da taxa de transmissão permite a entrada de novos clientes, mas pode gerar um impacto negativo sobre a qualidade de imagem (e.g., resolução), o que por sua vez pode levar ao abandono precoce dos usuários ou, em outras palavras, a redução de seus engajamentos. Ou seja, há um conflito de interesses entre o usuário, que deseja ter todos os recursos que requisita, e o provedor de vídeo, que precisa limitar esses recursos para aumentar a escalabilidade.

Com base na perspectiva apresentada, esta tese tem como objetivos gerais: (1) contribuir para a literatura existente sobre como se dá a relação entre desempenho de transmissão e engajamento para a produção de modelos preditivos e de comportamento de clientes que auxiliem nas tomadas de decisão de provedores de conteúdo e (2) explorar, a partir dos modelos desenvolvidos, alternativas de alocação de recursos que alcancem um melhor compromisso entre os interesses de clientes e provedor de vídeo, aumentando a economia de recursos para o provedor ao mesmo tempo em que minimiza a perda de engajamento de usuários no sistema. O caminho para atingir esse compromisso é a criação de restrições personalizadas que levem em consideração o desempenho de transmissão mínimo esperado por cada usuário do sistema de transmissão.

Os objetivos gerais acima são mapeados em quatro objetivos de pesquisa específicos e complementares, a saber: (1) caracterizar o desempenho de transmissão de clientes em vídeos ao vivo em larga escala e a correlação desse desempenho com o engajamento de usuários; (2) desenvolver um modelo de comportamento de clientes que leve em consideração o impacto do desempenho de transmissão no engajamento de seu usuário e nas

decisões de adaptação do cliente; (3) desenvolver modelos de descrição e previsão de engajamento para monitoramento ativo do engajamento em sistemas de transmissão de vídeo e (4) projetar um mecanismo para alocação de recursos no provedor, que analisa diversos cenários de alocação a fim de escolher o mais adequado para preservar o engajamento do usuário e reduzir o consumo de recursos no provedor de conteúdo.

As principais contribuições desta tese são: (1) uma caracterização do desempenho de transmissão em clientes de um evento com milhões de usuários simultâneos e uma avaliação do impacto desse desempenho no engajamento. Foi utilizado o conceito de cenários de desempenho, que considera a co-dependência entre métricas de desempenho e seu impacto no engajamento. Por exemplo, o aumento da taxa de transmissão é benéfico para o engajamento somente se as taxas de congelamento e adaptação são baixas. Também foi abordado o impacto de fatores contextuais nas métricas de desempenho e no engajamento. Foi constatado que o tipo de dispositivo, seu sistema operacional, seu provedor de conectividade e o período do jogo impactam tanto o engajamento quanto o desempenho de transmissão dos clientes. Além disso, o impacto da escala no desempenho de transmissão também foi investigado. Observou-se que, visando lidar com altas cargas de trabalho, a infraestrutura de transmissão limita a taxa de transmissão de seus clientes; (2) a criação e a validação de um modelo de comportamento de clientes sensível ao desempenho de transmissão experimentado, que demonstra que o engajamento do usuário (permanência, tempo entre sessões e número de retornos) e a atividade de adaptação do cliente são influenciados por métricas de desempenho de transmissão; (3) a criação de modelos de previsão e descrição de engajamento que avançam sobre o estado da arte em termos de precisão e acurácia. Esses modelos adotam uma nova abordagem, que descreve engajamento a partir da atividade de adaptação de um cliente, modelada por uma matriz de transição, e utiliza o conceito de cenários de desempenho para construir modelos especializados para diferentes níveis de desempenho experimentados pelos clientes. Nesse sentido, o modelo descritivo atingiu uma acurácia próxima de 90%, contra 65% de modelos que usam somente métricas clássicas de desempenho de transmissão, e o modelo preditivo obteve 80%, considerando o uso de cenários de desempenho; e (4) uma proposta de um mecanismo de alocação de recursos, que considera o impacto da adaptação no engajamento com o objetivo de adiar abandonos precoces e aliar redução do consumo de recursos com preservação de engajamento. Usando simulação por dados reais foi registrado um aumento de permanência e ganho médio de engajamento de 100%. Já no modo de economia de recursos, foi registrada uma economia de banda de centenas de gigabytes, com um impacto de menos de 0,4% no engajamento original.

Palavras-chave: Redes Multimídia. Internet. Transmissão de Vídeo. HTTP Live Streaming. QoS. Engajamento, QoE.

Abstract

Internet live streaming has reached large audiences. However, with the rise in popularity, fewer infrastructure resources become available to meet each user performance requirements. In other words, it becomes harder to conciliate high transmission performance and transmission scale growth.

One of the approaches to reacting to resource constraints and maintaining a minimum client transmission performance is the use of adaptation mechanisms, which adjusts the bitrate to the client device type and bandwidth. This dynamic bitrate adaptation reduces the probability of reproduction stalls, which have a negative perception by the users. On the other hand, the content provider needs to keep the system available for new users and uses for this goal resource allocation plans, which reduces, when necessary, the bitrate of its clients.

The bitrate reduction allows the entrance of new clients. However, it can produce a negative impact on the current users. As a result, they end by abandoning their sessions. In other words, the video bitrate reduction leads to a user engagement reduction. Therefore, there is a conflict of interest where the user always wants the maximum possible bitrate, and the content provider wants to maximize user engagement in both the number of users and client session duration, which may require client bitrate reduction.

Based on this perspective, this thesis has as main objectives: (1) to contribute to the current literature concerning the relationship between client transmission performance and engagement. This knowledge allows the creation of engagement and client behavior models that help content providers in infrastructure planning and performance monitoring, and (2) to explore, through these models, resource allocation alternatives to achieve a better tradeoff between user and content provider interests, that is, to increase resource saving in provider while it preserves engagement of the current users. The path to reaching this better tradeoff is the creation of personalized resource limitations that considers each client's transmission performance requirements.

This thesis map these expressed objectives in four research questions as follows: (1) to characterize client transmission performance in large-scale live streaming and the correlation of this performance with user engagement; (2) to develop a client behavior model that considers the impact of the client transmission performance on user engagement and the client adaptation regime; (3) to develop engagement descriptive and predictive models for active monitoring of the engagement in live video streaming a and (4) to project a mechanism for content provider resource allocation that evaluates various allocation sce-

narios to choose the most suitable for each client individually to preserve user engagement and reduce content provider resource consumption.

The main contributions of this thesis are: (1) a characterization of client transmission performance in a large-scale event with millions of simultaneous users and an evaluation of the impact of this performance on user engagement. We propose a concept of performance scenarios that show that the tolerance to a variation in a particular performance metric varies depending on the value of other performance metrics. For example, the rise in the client bitrate increases user engagement only if the stall and adaptation rates are low. We also addressed the impact of contextual factors on performance metrics and engagement. We found that the device type, platform, internet service provider, and transmission period influence client transmission performance and engagement. Besides this, we also investigated the impact of transmission scale on transmission performance. We verify that the transmission infrastructure applied bitrate limitations to deal with heavy workloads; (2) the creation and validation of a performance-aware client behavior model. This model revealed that client transmission performance impacts user engagement (permanence, time between sessions, and the number of sessions) and client adaptation regime; (3) the creation of descriptive and predictive engagement models that advance the state-of-the-art concerning precision and accuracy. These models introduce a new approach to describe client performance. Instead of using classical performance metrics like stall and adaptation rate, we used the client adaptation regime, stored in a transition matrix, associated with performance scenarios. Using this strategy, we constructed specialized models capable of reaching high accuracy in different client transmission performance levels. More specifically, the descriptive model has reached accuracy nearly 90% against 65% of the engagement model that uses classical performance metrics. The predictive model, in turn, obtained an 80% acuraccy in association with performance scenarios; and (4) The proposition of a resource allocation mechanism that considers the impact of the adaptation decisions on user engagement. We use the preservation of user engagement to guide allocation decisions which ensures, at the same time, user performance requirements and the reduction of client resource consumption. Using trace-based simulation, we verified an average gain of 100% in user engagement and a rise of hundreds of clients every minute. Considering the resource-saving mode, we registered a reduction of hundreds of gigabytes in bandwidth usage, with an impact of 0.4% in the original engagement.

Keywords: Multimedia Networks. Internet. Video Streaming. HTTP Live Streaming, QoS, User Engagement, QoE.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Plataformas de transmissão de vídeo ao vivo como o *Youtube* e *Twitch* têm alcançado grande popularidade na presente década. Estimativas dão conta de que em 2022, 82% de todo tráfego na Internet, e em 2023, 75% da rede móvel, serão usados para transmissões de vídeo [20, 30]. Essas plataformas têm em comum o uso de arquiteturas de transmissão adaptativas (*HTTP adaptive streaming* - HAS [96]). Em abordagens HAS, a mídia é disponibilizada em múltiplas taxas de transmissão, permitindo que um cliente possa trocá-la dinamicamente (i.e., adaptar), em resposta às variações na sua largura de banda corrente. Com isso, reduz-se a probabilidade de problemas de desempenho como longas latências de inicialização do vídeo e ocorrências de congelamentos (i.e., *stalls*).

Obter um alto desempenho de transmissão é importante porque ele estimula a permanência (i.e., engajamento) dos usuários [100, 124, 128]. Isto é, sabe-se que usuários em sessões com menor desempenho, onde existem, em particular, muitos congelamentos, podem abandonar mais cedo as suas sessões e demorar mais tempo para retornar [27, 72], o que diminui a viabilidade financeira de provedores de conteúdo.

Apesar das melhorias, o HAS, como implementado no estado-da-arte [94], não é capaz de sanar todos os problemas relativos ao desempenho de transmissão, que como mencionado, impactam diretamente o engajamento. Os algoritmos de adaptação existentes [10] possuem baixa eficiência em diversos contextos importantes, como transmissões ao vivo [9].

Um outro problema do HAS é que abordagens adaptativas não são cooperativas por padrão, porque requisitam uma taxa de transmissão sem levar em conta a perspectiva do provedor de conteúdo, que precisa garantir bom desempenho de transmissão para a maioria dos (idealmente todos) usuários, para que eles permaneçam por mais tempo e retornem mais vezes evitando, se possível, desperdício de recursos.

Diversos trabalhos têm apontado que a compreensão por parte do provedor da relação entre engajamento e métricas de desempenho de transmissão pode ajudar na criação de políticas de alocação que, ao mesmo tempo, preservem melhor o engajamento e diminuam o gasto de recursos [8]. No entanto, alcançar esse objetivo ainda é um desafio, uma vez que ainda não há um consenso sobre como as diversas métricas de desempenho de transmissão afetam o engajamento [9]. Evidências apontam que essa

influência pode abranger métricas de aplicação ou sistema (e.g., número de interrupções, taxa de transmissão), contextuais (e.g., dispositivo, conexão, hora do dia) e sociais (e.g., interesse no conteúdo) [103].

Nesta tese, será dado um enfoque mais evidente à classe de métricas de aplicação, que são amplamente utilizadas para caracterizar o desempenho de transmissão a partir da perspectiva do cliente. Elas são conhecidas também por terem um impacto, ainda que indireto, na qualidade percebida pelo usuário e, por consequência, em seu engajamento [27, 72, 3]. Exemplos dessas métricas são a *latência de inicialização*, *taxa de transmissão média*, *taxa de congelamentos* e *taxa de adaptações*. A título de simplificação, todas essas métricas serão classificadas como *métricas de desempenho de transmissão*. Além disso, ao longo do texto, serão propostas novas métricas, com o intuito de enriquecer o conhecimento da relação entre desempenho e engajamento. Exemplos de tais métricas são a *matriz de transição de adaptação* do cliente e sua *fração de segmentos em cada taxa de transmissão*. Os detalhes de cada métrica, assim como a heurística de cálculo usada em cada uma, serão introduzidos no Capítulo 4.

Com base no panorama apresentado, esta tese procura expandir o conhecimento sobre a relação entre métricas de desempenho de transmissão, como a taxa de congelamentos e taxa de transmissão, e o engajamento, a fim de propor abordagens práticas para obter um melhor compromisso entre o atendimento dos requisitos de desempenho dos usuários e o aumento do engajamento global e economia de recursos.

1.1 Motivação

O crescente uso da arquitetura *HTTP adaptive streaming* [96] aumentou o desempenho dos serviços de disseminação de vídeo na Internet [103]. No entanto, essa arquitetura não foi capaz de solucionar todos os episódios de baixo desempenho de transmissão, sobretudo em eventos ao vivo em larga escala, que registram uma taxa de chegada mais intensa e concentrada no tempo. A ocorrência de baixo desempenho faz com que os usuários se engajem menos e abandonem mais cedo suas sessões [103].

Para corroborar essa constatação, foi feita uma análise em partidas da Copa do Mundo FIFA de 2014 e 2018, buscando quantificar a quantidade de abandonos precoces (i.e., sessões com menos de 1 minuto) durante o pico de acessos. Foi constatado que a taxa de abandonos precoces alcançou um valor máximo de aproximadamente 64% nos dois eventos com uma média de 23% e 27% para 2014 e 2018, respectivamente. Ou seja, há uma parcela considerável de clientes novos que abandona suas sessões rapidamente, durante o período no qual o nível de interesse no conteúdo é muito grande.

Para entender as razões para a alta taxa de abandonos no pico de carga das transmissões, foi extraída a taxa de transmissão dos clientes fixos (i.e., computadores *desktop* e *smart-tv's*) que fazem parte do grupo de abandono precoce registrado. Esse grupo em particular utiliza telas de alta resolução, e por isso a demanda de recursos de seus usuários é maior. Foi observado que suas taxas de transmissão médias foram de 500 e 650 kbps nos eventos de 2014 e 2018, respectivamente. Ou seja, fica claro que uma das possíveis razões para o abandono se deve a um baixo desempenho, representado neste caso pela baixa taxa de transmissão. De acordo com [81, 70], uma taxa de transmissão de 500 kbps permite uma resolução de 240 linhas, insuficiente para a demanda de clientes em dispositivos de alta definição, que podem chegar hoje a resoluções de até 2160 linhas.

Vale ressaltar ainda que, além desses dois casos, outros eventos recentes também experimentaram episódios de baixo desempenho de transmissão, como por exemplo o seriado *Game of Thrones* [118] e o lançamento do serviço *Disney plus* [22].

Como pôde ser constatado pelo panorama exposto, existem situações em que a demanda por recursos já pode superar sua disponibilidade no provedor. Nesse tipo de situação, é comum que os provedores diminuam a taxa de transmissão do vídeo a fim de garantir um patamar mínimo de qualidade de serviço (QoS). O problema dessa abordagem é que a QoS muitas vezes está dissociada dos requisitos de desempenho de transmissão demandados pelos usuários [8, 10]. Por exemplo, como já mencionado, um vídeo em uma resolução de 240 linhas não é adequado para televisões de alta definição, mesmo que a transmissão tenha baixa latência e *jitter*. Esse descompasso entre QoS e os requisitos de desempenho dos usuários pode provocar uma redução de seus engajamentos.

A desconsideração de requisitos de desempenho pode afetar negativamente diversas etapas de uma transmissão adaptativa. Um exemplo é o planejamento da infraestrutura. O processo de transmissão de um grande evento precisa desta etapa para avaliar a demanda por recursos (e.g., largura de banda necessária, quantidade de servidores, entre outros). Uma das maneiras de fazê-lo consiste em gerar cargas sintéticas que, baseadas em modelos de comportamento realistas, extrapolam cargas previamente observadas para cenários de sobrecarga esperados.

No entanto, apesar das transmissões adaptativas serem usadas comercialmente há anos, ainda não existe um modelo que considere plenamente o impacto da variação de desempenho (e.g., taxa de transmissão) na carga imposta por seus clientes ao sistema. Além disso, a literatura mostra que a dinâmica de mudanças de taxa de transmissão têm impacto direto sobre o comportamento dos usuários [57, 73], fato esse que ainda não foi investigado amplamente. Com isso, cargas de trabalho sintéticas podem ser pouco realistas e levar a decisões de planejamento equivocadas.

Outro estágio importante é o de monitoramento e a alocação de recursos. Um dos objetivos mais importantes em um serviço de vídeo é garantir que seus clientes estejam em sessões com bom desempenho, livres de problemas como congelamentos e baixa taxa de

transmissão, já que isso impacta diretamente no engajamento dos usuários. No entanto, à medida que o número de clientes aumenta, fica mais difícil conciliar essa meta com o objetivo do provedor, que é atender ao maior número possível de clientes simultâneos.

Um dos caminhos para aumentar o número de clientes é a redução da taxa de transmissão dos clientes correntes. Nesse sentido, foi observado que mecanismos de alocação atuais não levam em consideração que uma redução de taxa impacta os usuários de formas diferentes, a depender de seu contexto [103]. Ou seja, é necessário levar em conta a relação entre desempenho de transmissão e engajamento, atentando-se também para fatores como tipo de dispositivo e conteúdo, para chegar a um melhor equilíbrio entre o interesse do cliente, que é obter o maior desempenho, e o interesse do provedor, que é manter seus clientes engajados por mais tempo e aumentar sua escala de distribuição.

Com base no exposto, tem-se que o fato motivador desta tese é a *constatação de que os mecanismos de planejamento e alocação de recursos têm apresentado uma eficácia parcial na tarefa de atingir um melhor compromisso entre o interesse do usuário, que é ter o melhor desempenho de transmissão possível, o que implica em consumo de recursos tanto quanto possível e requisitado, e o interesse do provedor, que é aumentar a escala da transmissão e o engajamento geral de todos os usuários [29]. Nesse sentido, uma das razões que dificulta a obtenção desse compromisso é que esses mecanismos não consideram de forma apropriada o impacto do desempenho de transmissão no engajamento dos clientes durante medidas de limitação de recursos.*

Assim, a partir dessa motivação, esta tese procura investigar *como o entendimento da relação entre as métricas de desempenho de transmissão e engajamento dos usuários pode auxiliar no aprimoramento dos mecanismos de planejamento de infraestrutura, para que seja possível atingir um melhor compromisso entre o interesse do usuário, que é obter alto desempenho de transmissão, e o interesse do provedor, que é aumentar o engajamento global (i.e., maior número de clientes, permanecendo por mais tempo)*

1.2 Hipótese e Questões de Pesquisa

O problema abordado nesta tese pode ser formalizado da seguinte maneira: seja L a banda total disponível no provedor P para uma dada transmissão e seja C um conjunto de clientes acessando esta transmissão.

Para que seja possível inferir o consumo de banda total dos clientes em C durante a transmissão, direcionando assim decisões de planejamento e gerenciamento de capacidade da plataforma provedora, é importante entender a fundo como clientes se comportam durante a transmissão, principalmente no que tange a questões relacionadas à adaptação

dinâmica da taxa de transmissão (dadas as opções em um conjunto de taxas B).

Em especial, é importante modelar o impacto que o desempenho de transmissão, medido por métricas de desempenho, tem na permanência do cliente na transmissão. Assim, a seguinte questão de pesquisa surge:

1) Como modelar o comportamento de clientes em transmissões ao vivo adaptativas levando em consideração o potencial impacto do desempenho de transmissão neste comportamento e engajamento do usuário?

Um modelo deste tipo permitiria a tomada de decisão sobre o valor mínimo de banda L necessário inicialmente para atender os requisitos de desempenho dos usuários com clientes em C (planejamento de capacidade) e preservar seu engajamento. Entretanto, dado o dinamismo natural do sistema durante a transmissão, em cada intervalo de tempo t, o provedor P pode estar em um dos seguintes estados:

a) **Estável:** o consumo total de banda pelos clientes em C está em um nível inferior ou igual a L. Nesse caso nada precisa ser feito, e cada $c \in C$ tem acesso a todas as taxas $b \in B$, podendo escolher qualquer uma delas. Uma limitação de taxas neste caso serviria para, por exemplo, economia de recursos;

b) **Sobrecarga:** a demanda de recursos está em um nível superior a L. Nesse caso o provedor deve obrigatoriamente limitar as taxas de transmissão para um ou mais clientes em C, caso queira admitir mais clientes. Mais especificamente, isso é feito escolhendo o melhor subconjunto de taxas $b \in B$ específico para cada cliente.

A premissa é que a escolha do subconjunto de taxas de transmissão para cada cliente, seja em estabilidade ou sobrecarga, deve ser feita de tal forma que a interferência no engajamento dos usuários seja minimizada. Em qualquer um dos estados citados, o uso de aprendizado de máquina para *previsão* do engajamento futuro de cada usuário com clientes em $c \in C$ é insumo chave para esta tomada de decisão.

O engajamento é função direta do desempenho de transmissão, estimado a partir de métricas de desempenho de transmissão [27, 72, 103]. Assim, durante o processo de realocação, as seguintes perguntas, relacionadas à restrição de recursos, precisam ser respondidas:

2) Como prever o engajamento, isto é, o tempo restante que cada usuário com cliente em $c \in C$ está disposto a permanecer no sistema (i.e., seu engajamento potencial) tomando como base o monitoramento contínuo de diferentes métricas de desempenho associadas na transmissão do cliente?

3) Como utilizar as previsões de engajamento para selecionar um cenário de realocação que mantenha ou mesmo aumente o engajamento dos usuários, ao mesmo tempo em que reduz o gasto de recursos, permitindo assim o ingresso de novos clientes?

Vale notar que as perguntas 2 e 3 devem ser respondidas periodicamente, isto é, em cada intervalo de tempo t, o sistema deve avaliar o engajamento potencial de todos os usuários correntes e redistribuir os recursos disponíveis a fim de melhorar o engajamento global e/ou reduzir o consumo de recursos.

1.3 Objetivos

Conforme discutido na seção anterior, o problema alvo a ser abordado nesta tese consiste em explorar mecanismo de aprendizado de máquina para a previsão de engajamento futuro de usuários, tomando como base o monitoramento de métricas de desempenho de transmissão, para realizar o gerenciamento dos recursos (notadamente largura de banda) do servidor, de forma a garantir um melhor compromisso entre interesses de usuários e do provedor.

A investigação deste problema foi dividida em 4 objetivos de pesquisa (OPs), projetados a fim de buscar respostas para as três perguntas propostas na seção anterior. São eles:

- **OP1 - Caracterização de desempenho de transmissão e sua relação com engajamento:** a capacidade por parte do provedor de avaliar a satisfação de seus usuários em relação ao serviço prestado é de vital importância para atrair mais receita e se manter viável. Isso pode ser feito assumindo o engajamento como um indicativo de satisfação do usuário com o serviço e avaliando o impacto do desempenho de transmissão neste engajamento. Por exemplo, já se sabe que congelamentos têm impacto negativo e devem ser evitados [27, 103, 23]. Apesar disso, existem outras métricas de desempenho cuja contribuição ainda está sendo avaliada. Por exemplo, não há uma compreensão sólida sobre a influência da dinâmica de adaptação e de fatores contextuais no engajamento. Também existem lacunas na compreensão da co-dependência entre métricas de desempenho de transmissão. Assim, esse objetivo de pesquisa visa aprofundar o conhecimento da correlação entre desempenho e engajamento e servir de base para os estudos que serão desenvolvidos nos objetivos de pesquisa posteriores.
- **OP2 - Modelagem do comportamento de clientes em transmissões adaptativas ao vivo:** modelos de comportamento ajudam a dimensionar os recursos necessários para transmissões de vídeo pela Internet. Contudo, a análise das abordagens atuais mostrou que elas dão suporte limitado às transmissões adaptativas, e que também não são sensíveis ao impacto do desempenho de transmissão sobre o engajamento dos usuários e o funcionamento do algoritmo de adaptação. Essas limitações podem levar a estima-

tivas de demanda por recursos imprecisas e aumentar a probabilidade de sobrecarga. Este OP tem como meta portanto projetar um modelo de comportamento com suporte arquiteturas de transmissão adaptativas e que considera o impacto do desempenho de transmissão no engajamento dos usuários. Além disto, pretende-se também implementar o modelo projetado em um gerador de cargas sintéticas que permita a criação de cenários de avaliação mais realistas. Desse modo, em termos gerais, este objetivo visa prover resposta para a primeira pergunta introduzida na definição do problema, que consiste em investigar como modelar o comportamento de clientes de transmissões ao vivo adaptativas, levando em consideração o potencial impacto do desempenho neste comportamento e no engajamento do usuário.

- **OP3 - Modelagem de engajamento para vídeos adaptativos ao vivo:** existem vários trabalhos dedicados a modelar a satisfação (e.g., engajamento) de usuários a partir de métricas de desempenho de transmissão, contextuais e de QoS [8, 116, 114]. Modelos desse tipo são úteis em particular para tarefas preditivas, como previsão de abandonos, e tarefas descritivas, como a descoberta dos fatores de desempenho que mais se relacionam à satisfação do usuário. No entanto, apesar de todo esforço empreendido, ainda não existe um modelo holístico, capaz de explorar múltiplas métricas de desempenho a fim de prever/descrever principalmente engajamento com precisão [9]. Nesse objetivo de pesquisa, o foco é utilizar a compreensão obtida nos objetivos de pesquisa 1 e 2 para responder à segunda questão proposta, a fim de investigar se é possível modelar com acurácia o engajamento por meio de métricas de desempenho de transmissão e métricas relativas à adaptação.
- **OP4 - Melhoria da alocação de recursos em transmissões adaptativas ao vivo:** Nesta etapa, serão estudadas novas estratégias de alocação de recursos que atinjam um melhor compromisso entre aumento de escalabilidade e economia de recursos (interesse do provedor) e garantia de desempenho individual (interesse dos usuários). Isso é motivado pelo fato de que os mecanismos adaptativos tradicionais se concentram apenas no atendimento dos interesses dos usuários. Para alcançar esse compromisso, é empregado um monitoramento contínuo do desempenho de transmissão das sessões em andamento dos clientes. A partir desse monitoramento, algoritmos de aprendizado de máquina são utilizados para prever o engajamento futuro dos usuários e estabelecer estratégias de alocação de recursos que permitam aumentar o número de clientes simultâneos e minimizar a ocorrência de potenciais abandonos de sessões por parte dos usuários.

1.4 Publicações

Os achados e contribuições, referentes aos objetivos de pesquisa propostos nesta tese, foram também documentados em artigos que foram publicados em diversas conferências e periódicos, conforme a listagem a seguir:

1. Guarnieri, T., Ítalo Cunha, Almeida, J., Drago, I., and Vieira, A. B. (2017). Characterizing QoE in large-scale live streaming. In Proc. of the IEEE GLOBECOM. (**OP1**)
2. Guarnieri, T., Drago, I., Ítalo Cunha, Almeida, B., Almeida, J., and Vieira, A. B. (2021). Modeling Large-Scale Live Video Streaming Client Behavior. Multimedia Systems. (**OP2**)
3. Guarnieri, T., Vieira, A., Cunha, I., and Almeida, J. (2018). Previsão de engajamento de usuários durante transmissão adaptativa de vídeo ao vivo. In Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos.(**OP3**)
4. Guarnieri, T., Vieira, A. B., and Almeida, J. (2019). Um modelo sensível a adaptação para previsão de qualidade de experiência em vídeos na Internet. In Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos.(**OP3**)
5. Guarnieri, T., Almeida, J., and Vieira, A. (2019). An adaptation aware model to predict engagement on http adaptive live streaming. In 2019 IEEE Symposium on Computers and Communications (ISCC). (**OP3**)

Vale ressaltar também que os resultados desta tese já estão sendo utilizados na indústria. Esse é o caso do gerador de cargas sintéticas, produzido no objetivo de pesquisa 2, que modela o comportamento de clientes em transmissões adaptativas ao vivo. O gerador foi cedido para ser usado em projetos internos da *Samsung Research Brazil*[1], no ano de 2020. Ao longo do desenvolvimento da tese, foram desenvolvidos geradores baseados na edição de 2014[2] e 2018[3] do evento considerado. Esses projetos estão disponíveis publicamente para a comunidade acadêmica. Vale salientar também que os artigos publicados já contam com mais de 20 citações, todas em publicações e conferências importantes de instituições como a IEEE e ACM.

[1] https://research.samsung.com/srbr
[2] https://github.com/thiagoguarnieri/adpgen-adaptive-workoad-generator-2014
[3] https://github.com/thiagoguarnieri/adpgen-adaptive-workoad-generator-2018

1.4.1 Sumário

O restante desta tese está organizado em 8 capítulos. O Capítulo 2 detalha de forma estruturada os conceitos relacionados com esta tese, como por exemplo o processo de transmissão adaptativa. O Capítulo 3, por sua vez, apresenta uma revisão da literatura relacionada, ressaltando as lacunas de pesquisa relacionadas com cada objetivo de pesquisa proposto, e o Capítulo 4 apresenta uma visão da infraestrutura de transmissão, detalhes de pré-processamento dos dados e heurísticas de cálculo de métricas de desempenho e engajamento.

O Capítulo 5 inicia a exploração dos objetivos de pesquisa propostos. Mais especificamente, ele apresenta uma caracterização do desempenho de transmissão de um grande evento transmitido ao vivo em larga escala, além da investigação da relação entre métricas contextuais e de desempenho de transmissão com o engajamento de usuários, contemplando com isso o OP1.

Já o Capítulo 6 explora o OP2 e apresenta uma proposta de modelo de comportamento de clientes para transmissões adaptativas ao vivo em larga escala, que tem como objetivo descrever, além de aspectos de permanência, o regime de adaptação do cliente e também as mudanças no comportamento induzidas pelo nível de desempenho experimentado.

Em seguida é explorado o OP3, no Capítulo 7. Esse capítulo apresenta propostas de modelos preditivos e descritivos de engajamento. A principal diferença do modelo proposto reside no fato de que suas variáveis independentes são baseadas na atividade de adaptação dos clientes, extraída dos *logs* do provedor, reduzindo a dependência de informações advindas do cliente, como seu nível de *buffer*, que são essenciais em modelos clássicos da literatura em geral. Além disso, é utilizado o conceito de cenários de desempenho, que combina a contribuição das métricas de desempenho clássicas e as novas métricas de adaptação, visando capacitar o modelo para previsões em clientes com diferentes níveis de desempenho registrados.

O Capítulo 8 fecha a exploração dos objetivos de pesquisa apresentando as contribuições para o OP4. É apresentada uma proposta de mecanismo de alocação de recursos para transmissões adaptativas ao vivo. Esse mecanismo visa conciliar os interesses de usuários e provedor de vídeo no consumo de recursos de transmissão.

Por fim, no Capítulo 9, é apresentada conclusão da tese, com as principais contribuições em cada objetivo de pesquisa e os trabalhos futuros que podem ser explorados a partir das contribuições da tese.

Capítulo 2

Referencial Teórico

Este capítulo apresenta os conceitos que servem de base para o trabalho desenvolvido na tese. Em primeiro lugar, a Seção 2.1 apresenta uma descrição a respeito de transmissões adaptativas via HTTP. Em seguida, na Seção 2.2, são descritos os conceitos relativos à medidas de aceitação, como o engajamento, e sua relação com desempenho de transmissão.

2.1 Transmissão Adaptativa via HTTP

A transmissão de vídeo adaptativa via protocolo HTTP (*HTTP adaptive streaming* – HAS) [96] é uma abordagem de transmissão de vídeo adotada pelos maiores serviços de vídeo atualmente. Ela é composta de uma arquitetura que define como a mídia é produzida e segmentada e um protocolo que define a troca de mensagens passadas entre clientes e provedores de conteúdo. Existem hoje diversas implementações comerciais desse padrão, como o *Apple HTTP live streaming* [98], *Microsoft Smooth Streaming* [126], *Adobe HTTP Dynamic Streaming* [1] e também *MPEG-DASH* [111], que é um padrão aberto criado pela *International Organization for Standardization* (ISO).

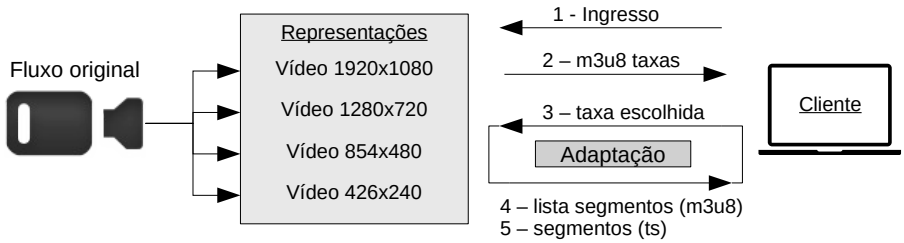
O objetivo do HAS é permitir que um fluxo de vídeo se ajuste a diferentes tipos de dispositivo e larguras de banda disponíveis. Ou seja, o HAS permite que uma mesma transmissão se adapte tanto a um cliente móvel de baixa resolução, acessando por meio de uma conexão móvel instável, quanto a um cliente em tela de alta definição com conexão de fibra óptica de alta velocidade.

Para isso, o padrão determina que o provedor codifique o vídeo em diversas *representações*, que são versões de um vídeo com taxas de transmissão e resoluções diferentes. A Figura 2.1 apresenta um exemplo de conjunto de representações para essa finalidade. Uma vez que o vídeo é codificado nas representações escolhidas, seu conteúdo é particionado em *segmentos* de tamanho fixo e disponibilizado para requisição. Quando um cliente $c \in C$ ingressa no sistema (etapa 1), ele recebe um arquivo de extensão *m3u8* com a lista de representações (etapa 2). Em seguida, na etapa 3, c escolhe uma dessas representações

e passa a receber listas de segmentos (arquivos *.ts* - etapas 4 e 5).

O tamanho das listas *m3u8* não têm um valor fixo. Por exemplo, foram feitas observações preliminares em dados no lado do cliente, referentes às transmissões ao vivo de jogos de futebol da Copa do Mundo de 2022, que mostraram que eventos ao vivo têm listas com 120 segmentos, contra 80 em vídeos sob demanda. O uso de listas maiores aumenta o intervalo de requisição das mesmas. Ou seja, diminui a chance de sobrecarregar o servidor de mídia pelo excesso de conexões abertas, principalmente em eventos que têm o potencial de atrair milhões de clientes simultâneos, como é o caso das transmissões dos jogos da Copa do Mundo.

Figura 2.1: Exemplo de conjunto de representações adequado a diferentes contextos



Fonte: Elaborado pelo autor.

Dando continuidade à descrição do processo de adaptação, vale ressaltar que as etapas 3, 4 e 5 ocorrem de maneira periódica e o cliente pode mudar de taxa no início de cada repetição ou mesmo a cada segmento. Essa mudança é orientada por um *algoritmo de adaptação*, que tem como objetivo monitorar variáveis como a largura de banda e *buffer* do cliente e ajustar (i.e., adaptar) a taxa de transmissão do vídeo do cliente de acordo com essas variáveis. Com isso, o usuário poderá perceber uma degradação na resolução de imagem caso haja uma adaptação para uma taxa de transmissão inferior à atualmente em uso. Entretanto, essa redução evita uma interrupção na reprodução, o que é benéfico, visto que, em termos de perda de desempenho de transmissão, a queda de qualidade de imagem é mais tolerada do que congelamentos de reprodução [67, 85, 92].

Ainda em relação à adaptação, vale destacar que também existem diferenças dependendo do tipo de conteúdo transmitido. Retornando à análise preliminar das transmissões dos jogos da Copa do Mundo de 2022, foi observado que os segmentos da transmissão ao vivo são menores do que os de vídeo sob demanda, indicando que o algoritmo de adaptação pode trocar de taxa mais vezes em vídeos ao vivo. Especula-se que a razão é que vídeos ao vivo podem passar por instabilidades mais frequentes e os clientes devem reagir mais rapidamente a elas. Essa observação também foi feita para transmissões ao vivo dos jogos da edição anterior do mesmo evento, ocorrida em 2018.

2.1.1 Arquitetura do Módulo de Adaptação

Para que uma aplicação suporte troca de representações, são necessários dois componentes: o *buffer* e o *algoritmo de adaptação*. O *buffer* armazena um certo número de segmentos para reprodução num futuro próximo. A função deste componente é evitar que o cliente fique sem segmentos para reproduzir em períodos em que a taxa de recepção de segmentos é menor que a velocidade em que esses segmentos são lidos pelo cliente. No entanto, se esse período for longo, pode haver um esvaziamento total do *buffer* e o consequente congelamento da reprodução. Assim, para evitar que isso aconteça, deve haver uma redução na taxa de transmissão, a fim de equiparar a taxa de recepção de segmentos com o regime de consumo dos mesmos pelo cliente. A tarefa de controle da taxa de transmissão é de responsabilidade do *algoritmo de adaptação*.

O algoritmo de adaptação é responsável pela manutenção do nível do *buffer* em um patamar mínimo. Mais especificamente, sua função é efetuar as trocas de representação ao longo da reprodução e sua lógica consiste no monitoramento de variáveis-chave como o *nível de buffer* do cliente e sua *vazão de rede*. A partir desses indicadores, há a geração de um *conjunto de regras* para determinar os casos em que se deve efetuar trocas de representação. Existem diversas formas de construir o conjunto de regras de adaptação [10]. Essas regras podem ser estáticas [82], ou dinamicamente aprendidas por meio de algoritmos de aprendizado de máquina, que incluem redes neurais [86], aprendizagem por reforço [16], entre outros.

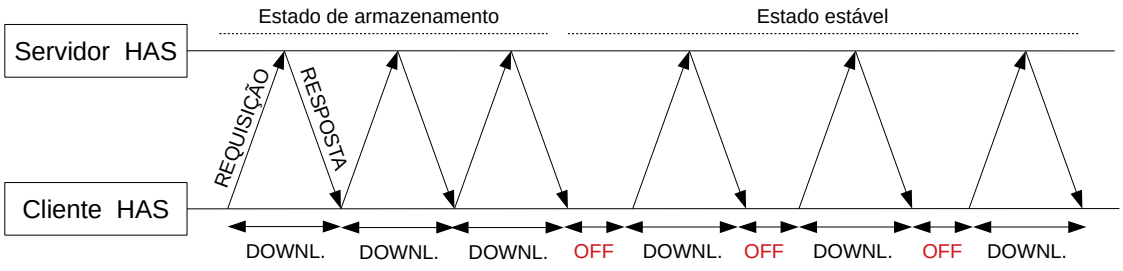
Ainda em relação ao funcionamento dos algoritmos de adaptação, existem diferenças quanto às variáveis que interferem nas decisões de troca de taxa de transmissão. Em alguns casos, essas decisões são baseadas em uma previsão de largura de banda [125, 112, 69], que pode ser imprecisa devido à natureza do próprio esquema de adaptação (ver Seção 2.1.3), ou no monitoramento do nível de *buffer* [58, 109, 68]. Algoritmos que consideram somente a ocupação de *buffer* também têm limitações, que incluem baixo desempenho de transmissão e problemas de instabilidade, especialmente em casos onde existe uma flutuação persistente na largura de banda [10]. Por fim, há também casos em que tanto a largura de banda quanto o *buffer* são monitorados, tentando conciliar as vantagens de ambas as abordagens [79, 108].

2.1.2 Regimes de Requisição

Uma sessão adaptativa pode alternar entre dois estados com regimes específicos de requisição de segmentos [60]. O primeiro é chamado de *estado de armazenamento* (*buffering state*). Nele, os segmentos requisitados são de baixa taxa de transmissão e toda largura de banda disponível é usada. Esse estado é utilizado quando o cliente está preenchendo seu *buffer* de segmentos ou quando seu nível se encontra abaixo de um patamar pré-estabelecido. Nesse caso, a escolha de segmentos de baixa taxa de transmissão tem a finalidade de reduzir o tempo necessário para o preenchimento do *buffer* e acelerar o início ou a retomada da reprodução do conteúdo.

Com o *buffer* acima do patamar definido, o cliente entra no *estado estável* (*steady state*). Nesse estado a taxa de requisições diminui para 1 segmento a cada n segundos, sendo n a duração do segmento [4, 10]. Isso faz com que o cliente alterne ciclos de transferência (*on*) e inatividade (*off*) e permite também que haja uma economia de banda na conexão de Internet do cliente. Porém, se o nível de *buffer* retornar a um valor abaixo do patamar pré-definido, há um retorno ao estado de armazenamento. É também durante o estado estável que o algoritmo de adaptação monitora os indicadores de recursos do cliente e pode decidir mudar a representação. A Figura 2.2 mostra a alternância entre os dois estados em uma sessão. Notar que no estado de armazenamento não há intervalos entre requisições consecutivas. Já no estado estável, é possível perceber que há um tempo de inatividade (*off*) entre pares de requisições.

Figura 2.2: Requisições de um cliente durante uma transmissão HAS



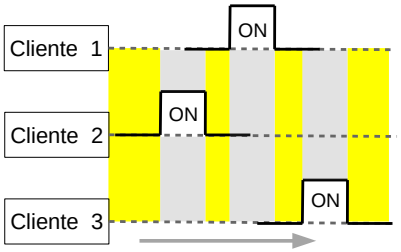
Fonte: Elaborado pelo autor.

2.1.3 Limitações do HAS

A despeito dos avanços, o HAS também possui deficiências. O primeiro é abordado por [4] e é ilustrado na Figura 2.3, que apresenta três clientes HAS compartilhando uma mesma conexão com a Internet. Assumindo clientes com *hardware* similar e uma banda disponível de x kbps, é desejável que cada cliente obtenha uma fração justa de, no máximo, $x/3$ kbps. A figura mostra como cada cliente, durante o estado estável, alterna entre períodos de transferência (*on*), em cinza e períodos de inatividade (*off*), em amarelo. Note que, no exemplo da figura, se pelo menos um dos clientes estimar a largura de banda disponível em qualquer um dos intervalos em amarelo, ele terá a falsa impressão de haver uma vazão disponível de x kbps, tendo em vista que, neste intervalo, nenhum cliente está requisitando segmentos de vídeo do servidor remoto. Com isso, este cliente tenderá a solicitar uma taxa acima de $x/3$ kbps, fazendo com que os outros clientes tenham que reduzir suas taxas para um valor abaixo de $x/3$ kbps para evitar congelamentos.

Notar ainda que essa situação se agrava à medida que mais clientes estimam a vazão nos períodos em amarelo. No pior caso, todos os clientes irão interpretar a vazão disponível como igual a x kbps e poderão solicitar o vídeo nesta taxa, levando a uma superestimação de banda três vezes maior do que a realmente disponível para esta conexão em particular. Como consequência, pode haver um aumento das chances de ocorrência de esvaziamento de *buffer* e congelamento de reprodução.

Figura 2.3: Exemplo com três clientes em estado estável na mesma transmissão



Fonte: Elaborado pelo autor.

O segundo problema se refere ao armazenamento (*caching*) de segmentos em transmissões ao vivo [84]. Um provedor pode distribuir geograficamente servidores secundários (*proxies*), com cópias de seu conteúdo, para aliviar a carga sobre o ponto central de distribuição. Com isso o serviço de vídeo se torna mais escalável. No entanto, transmissões adaptativas são disseminadas em diversas taxas de transmissão, o que requer uma quantidade muito maior de armazenamento nesses servidores. Além disso, uma transmissão ao vivo gera novos segmentos de forma contínua, o que faz com que estes segmentos tenham uma validade muito curta na memória dos servidores secundários e tenham que ser

continuamente substituídos. Como resultado, mais transferências a partir do servidor de origem são necessárias, o que aumenta as chances de sobrecarga em um ponto central.

Por fim, há um terceiro problema, relacionado ao caráter não-cooperativo dos clientes HAS. O objetivo central do algoritmo de adaptação é maximizar a taxa de transmissão disponível para seu cliente, sem levar em consideração a necessidade do provedor, que é oferecer desempenho de transmissão adequado para todos os clientes por meio dos recursos que dispõe e maximizar sua escala de transmissão. Essa característica pode levar a uma redução nos ganhos financeiros do provedor de mídia, decorrente de uma subutilização de seu *link* de conexão, e uma consequente redução do número de clientes atendidos.

Esta tese aborda em particular o terceiro problema acima mencionado, propondo mecanismos que visam alcançar um melhor compromisso entre o objetivo do algoritmo de adaptação, que é maximizar a taxa de transmissão do cliente, capturando portanto a perspectiva de cada cliente individualmente, e do provedor, que é ter mais clientes concomitantemente. Isso é feito por meio de conjuntos de taxas de transmissão personalizadas, baseadas no conhecimento dos requisitos de desempenho de cada perfil de usuário e uma coordenação mais global do compartilhamento dos recursos em transmissões adaptativas ao vivo.

2.2 Desempenho de Transmissão e Aceitação

A academia tem buscado formas de descrever a satisfação de um cliente a partir do desempenho de transmissão que o mesmo desfruta em sua sessão. O caminho mais comum para isso é construir um modelo que correlacione as métricas que descrevem tal desempenho com uma medida específica de satisfação. Isso permite priorizar os aspectos de desempenho mais importantes para os usuários.

As métricas de desempenho serão descritas na Seção 2.2.1. Já a Seção 2.2.3 apresenta o conceito de *aceitação de serviço*, que visa justamente descrever maneiras de medir a satisfação de um cliente por meio de uma métrica quantificável.

2.2.1 Fatores de Desempenho de Transmissão

De acordo com [12], um fator de desempenho é qualquer característica de serviço, aplicação ou contexto cujo estado ou configuração atual pode influenciar a satisfação de um

usuário. Nesse sentido, os autores em [106] classificam fatores de influência para aceitação em fatores de *sistema, humanos, contextuais e de conteúdo*. Os fatores de *sistema* são relacionados aos aspectos da aplicação (e.g., evento de troca de representação), rede (meio de acesso, largura de banda, latência) e cliente (lógica de adaptação, resolução, capacidade de processamento). Já os fatores *humanos* por sua vez são os relacionados com o histórico de uso e contexto social, psicológico e econômico. Os fatores *contextuais*, por sua vez, se referem ao aspecto ambiental, como a hora do dia, dispositivo, tipo de conteúdo, entre outros. Por fim, fatores de *conteúdo* se referem a características do vídeo, como duração, taxa de quadros e quantização.

A partir dos fatores descritos, é possível construir um conjunto de métricas que posteriormente podem compor um modelo que visa descrever uma medida de aceitação que caracteriza a satisfação do usuário com a transmissão. No contexto de transmissões adaptativas de vídeo há um conjunto de métricas amplamente abordado que são as métricas de Qualidade de Serviço (QoS). Exemplos de métricas desse tipo são a latência, que é o tempo que um pacote leva da origem ao destino, e o *jitter*, que é a variação da latência ao longo do tempo.

Um dos problemas do uso de métricas de QoS é que uma transmissão com boa Qualidade de Serviço pode não ser necessariamente atrativa do ponto de vista do usuário. Como já foi exemplificado, transmissões com baixa resolução podem ter alta QoS mas não oferecem uma boa experiência de exibição para os usuários, principalmente para aqueles que utilizam clientes executando em dispositivos de alta definição.

2.2.2 Métricas de Desempenho de Transmissão

A partir do entendimento das limitações das métricas de QoS de rede, foi proposto um outro conjunto de métricas, chamadas de *métricas de QoS no nível de aplicação*, nomeadas nesta tese como *métricas de desempenho de transmissão*, a título de simplificação de leitura. Essas métricas possibilitam a medição de aspectos que se relacionam melhor com a percepção de qualidade do usuário.

No começo, as métricas de desempenho de transmissão se relacionavam primariamente com a dinâmica de preenchimento de *buffer*. Quando o *buffer* esvazia, há a ocorrência de um congelamento, que é um evento capturado pelo sistema sensorial do usuário de maneira negativa e impacta negativamente em seu engajamento [27, 72]. As principais métricas relacionadas com a dinâmica de *buffer* abordadas nesta tese são a *latência de inicialização* e a *taxa de congelamentos por minuto*. A latência de inicialização se refere ao intervalo de tempo gasto para preencher o *buffer* do cliente pela

primeira vez, no início da reprodução. Já a taxa de congelamentos se refere a quantidade de congelamentos dividida pela duração da sessão em minutos.

Com o advento das transmissões adaptativas, novas métricas têm sido propostas. Essas métricas procuram quantificar fatores relacionados às trocas de taxa de transmissão do cliente e seu impacto para o engajamento. Esse conjunto de métricas é mais recente e seu impacto no engajamento ainda é pouco investigado. Nesse sentido, esta tese explora a *taxa de transmissão média* e a *taxa de adaptações*. A taxa de transmissão média é calculada tomando a média aritmética da taxa de transmissão de todos os segmentos. Já a taxa de adaptações se refere a quantidade de trocas de taxa de transmissão dividido pela duração da sessão em minutos.

Adicionalmente, foram propostas novas métricas de adaptação para dar suporte aos mecanismos propostos nesta tese. Exemplos dessas métricas são a *fração de permanência nas taxas de transmissão* e a *matriz de transição de taxas de transmissão*. Como será visto ao longo desta tese, a fração de permanência em cada taxa se refere a um conjunto de atributos que apresentam a porcentagem de tempo em cada taxa de transmissão. Essa métrica é mais expressiva que a taxa de transmissão média, uma vez que uma mesma média pode ser derivada de diferentes distribuições de valores dentro do conjunto de atributos mencionado. Por exemplo, assumindo duas sessões A e B, com uma taxa de transmissão média de ≈ 1430 Kbps e 3 segmentos cada. A taxa de transmissão dos segmentos de A pode ser $\{264, 891, 3127\}$ e a de B pode ser $\{891, 1337, 2085\}$. Nessas configurações, é possível perceber que A irá experimentar um desempenho menor que B porque A permaneceu mais tempo em taxas de transmissão menores. Já a matriz de transição de taxas de transmissão, por sua vez, representa um grafo direcionado ponderado, que descreve em detalhes a atividade de adaptação do cliente ao longo da sessão. As descrições das heurísticas de cálculo de cada métrica estão descritas no Capítulo 4, na Seção 4.4.

2.2.3 Aceitação de um Serviço

Em [33], os autores apresentam o conceito de *aceitação* (*acceptability*) de um serviço. De acordo com este conceito, um serviço tem maior aceitação pelo usuário, ou produz maior satisfação, à medida em que atende às suas expectativas de desempenho. Vale ressaltar que essa expectativa tem a ver com a forma com a qual cada usuário considera o que é alto desempenho. Nesse sentido, a relação entre a maneira que o usuário enxerga desempenho e como os requisitos desse desempenho são atendidos é denominado *Qualidade de Experiência (QoE)* [12].

Um dos grandes desafios no contexto das transmissões de vídeo é desenvolver uma métrica que permita monitorar a aceitação dos usuários de um provedor de vídeo. De maneira geral a aceitação é inferida a partir de um modelo que tenta construir uma visão geral da QoE do usuário, tomada por uma combinação de métricas de desempenho, que é em seguida expressa por uma medida unificada de aceitação.

Na literatura, a métrica de aceitação mais popular é o *Mean Opinion Score* ou *MOS* [62]. Nessa abordagem de medição, cada usuário que assiste ao vídeo deve, ao final da reprodução, preencher um formulário no qual dá notas para diversos aspectos relacionados ao desempenho de transmissão. Em seguida, um valor agregado (e.g., média) é calculado. Ou seja, o MOS é uma métrica qualitativa e, por essa razão, existe uma dificuldade inerente de se construir uma base de referência suficientemente grande, que possa generalizar para todos os perfis de usuários e para suas relações com desempenho da transmissão que podem existir em um sistema.

Para superar as deficiências em relação ao MOS e para que seja possível avaliar aceitação em larga escala foi proposta a noção de *engajamento*. O engajamento é um indicador quantitativo, que assume que a aceitação de um serviço por um usuário está diretamente relacionada à duração de sua permanência [27, 72, 3]. Isto é, quanto maior for a permanência do usuário em sua sessão de vídeo, maior é a sua aceitação. Desta forma, o engajamento pode ser avaliado em larga escala, bastando que os provedores de serviço consigam extrair dos seus *logs* a permanência total de cada cliente. Apesar dessas vantagens, o engajamento é uma medida fortemente influenciada por fatores contextuais como, por exemplo, o interesse do usuário no conteúdo sendo transmitido. Isto é, o engajamento pode ser baixo, mesmo com um desempenho satisfatório, simplesmente pelo fato de o conteúdo não ser do interesse do usuário. Devido a isso, o MOS e suas variações ainda são mais explorados na construção de modelos de aceitação do que o engajamento [9].

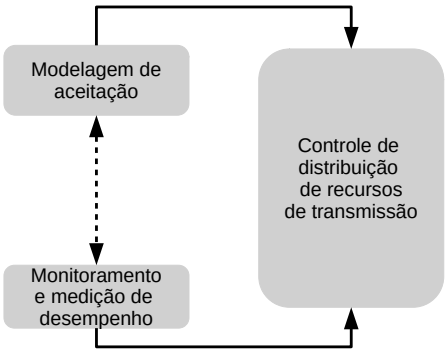
Outras métricas para medir aceitação e/ou construir uma visão da QoE do usuário têm sido utilizadas. A primeira delas é a VMAF [78], focada em aspectos de desempenho relacionados à codificação do vídeo, e a segunda é descrita no padrão ITU.T P.123 [63] que usa, além de características de codificação, métricas relacionadas a interrupções e taxa de transmissão.

2.2.4 Modelagem de Aceitação

Conforme enunciado em [33], a aceitação de um serviço (inferida a partir de MOS ou do engajamento do usuário) está vinculada ao desempenho. Esta é uma motivação para construir um modelo que correlacione métricas de desempenho e aceitação. Como a

Figura 2.4 mostra, um modelo acurado de aceitação é um componente crucial para um sistema de planejamento de recursos com foco na aceitação.

Figura 2.4: Gerenciamento de Desempenho



Fonte: Elaborado pelo autor.

São várias as vantagens de se conhecer o mapeamento entre métricas de desempenho e aceitação. Em [9], essas vantagens são classificadas de acordo com as diferentes partes interessadas. Em primeiro lugar, provedores de acesso e de serviço podem identificar os fatores de desempenho mais importantes para a aceitação de serviço, de forma a configurar adequadamente seus mecanismos de alocação e provisionamento para evitar perda dessa aceitação. Além disso, modelos de aceitação podem ser usados em tarefas antecipatórias como por exemplo a previsão de engajamento futuro. Em segundo lugar, modelos de aceitação podem ajudar no desenvolvimento de novos dispositivos de mídia, no sentido de que seu design garanta a preservação de fatores mais importantes para a aceitação (e.g., *buffers* maiores, adaptação otimizada para dispositivos com limitações de recursos, etc.). Em particular, o desenvolvimento de modelos de previsão e estimação do engajamento como medida de aceitação tem ganhado atenção da comunidade acadêmica [8, 123, 115, 114]. No entanto, alguns desafios permeiam esta tarefa:

- **Fatores contextuais:** variáveis como o interesse do usuário, expectativas, contexto ambiental e social podem exercer um impacto significativo sobre o engajamento. No entanto, esses fatores são de difícil monitoramento, coleta e inferência e, portanto, são pouco abordados [9].
- **Complexidade:** os modelos de previsão de aceitação existentes não foram pensados para execução em dispositivos com restrições de recursos, tendo em vista que utilizam algoritmos de alto custo computacional. Esse problema é particularmente importante, em função de já existirem relatos de eventos em larga escala nos quais mais da metade dos clientes executam em dispositivos móveis [85]. Outro problema é que muitos modelos necessitam de informações do conteúdo do vídeo [49, 105], o que inviabiliza sua aplicabilidade em transmissões criptografadas.

– **Localização do modelo**: muitos modelos de aceitação são implantados no cliente. No entanto, alterações podem ser feitas nesses modelos pelos usuários a fim de obter vantagens na distribuição de recursos. Exemplos são a modificação da prioridade de recepção de segmentos em sistemas descentralizados par-a-par e a alteração do regime de recepção de segmentos em algoritmos de adaptação. A implantação do modelo no provedor pode mitigar este problema. No entanto, o acesso a certos fatores de desempenho, como a ocorrência de congelamentos, é mais difícil a partir do provedor, pois têm a ver com estruturas de dados do cliente, como seu *buffer* de recepção.

2.2.5 Modelos Antecipatórios de Aceitação

A presente tese irá abordar um tipo específico de modelo de aceitação, que se destina à *previsão* e à *descrição de engajamento*. Mais especificamente, o objetivo é a construção de modelos que permitam descrever e prever engajamento a partir dos valores de métricas de desempenho de transmissão. Tais modelos permitem, por exemplo, antecipar possíveis abandonos e adotar medidas para mitigá-los, ou estabelecer quais fatores de desempenho devem ser priorizados para melhorar o engajamento dos usuários. Vale ressaltar que mecanismos antecipatórios, seja para prever engajamento ou outra variável, como nível de *buffer* e vazão, são particularmente úteis em transmissões ao vivo, nas quais o *buffer* do cliente é menor e, por isso, menos capaz de auxiliar o algoritmo de adaptação na tarefa de evitar falhas de desempenho e abandonos de sessão.

A construção dos modelos de previsão e descrição usa como base o ferramental teórico de aprendizado de máquina supervisionado e não-supervisionado. No aprendizado supervisionado, os modelos aprendem a relação entre métricas de desempenho e engajamento por meio de um treinamento, que é feito com base em um subconjunto dos dados. Em um momento posterior, esse aprendizado é validado na etapa de *teste*, na qual outro subconjunto, disjunto ao da etapa de treino, é utilizado para verificar se o aprendizado se generaliza para novas instâncias dos dados.

Alguns dos modelos propostos também utilizam, em maior ou menor grau, uma etapa prévia de *agrupamento* [8, 114], que divide os clientes em região geográfica, dispositivo ou outro atributo. Nesse caso, o objetivo é detectar diferentes tipos de perfis de usuários dentro do sistema e produzir modelos mais especializados para cada perfil. Nesta tese, essa abordagem é associada a um algoritmo de aprendizado não supervisionado, que visa a descoberta de grupos que não estão explícitos diretamente nos dados, mas sim por meio de arranjos específicos das métricas de desempenho. Os conceitos relativos ao treinamento e teste dos modelos serão melhor detalhados no Capítulo 7 e o agrupamento

para a construção dos perfis será explorado nos Capítulos 5, 6 e 7.

A descrição das métricas de desempenho utilizadas como atributos dos modelos, bem como os procedimentos de coleta, estão descritos no Capítulo 4. Foram utilizadas métricas tradicionais como taxa de congelamentos e latência de inicialização, assim como métricas mais recentes, relacionadas à dinâmica de adaptação do cliente. Também foi proposta a utilização de uma matriz de transições, como complemento para aumentar a acurácia dos modelos descritivos e preditivos (Seção 6.1).

2.3 Sumário

Este capítulo apresentou o referencial teórico que serve de contexto e motivação para o desenvolvimento da tese. Mais especificamente, na Seção 2.1 foram apresentados os conceitos relativos ao protocolo de transmissão adaptativa via HTTP. Já na Seção 2.2, foram apresentados os conceitos relativos à medição de desempenho em sessões de vídeo adaptativo e aceitação de um serviço. No próximo capítulo serão apresentados os trabalhos relacionados aos objetivos de pesquisa abordados nesta tese.

Capítulo 3

Trabalhos Relacionados

Este capítulo discute os trabalhos relacionados a cada um dos quatro objetivos de pesquisa desta tese. Na Seção 3.1 são descritos os trabalhos que analisam a relação entre métricas de desempenho e o engajamento (OP1). Em seguida, na Seção 3.2, são apresentados os trabalhos relacionados à caracterização de comportamento de clientes em sistemas de vídeo (OP2). Na Seção 3.3, por sua vez, são apresentadas as principais abordagens de modelagem de engajamento (OP3). Por fim, na Seção 3.4, são descritos os trabalhos da literatura que tratam de alocação de recursos em sistemas de transmissão de vídeo (OP4).

3.1 A Relação entre Desempenho de Transmissão e Engajamento

Diversos trabalhos têm constatado que a análise da relação entre engajamento e métricas de desempenho oferece intuições úteis das razões que levam um usuário a ficar satisfeito ou não com um serviço [27, 72, 103]. Além disso, estar ciente dessas razões pode facilitar os provedores a planejar transmissões que atendam mais eficazmente aos requisitos de desempenho de seus usuários. No entanto, o conhecimento da relação entre engajamento e desempenho ainda está em processo de construção, com estudos abordando o mesmo conjunto de fatores chegando a conclusões diferentes, dependendo da metodologia e os dados considerados [26, 94, 72].

O estudo do desempenho e sua relação com engajamento foca geralmente em um pequeno conjunto de métricas de desempenho, principalmente referentes à dinâmica do *buffer*, como a latência de inicialização, e quantidade, posição e duração de congelamentos [27, 72, 8, 3, 115]. Porém, com o advento das transmissões adaptativas, houve o desenvolvimento de novas métricas, sendo a taxa de transmissão média a mais utilizada [27, 72]. Além disso, observa-se que o estudo tem se expandido para métricas relacionadas ao funcionamento do algoritmo de adaptação propriamente dito. Nessa li-

nha, a métrica mais explorada, embora ainda em uma quantidade menor de estudos, é a taxa de adaptações [114].

Com base nesse contexto, esta seção apresenta uma revisão dos trabalhos relacionados ao OP1, que se refere à caracterização da relação entre desempenho de transmissão e engajamento. A seção foi dividida por tipo de métrica de desempenho, sendo a Seção 3.1.1 dedicada a relação das métricas de desempenho relacionadas ao *buffer* do cliente com o engajamento, e a Seção 3.1.2 dedicada às métricas de adaptação de taxa de transmissão, que ainda são menos exploradas na literatura. Por fim, na Seção 3.1.3, é apresentado os trabalhos que estudam o impacto de métricas contextuais para o engajamento.

3.1.1 Métricas de Buffer

A primeira métrica abordada nesta tese é a *latência de inicialização*. Os autores de [27] mostram que a latência de inicialização desestimula o retorno do usuário para novos vídeos. Além disso, o trabalho em [55] aponta que a latência tende a ser mais tolerada que congelamentos. Isso ocorre porque o congelamento é um evento inesperado que é interpretado pelo sistema sensorial do usuário de uma maneira mais negativa. Essa conclusão é corroborada em [94] e também em [26] para clientes móveis. Os autores em [72], por sua vez, refutam essa conclusão e indicam que uma latência de inicialização acima de 2 segundos já interfere negativamente no engajamento. Essa métrica impacta ainda mais negativamente serviços com uma grande variedade de conteúdos como o *Youtube*, nos quais um usuário começa a assistir vários vídeos e sai precocemente até encontrar algum cujo conteúdo lhe interessa [17]. Nesse caso, a tolerância a uma longa espera inicial é muito menor do que em um vídeo cujo conteúdo já é conhecido e está sendo buscado pelo usuário (e.g., uma transmissão de uma partida de futebol).

O segundo fator de influência abordado na tese é o *congelamento da reprodução*. Ele ocorre devido ao esvaziamento completo do *buffer* do cliente. Existe um consenso na literatura sobre o impacto negativo deste fator. Por exemplo, os autores em [100] mostram que interrupções longas são mais toleradas do que várias interrupções curtas. Já em [113] é mostrado que a posição da interrupção no vídeo também interfere na aceitação, enquanto o trabalho apresentado em [59] mostra que interrupções em intervalos irregulares são menos toleradas que em intervalos periódicos. Por fim, os autores de [56] concluem que usuários têm boa tolerância a apenas 1 congelamento por vídeo, desde que sua duração seja pequena. O baixo impacto de congelamentos curtos também é constatado em [74].

3.1.2 Métricas de Adaptação de Taxa de Transmissão

As métricas de adaptação, como já mencionado, são extraídas da dinâmica de trocas de taxa de transmissão do cliente. Uma das primeiras métricas que foram estudadas é a *taxa de transmissão média*. A literatura atesta que essa taxa tem impacto positivo na aceitação de um serviço por parte dos usuários, isto é, ao desfrutar de uma alta taxa de transmissão, os usuários tendem a permanecer por mais tempo em suas sessões e retornar mais vezes [27, 72, 3]. No entanto, a taxa de transferência média não possui uma correlação estritamente crescente com o engajamento [8]. Isso se deve ao fato de que ela omite diversos aspectos do desempenho, como por exemplo o tempo de permanência em cada taxa. Alguns poucos segmentos em uma taxa alta podem elevar a taxa de transmissão média para um valor que na verdade não representa o desempenho real da sessão.

Outro fator de influência relevante se refere à dinâmica de variação da taxas de transmissão (i.e., *adaptações*) ao longo de uma sessão. Existem ainda poucos estudos a respeito deste fator, pelo fato de que a taxa de transmissão dos segmentos de vídeo é raramente informada nos conjuntos de dados existentes. Dentre os trabalhos disponíveis na literatura atual, os autores de [124] mostram que o uso de abordagens adaptativas leva a uma redução de 80% dos congelamentos em cenários de redução de banda. Já em [95, 94] é mostrado que o número de abandonos aumenta mais devido às reduções de taxa do que nos aumentos. Além disso, uma taxa elevada de trocas também aumenta o número de abandonos [3, 74]. Por fim, os autores de [71] fazem um estudo do conjunto de representações em diversas plataformas de vídeo e mostram que nem sempre as taxas disponíveis atendem aos requisitos de desempenho dos clientes.

3.1.3 Métricas de Contexto

Um terceiro conjunto de estudos prévios focam no impacto de fatores contextuais na aceitação dos clientes. Um trabalho nessa categoria é o estudo apresentado em [92], que afirma que usuários de clientes móveis são mais tolerantes às interrupções. Já em [27], os autores mostram que uma alta taxa de transmissão é mais crucial em eventos ao vivo, enquanto que os autores de [25] argumentam que existe uma tendência dos vídeos serem assistidos em uma taxa de bits mais elevada em lugares onde o IDH é maior. Em [8], os autores mostram que o tipo de dispositivo usado, tipo de vídeo, conexão e horário do dia interferem no engajamento dos usuários. Os autores de [23], por sua vez, constatam

que usuários da plataforma *Android* têm menor desempenho se comparado àqueles em sistemas *Apple*. Além disso, usuários em conexões móveis tendem a ter engajamento mais alto, apesar da menor velocidade de conexão. Por fim, o provedor do usuário também já foi proposto em [2] como um atributo categórico para melhorar a detecção de falhas de desempenho.

O interesse no conteúdo também é um fator contextual que tem o potencial de afetar o engajamento. Em particular, há um maior impacto do interesse sobre o engajamento em vídeos sob demanda onde o usuário tem múltiplas opções de conteúdo e troca rapidamente entre elas [17]. Plataformas como o *Youtube* e *Twitch* se encaixam nessa descrição. Por outro lado, o estudo em [74] mostra que o impacto do interesse sobre o engajamento é menor em vídeos ao vivo, onde os usuários permanecem assistindo ao conteúdo por longos períodos. Além do tipo de transmissão, também há estudos que demonstram que o interesse impacta de formas diferentes conteúdos diferentes. Nesse sentido, o estudo em [18] mostra que o engajamento em conteúdos esportivos é menos afetado pelo interesse, isto é, saídas precoces nesse tipo de conteúdo se devem mais a falhas de desempenho do que falta de interesse.

Como se pode notar pela discussão apresentada, há muitos trabalhos que avaliam o impacto de fatores de desempenho na aceitação de clientes. De forma geral, esse impacto é estudado individualmente. Desta forma, ainda é limitado o conhecimento sobre possíveis interferências entre estes fatores e como os mesmos impactam em conjunto a aceitação dos clientes. Este é um dos objetivos explorados nesta tese (Capítulo 5). Outra questão em aberto abordada no presente trabalho é entender como se dão as escolhas das taxas de transmissão nos clientes em resposta aos diferentes níveis de desempenho experimentados (Capítulo 6). Como será apresentado no Capítulo 7, a utilização adequada desse conhecimento pode levar a modelos de engajamento mais precisos.

3.2 Modelagem do Comportamento de Clientes

Atualmente, uma transmissão ao vivo em larga escala pode facilmente alcançar centenas de milhares de usuários simultâneos [80, 85], impondo sobre os provedores uma alta carga de trabalho, nem sempre dimensionada de maneira satisfatória. Para calcular os recursos de maneira mais adequada, diversos estudos procuram desenvolver modelos que descrevem de forma precisa os padrões de comportamento desses usuários e dos clientes associados. Nesse sentido, esta seção apresenta os trabalhos relacionados com o segundo objetivo de pesquisa, que consiste precisamente na modelagem de comportamento de

clientes em sistema de transmissão de vídeo adaptativo.

No início dos anos 2000, quando a transmissão de vídeo começava a se tornar uma aplicação viável, tinha-se que esta se destinava a alguns nichos específicos como a área acadêmica, e atingia um público limitado, visto que as caras conexões *dial-up*[1] ainda eram a única forma de acesso para a maioria das pessoas. Arquiteturas cliente-servidor, amplamente adotadas pelas plataformas servidoras, também limitavam a escala de transmissão. Esse cenário impactou no comportamento dos usuários, de tal forma que a probabilidade de um vídeo curto ser assistido integralmente era muito maior que a de vídeos longos [5].

Um dos primeiros trabalhos destinados à modelagem de comportamento de usuários em plataformas de vídeo foi apresentado em [97], ainda no final da década de 90. Neste trabalho, os autores modelam a atividade de um usuário por meio de intervalos de download (*on-times*) intercalados com períodos de inatividade (*off-times*). Os autores de [5] estendem este modelo para capturar também a taxa de chegada de clientes e a popularidade dos vídeos. Os autores de [119], por sua vez, utilizam o modelo *ON-OFF* para descrever uma transmissão ao vivo com $3,5$ milhões de requisições adotando dois níveis de granularidade. No primeiro nível, de transferência, os *on-times* e *off-times* se referem à atividade do cliente durante uma sessão. Já no nível de sessão, essas métricas se referem à dinâmica do usuário entre sessões. Já em [24], o modelo *ON-OFF* é usado na análise de um conjunto de transmissões mais diversificado. Os dados são compostos de mídias de áudio e vídeo para conteúdos educacionais e de entretenimento. Tanto o estudo em [119] quanto em [24] caracterizam a duração das sessões, assim como a taxa de acessos ao longo do dia e semana.

A partir dos anos 2000, as redes de disseminação de vídeo passaram a adotar uma arquitetura descentralizada, do tipo *peer-to-peer* (P2P), em que os clientes formam uma rede sobreposta para compartilhar segmentos de vídeo entre si. Essa abordagem reduz a sobrecarga no servidor, aliviando, assim, a limitação de escala da abordagem cliente-servidor. Por outro lado, a topologia de uma rede P2P pode mudar dinamicamente, à medida em que clientes entram e saem do sistema (*churn*), o que interfere na sua capacidade de disseminação. Com isso, é mais difícil estabelecer garantias mínimas de desempenho ou até mesmo assegurar a disponibilidade do conteúdo. Além disso, os clientes de redes P2P podem estar espalhados ao longo do globo, o que impacta diretamente no custo por *bit* transmitido. Ainda assim, as redes P2P contribuíram para o aumento da eficiência na transmissão em larga escala de vídeos [76, 101]. Com isso, houve um crescimento da duração de sessões, assim como a quantidade de vídeos assistidos por usuário.

Um dos primeiros estudos que procura modelar o comportamento de usuários em um sistema P2P de vídeo ao vivo é apresentado em [50]. Neste estudo os autores analisam um evento de longa duração, transmitido pela plataforma *PPlive*, com audiência de 200

[1] http://xahlee.info/comp/bandwidth.html

mil usuários. Esse estudo foi estendido em [51] para capturar também a quantidade de tráfego redundante e propriedades das parcerias estabelecidas entre os *peers*. Já em [11], os autores caracterizam o comportamento de usuários do sistema *Sopcast*, também em uma transmissão ao vivo. Mais especificamente, são apresentadas as distribuições referentes ao regime de chegada dos clientes, número e duração das sessões, além aspectos relacionados à rede sobreposta, como por exemplo a quantidade e duração médias das conexões entre clientes. Um estudo mais recente [80] aborda o serviço chinês *PPTV*. Os autores apresentam uma análise de 8 milhões de usuários de vídeos pré-armazenados. Além disso, clientes móveis e fixos são analisados separadamente, o que permitiu evidenciar as distinções referentes ao comportamento de seus usuários. Em especial, foi constatado que usuários em clientes móveis permanecem por menos tempo e a probabilidade de falha de inicialização também é maior quando ele ingressa por meio de conexões móveis. Outra contribuição relevante foi a introdução do conceito de *sessão problemática*. Este tipo de sessão ocorre quando um usuário tem grande interesse no conteúdo transmitido, mas experimenta baixa QoS. Como resultado, esse tipo de sessão é caracterizada por uma grande quantidade de tentativas de conexão num curto espaço de tempo.

À medida em que serviços de vídeo foram alcançando maturidade, houve a necessidade de se ter um controle mais centralizado para garantir desempenho de transmissão mínimo para os usuários. Isso fez com que os provedores voltassem a investir nas arquiteturas cliente-servidor. Diversas melhorias foram propostas para lidar com as limitações dessa abordagem. Em relação à escalabilidade, intensificou-se o uso das redes de distribuição de conteúdo (*content distribution networks* - CDN). Já no lado do cliente surgiu a arquitetura de transmissão adaptativa, que foi apresentada na Seção 2.1.

Uma das primeiras iniciativas para caracterizar o comportamento de usuários neste cenário é apresentada em [43]. Neste trabalho, foram analisadas 8 milhões de sessões ao vivo e sob demanda provenientes de 485.000 usuários de uma rede de celular francesa. Os autores apresentam uma caracterização da duração da permanência dos clientes e popularidade de conteúdos, e propõem uma política de armazenamento de segmentos que aumenta a taxa de acertos em vídeos pré-armazenados. Mais recentemente, o trabalho em [77] caracteriza o comportamento de clientes móveis no serviço *Youku*. Neste trabalho, os autores caracterizam o comportamento de clientes com relação à sua frequência de acesso e evidenciam distinções nos padrões de comportamento entre usuários ocasionais e usuários frequentes, especialmente no que tange a permanência e preferências de conteúdo.

Por fim, em [85] é apresentada uma análise da evolução do comportamento dos clientes das Copas do Mundo FIFA de 2014 e 2018. São mostradas as diferenças em relação à duração da permanência dos clientes móveis e fixos, bem como a evolução de métricas de desempenho de transmissão, como taxa de congelamento e taxa de transmissão média. Apesar dos avanços, o trabalho faz admissões importantes que podem não se verificar na prática. Um exemplo é a de que os crescimentos e quedas de taxa de transmissão

têm o mesmo impacto para o engajamento. No Capítulo 5, é mostrado que adaptações positivas e negativas impactam de formas diferentes o engajamento dos usuários e que esse impacto depende ainda do desempenho geral do cliente. Outro aspecto é a admissão de um comportamento único generalizado para todos os clientes que, como mostrado no Capítulo 6, pode gerar dados sintéticos pouco realistas.

Em suma, os trabalhos apresentados têm foco em parâmetros específicos de comportamento de usuários como a duração de seu comportamento, bem como seu regime de chegada no sistema. Características do conteúdo como popularidade também são comuns nestes trabalhos. Por outro lado, aspectos atualmente relevantes como as decisões de adaptação do cliente ou o impacto do desempenho de transmissão no comportamento dos usuários ainda foram pouco explorados na literatura. Um dos objetivos desta tese é modelar o impacto do desempenho de transmissão no comportamento dos usuários e na dinâmica de adaptação dos clientes. Como será mostrado, é possível gerar cargas de trabalho sintéticas mais realistas que podem servir de base em tarefas de planejamento de consumo de recursos.

3.3 Modelagem de Engajamento

Conforme discutido na Seção 2.2, um modelo capaz de explicar a aceitação (e.g., engajamento) de um usuário a partir de métricas de desempenho de transmissão é útil para a criação de políticas de gerenciamento de recursos mais sensíveis à satisfação dos usuários. Nesse sentido, esta sessão se dedica a apresentar os trabalhos relacionados com o OP3, que é a criação de modelos descritivos e preditivos de engajamento. No entanto, sua criação é muito desafiadora devido à complexidade da interação entre os diversos tipos de métricas de desempenho existentes.

A maior parte dos esforços presentes na literatura para modelos de aceitação visa a métrica MOS. Um dos primeiros trabalhos de modelagem de MOS em transmissões de vídeo pela Internet foi apresentado em [91]. Nele, os autores propõem uma função de estimação de MOS que possui como entrada a latência de inicialização, frequência e duração de interrupções. No mesmo ano, o modelo foi expandido para considerar as reações do usuário a problemas de reprodução [90]. Ambos os modelos ignoram o impacto da adaptação. Já em [102], os autores propõem uma função com variáveis de entrada relacionadas a congelamentos na reprodução. O modelo foi expandido em [127] para admitir latência de inicialização e trocas de taxa de transmissão. Por fim, um modelo que retrata o MOS em termos da fração de tempo em cada taxa de transmissão e a amplitude

das trocas de taxa é proposto em [57]. Todos esses modelos têm em comum o fato de a validação ter sido feita com uma base de referência contendo uma pequena quantidade de usuários. Com isso, pode haver uma dificuldade de extrapolar os resultados obtidos para contextos mais realistas.

Ferramentas de mineração de dados e aprendizado de máquina também podem ser utilizadas para construir a relação entre MOS e desempenho de sessão. Isso é visto em [116], trabalho em que uma combinação de métricas de desempenho e de rede é empregada para treinamento de uma árvore de decisão. Apesar de ter sido validado em cenário mais realista, a partir de uma coleção de dados reais, este modelo não leva em conta métricas relacionadas à adaptação de taxa de transmissão.

Alguns estudos têm sugerido a utilização do *engajamento* para superar o problema de não haver uma medida de aceitação aplicável em larga escala. No entanto, como já mencionado, o engajamento é fortemente influenciado por fatores contextuais. Por isso, sua modelagem é ainda mais difícil. Um dos primeiros esforços de modelagem de engajamento pode ser encontrado em [8]. Nele, os autores utilizam métricas de desempenho e métricas contextuais para produzir um modelo descritivo de engajamento, treinado a partir de uma árvore de decisão, para classificar usuários de acordo com seu engajamento. Neste trabalho, o engajamento é relativo ao tempo total de transmissão é dividido em 10 classes (i.e., $\{10\%, 20\%, ..., 100\%\}$). Como resultado, os autores reportam uma acurácia de $\approx 69\%$. Ainda nessa linha, os autores de [114] apresentam um modelo de classificação ternária para determinar a razão de abandono de um usuário de sua sessão, que pode ser por falta de interesse, baixo desempenho ou por ambos. Neste trabalho, os autores utilizam máquinas de vetores de suporte – suport vector machine (SVM) – e reportam uma acurácia de classificação de tipo de abandono de $\approx 69\%$.

Já em [115], os autores incluem o interesse do usuário como variável independente. Para estimar o interesse, os autores utilizam o histórico de acesso dos clientes e filtragem colaborativa. Com essa abordagem, é reportado um erro médio de previsão de $\approx 29\%$. Em [123], é apresentado um método para prever o engajamento de vídeos no *Youtube* por meio de informações como a reputação, categoria e linguagem de um canal. Esse previsor permite estimar o engajamento médio que um vídeo terá ao longo de um mês, com um erro absoluto médio de menos de 8%. No entanto, este modelo não prevê o engajamento dentro de uma sessão individual. Os autores de [122], por sua vez, propõem um modelo que prevê o número de usuários concomitantes e seu engajamento em uma transmissão de futebol com base em eventos chave como, por exemplo, a marcação de um gol.

Além de [123], outros trabalhos como [99, 34] seguem linha similar, estudando a previsão de popularidade de vídeos publicados. De forma geral, essas abordagens se baseiam na avaliação de características relativas ao conteúdo. A abordagem desta tese, por outro lado, se concentra em lidar com aspectos relativos ao desempenho do sistema de transmissão, independente do tipo de conteúdo transmitido.

Apesar dos avanços apresentados nos trabalhos de modelagem de engajamento citados, nota-se pontos que podem ser explorados com maior profundidade. Em primeiro lugar, nenhum dos modelos explora adaptação e seu impacto no engajamento de usuários. Outro fato é que muitos deles assumem que existe informação sobre o histórico ou interesse do usuário. Também são trabalhos voltados a conteúdo transmitido sob demanda, cenário em que os acessos ao vídeo tendem a ser mais distribuídos durante um período de tempo mais longo, ao contrário de vídeos ao vivo em que os acessos tendem a se concentrar mais no lançamento do conteúdo [80]. Diferentemente desses trabalhos, nosso objetivo é a obtenção de um modelo preditivo de engajamento específico para eventos ao vivo, que não leva em conta o tipo de conteúdo e histórico de acessos de um cliente a outros vídeos, e que possa ser usado enquanto a sessão ou evento ainda está em andamento, para permitir o suporte a tomadas de decisão para manutenção de engajamento global por parte dos provedores.

3.4 Compromisso entre Usuários e Provedores no Uso de Recursos de Transmissão

Muitas aplicações na Internet precisam lidar com a existência de conflitos de interesses entre usuários e provedor no consumo de recursos do próprio provedor e da infraestrutura de transmissão em geral. Mais especificamente, esse conflito ocorre porque um usuário tem como meta obter para seu cliente todos os recursos que requisita do provedor a fim de atender a seus interesses particulares de desempenho. Já o provedor, por outro lado, precisa limitar a disponibilidade destes recursos em uma parcela dos seus clientes com o objetivo de atender a seus próprios interesses, que podem ser um aumento de escala de transmissão ou a priorização de clientes pagantes, por exemplo.

A literatura atesta que achar o ponto de equilíbrio entre os interesses de clientes e provedor é uma tarefa desafiadora e não trivial, sendo estudada em diversos contextos. Com base nessa afirmação, essa seção visa apresentar os mecanismos propostos na literatura relacionados ao objetivo de pesquisa 4, que aborda o conflito de interesses entre clientes e provedor no uso de recursos de transmissão.

No contexto geral de rede de distribuição de dados, os autores em [61] propõem um *framework* baseado na teoria dos jogos para otimizar o uso da infraestrutura ao mesmo tempo em que minimiza os custos. Para isso é sugerido que a precificação seja de acordo com os requisitos de cada carga. Já os autores em [104] analisam este problema do

ponto de vista da alocação de banda em *datacenters*. Por meio da análise de jogos não cooperativos, os autores mostram que uma precificação estática, que não leva em conta consumo de banda, juntamente com políticas de melhor esforço, pode levar à competição por recursos dos enlaces e a congestionamentos. Para resolver o problema, é sugerido um esquema de precificação de acordo com a utilização do enlace. Por fim, em [42], os autores analisam o conflito de interesses no contexto da computação em nuvem e mostram que parte da tarefa de transferência pode ser delegada aos usuários, reduzindo os custos para os provedores.

Considerando o escopo das transmissões adaptativas, a forma mais comum de guiar a alocação de recursos é interferir na escolha da taxa de transmissão do cliente, em particular no ingresso ao sistema. Ou seja, a liberdade do cliente de escolher sua taxa vai ser maior ou menor dependendo se essa escolha atende também aos interesses do provedor. Nesse sentido, os autores de [83] propõem um algoritmo que analisa o desempenho de transmissão médio em cada servidor de uma CDN a fim de selecionar aquele que produz o melhor desempenho de transmissão para o usuário. O desempenho é avaliado por meio de um algoritmo que compara cada combinação de {*servidor, taxa de transmissão*} buscando a que registra menor taxa de congelamentos e latência de inicialização. Esse algoritmo foi posteriormente aprimorado com a substituição da média do histórico de desempenho por abordagens de previsão que exploram técnicas de aprendizado de máquina [64] e, posteriormente, aprendizado por reforço [65].

Ainda em relação à escolha da melhor taxa de transmissão e servidor de CDN, é reportado o estudo em [8] que, ao contrário dos outros trabalhos, embasa sua decisão na combinação que proporciona maior engajamento futuro. Neste caso, otimizar o engajamento é um avanço em relação a simplesmente melhorar métricas de desempenho individuais, porque o engajamento é uma medida unificada de satisfação, construída com base no arranjo de múltiplos fatores de desempenho e seus níveis específicos de influência para cada tipo de contexto. Com isso, encontrar a combinação de servidor de CDN e taxa de transmissão que proporciona melhor engajamento faz com que o interesse do provedor, que é manter mais clientes concomitantes, seja melhor atendido. No entanto, apesar deste avanço, o trabalho mostra que a previsão de engajamento futuro tem ainda pouca precisão, ficando próximo de 69%. Com isso o engajamento previsto pode não refletir o engajamento que seria observado numa transmissão real.

Por fim, em [29], os autores abordam o problema do conflito de interesses em sistemas de vídeo adaptativo. Os autores sugerem um algoritmo que devolve ao cliente a taxa mais próxima da requisitada, respeitando restrições de recursos do provedor. A solução não considera que os clientes possam ter requisitos de desempenho diferentes, conforme argumentado na Seção 3.1. Assim, uma limitação pode impactar diferentemente clientes com contextos diferentes. Além disso, essa abordagem também necessita de acesso ao conteúdo do vídeo e não é adequada a transmissões com conteúdo criptografado.

Em suma, os trabalhos aqui apresentados oferecem um panorama das abordagens que são utilizadas para tentar amenizar os conflitos de interesses entre provedores e usuários. No contexto da distribuição de vídeo adaptativo, existe uma preocupação geral com o desempenho experimentado pelo cliente, mas não com o interesse do provedor, que é atender mais clientes concomitantemente. Nesse sentido, essa tese apresenta uma abordagem que trata diretamente da preservação de engajamento no estudo de conflito de interesses entre provedores de mídia e usuários. A garantia de engajamento, mesmo em cenários de restrição de recursos, é importante para que provedores de conteúdo melhorem suas métricas de lucro diretamente associadas à permanência do usuário em suas transmissões.

3.5 Sumário

Este capítulo detalhou os trabalhos relacionados a cada objetivo de pesquisa proposto. Mais especificamente, na Seção 3.1 foi apresentado um sumário do estado da arte no que tange a avaliação da relação entre métricas de desempenho e a aceitação, expressa principalmente pelo engajamento (OP1). O próximo capítulo (Capítulo 5) apresenta as contribuições desta tese em avançar este estado da arte, a partir da avaliação do desempenho oferecido em um conjunto de transmissões em larga escala e o impacto de métricas de desempenho no engajamento dos usuários. Além disso, também é apresentado um estudo dos de variações das métricas dentro de arranjos específicos denominados cenários de desempenho.

Na Seção 3.2 foi apresentada uma revisão da literatura relacionada à modelagem de comportamento de clientes em sistemas de vídeo (OP2). Foi possível observar que esses trabalhos focam em alguns aspectos chave, como a popularidade de conteúdos e perfil de permanência dos clientes, mas não consideram, ou consideram de forma limitada, a modelagem das escolhas de taxa de transmissão dos clientes e o impacto do desempenho de transmissão nessas escolhas. As contribuições desta tese no sentido de preencher essas lacunas estão descritas no Capítulo 6

Na Seção 3.3, por sua vez, foi discutido o estado-da-arte em modelos de aceitação (OP3). Esses modelos são especialmente úteis para quantificar a importância de fatores de desempenho para o engajamento e para o desenvolvimento de abordagens preditivas, que visam a antecipação de abandonos precoces e eventuais falhas de desempenho. No entanto, como a literatura atesta, ainda não existem modelos capazes de estabelecer uma correlação precisa entre métricas de desempenho e engajamento [103, 9]. As propostas para novos modelos de engajamento estão descritas no Capítulo 7.

Por fim, na Seção 3.4, foram discutidos os trabalhos da literatura que têm por objetivo o estudo do conflito de interesses na alocação de recursos em serviços de vídeo na Internet. Existem ainda poucos estudos que procuram conciliar os interesses de clientes e provedor em sistemas de vídeo adaptativo, especialmente usando engajamento como um critério de alocação de recursos. Contudo, os exemplos de outras áreas servem de insumo para a proposição do mecanismo de alocação de recursos que será apresentado no Capítulo 8, no contexto do objetivo de pesquisa 4.

Capítulo 4

Infraestrutura e Conjunto de Dados

Essa seção contempla dois objetivos fundamentais. O primeiro é dar detalhes a respeito da infraestrutura de transmissão e as tecnologias associadas à geração e distribuição da mídia adaptativa no contexto desta tese. Já o segundo objetivo é oferecer um primeiro panorama a respeito dos dados coletados no provedor de vídeo, bem como uma descrição sobre as heurísticas utilizadas para cálculo das métricas de desempenho, adaptação e engajamento que serão exploradas ao longo do trabalho.

Mais especificamente, na Seção 4.1 é mostrada a infraestrutura de transmissão e as tecnologias associadas a ela. Na Seção 4.2, por sua vez, é abordado o processo de individualização da atuação dos clientes a partir dos logs coletados do servidor. Na Seção 4.3 é mostrado em seguida uma visão geral a respeito dos dados do conjunto de transmissões considerado. Já na Seção 4.4, são detalhadas as métricas que foram utilizadas para medir o desempenho de transmissão e a qualidade de experiência dos usuários. Por fim, o capítulo é finalizado com o sumário, na Seção 4.5.

4.1 Infraestrutura de transmissão

Esta seção apresenta a arquitetura e organização dos componentes relacionados à transmissão do evento, bem como as tecnologias utilizadas. Essa estrutura foi utilizada para transmitir um grande evento global, a Copa do Mundo Fifa de 2018. O evento foi transmitido pela Rede Globo, tanto de maneira tradicional, pela sua rede de televisão, quanto pela internet, por meio de seu provedor Globo.com, entre 15 de junho e 15 de julho de 2018.

4.1.1 Tecnologias de Transmissão Adaptativa

A Globo.com transmite vídeo utilizando uma implementação do *HAS (HTTP Adaptive Streaming)* chamada *Apple HTTP Live Streaming* (HLS) [98].

Nessa implementação do HAS, a dinâmica de interação do cliente com o sistema obedece a um conjunto bem determinado de etapas, similar ao apresentado na Seção 2.1.

Depois que um cliente requisita o início da reprodução de um vídeo, ele recebe um arquivo extensão *.m3u8* que tem o objetivo de informar as características de cada representação disponível na transmissão (Figura 4.1). Essas características incluem a taxa de transmissão (*kbits*/segundo), a resolução, bem como a URL do segundo arquivo .m3u8, que contêm a lista de segmentos de vídeo (arquivos .ts) a serem transferidos para o cliente.

Figura 4.1: Exemplo de arquivo m3u8 inicial com a lista de representações

```
#EXTM3U
#EXT-X-VERSION:3
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=264000,RESOLUTION=384x216 estudiocgjrj_264/playlist.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=396000,RESOLUTION=512x288 estudiocgjrj_396/playlist.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=594000,RESOLUTION=512x288 estudiocgjrj_594/playlist.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=891000,RESOLUTION=640x360 estudiocgjrj_891/playlist.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=1337000,RESOLUTION=768x432 estudiocgjrj_1337/playlist.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=2085000,RESOLUTION=1280x720 estudiocgjrj_2085/playlist.m3u8
#EXT-X-STREAM-INF:PROGRAM-ID=1,BANDWIDTH=3127000,RESOLUTION=1280x720 estudiocgjrj_3127/playlist.m3u8
```

Fonte: Elaborado pelo autor.

O segundo arquivo .m3u8 (Figura 4.2), como dito, possui uma lista de segmentos que são baixados sequencialmente. Essa lista também contém informações a respeito da duração de cada segmento e também um número que identifica a lista corrente na sequência de listas de segmentos geradas pelo provedor de vídeo. Ou seja, é necessário requisitar uma nova lista sempre que todos os segmentos da lista corrente são consumidos.

Figura 4.2: Exemplo de trecho de arquivo m3u8 com Lista de segmentos

```
#EXTM3U
#EXT-X-VERSION:3
#EXT-X-TARGETDURATION:4
#EXT-X-MEDIA-SEQUENCE:23002
#EXT-X-PROGRAM-DATE-TIME:2018-06-14T16:13:52.000+00:00
#EXTINF:4.0000, estudiocgjrj_264-1528992832-385147162800.ts
#EXTINF:4.0000, estudiocgjrj_264-1528992836-385147522800.ts
#EXTINF:4.0000, estudiocgjrj_264-1528992840-385147882800.ts
#EXTINF:4.0000, estudiocgjrj_264-1528992844-385148242800.ts
```

Fonte: Elaborado pelo autor.

No conjunto de dados da Copa de 2018, os segmentos têm a duração de 4 segundos e são codificados em 7 representações (i.e., taxas de transmissão) que variam de 264

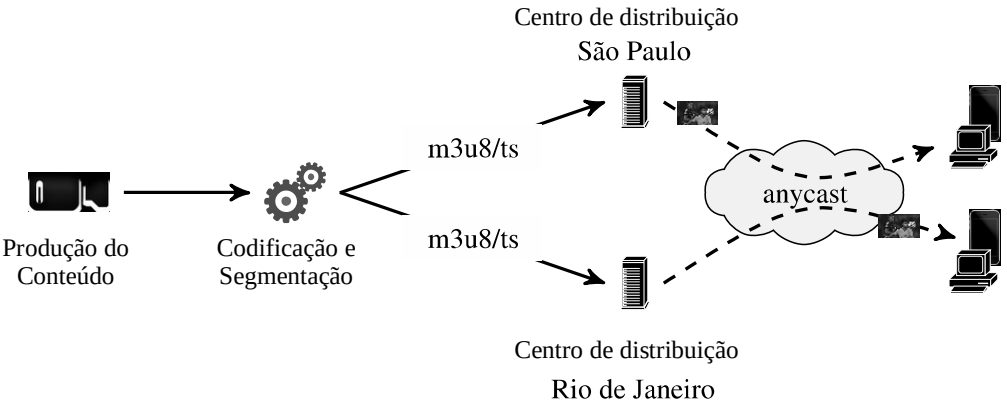
a 3127 kbps. Nesse sentido, vale dizer que a reprodução inicia sempre na menor taxa disponível e só aumenta após a avaliação da largura de banda do cliente.

No lado do Cliente, os usuários podem assistir aos conteúdos pela Internet a partir de um navegador na *web*, ou por aplicativo de dispositivos específicos como *smartphones*, *tablets* e *smartTVs*. O tocador de mídia utilizado pelo sistema da Globo.com é o *Clappr* [21], um programa de código aberto que é compatível com múltiplos navegadores. Nas transmissões ao vivo, os clientes não podem alterar a representação (isto é, as taxas de transmissão e resolução) utilizadas.

4.1.2 Arquitetura de Transmissão

Para distribuição de vídeo, é usada uma arquitetura cliente-servidor comum, ilustrada na Figura 4.3. Os dois centros de distribuição se localizam em São Paulo e no Rio de Janeiro. Eles estão ligados em pontos de troca de tráfego, onde podem se comunicar com outros provedores. Para associar os clientes ao centro mais próximo em número de enlaces, é utilizado o *Anycast* [88], que permite que um mesmo conjunto de prefixos possa apontar para os dois centros de distribuição simultaneamente. Comparado com 2014, a arquitetura em 2018 foi estendida pela contratação de uma rede de distribuição de conteúdo (CDN) externa para entregar conteúdo com resolução 4K [85].

Figura 4.3: Infraestrutura de transmissão da Globo.com



Fonte: Elaborado pelo autor.

Tanto os segmentos de vídeo quanto os metadados são transmitidos para dois centros de distribuição. Nestes centros, servidores executando a aplicação `Nginx` processam e respondem às requisições *HTTP* dos clientes. Essas requisições *HTTP* são registradas em formato *NCSA*, contendo a *data de requisição, horário, IP de origem, URL, status HTTP,*

bytes enviados e o user agent. No caso específico deste provedor, a taxa de transmissão do segmento faz parte da URL (como visto na Figura 4.2), assim como, em alguns casos, o nome da partida a qual o segmento pertence. Ter a informação da taxa de transmissão de forma explícita torna muito mais fácil a investigação da dinâmica de adaptação em uma sessão de vídeo.

No final de cada dia, estas mensagens de *log* são escritas em vários arquivos de texto (*plain/text*). Cada arquivo contém na casa de alguns milhões de registros, e possuem até 2 GB, variando entre 50 MB até 350 MB em formato compactado.

4.2 Delimitando a Atuação do Usuário na Transmissão

A delimitação do período de atuação do usuário no sistema é calculada pelo intervalo de tempo decorrido entre a recepção de seu primeiro e último segmentos. Nesse intervalo, um cliente pode sair e voltar múltiplas vezes, sendo que cada reentrada dá origem a uma nova *sessão* de cliente.

Uma sessão é a ordenação sequencial de segmentos de um cliente no tempo. Essa ordenação não existe nos *logs* de requisição de segmentos do provedor. Para obtê-la foi necessária uma etapa de pré-processamento, onde os *logs* foram unificados e armazenados na forma de um banco de dados relacional. Na geração do banco também foi feito o primeiro estágio de limpeza e *parsing* das linhas dos arquivos de *log*, com o objetivo de separar as informações importantes do cliente, como seu IP e *user-agent*. Em uma segunda etapa, durante a construção das informações agregadas de sessão, também foram determinados fatores contextuais adicionais como a localização geográfica, sistema autônomo, provedor, entre outros, com o auxílio de bibliotecas adicionais [87].

Dentro de uma sessão, requisições consecutivas de segmentos são separadas por um intervalo de tempo fixo igual à t, que no caso dos dados da tese é de 4 segundos, conforme explicado na Seção 2.1.2. Entretanto, um aumento de t é esperado no fim de uma sessão, pois o cliente só irá fazer novas requisições quando reingressar na transmissão, e também na ocorrência de congelamentos. Ou seja, não há como determinar com precisão se o aumento de t se deve a um abandono de sessão ou a um congelamento. Com base em estudos da literatura [103], foi adotada a premissa de que, em geral, um congelamento provoca a saída de um cliente em poucos segundos, e que basta considerar um tempo t significativamente longo como parâmetro para decretar com segurança que a ausência de requisições se deve ao fim de uma sessão. Por essa razão, foram avaliados tempos de 30 a

180 segundos, que levaram a resultados similares, sendo adotado então o valor definitivo de 120 segundos.

Para identificar os clientes de forma individual, foram utilizados o seus endereços *IP* e *user agents*, tendo em vista que poucas transmissões do conjunto implementaram chaves de identificação única por sessão. Essa abordagem ajuda a aliviar o impacto do uso de *NAT (Network Address Translation)*, porque clientes com navegadores diferentes podem ser identificados mesmo quando seu IP público é o mesmo.

Além da identificação por *IP+User Agent*, foram excluídas também as sessões com altas taxas de segmentos idênticos (acima de 5%). Isto é, são esperadas retransmissões de segmentos por alguma falha de recepção. Contudo um número muito alto e constante de segmentos repetidos é evidência de múltiplos clientes com mesmo *IP* e *user agent* assistindo ao mesmo conteúdo. No geral, menos de 5% das sessões em nosso conjunto de dados foi afetado por esse fenômeno.

4.3 Características Gerais dos dados

Esta seção apresenta detalhes gerais da carga de trabalho usada nesta tese. Esses dados dados representam ainda hoje uma perspectiva atualizada de um cenário de larga escala, ao vivo, que usa as mais recentes técnicas de transmissão adaptativa disponíveis.

O número de clientes e carga para cada partida está listado na Tabela 4.1[1]. Partidas que ocorreram aos fins de semana estão marcadas com a letra *W*. O conjunto de dados possui um total de 62 milhões de sessões provenientes de 38 milhões de clientes distintos, que geraram um tráfego 35 PB. Além disso, cada transmissão atraiu em média 837 mil sessões e produziu um consumo de banda médio de 769 TB, que representa um aumento de mais de 10 vezes no tráfego gerado em relação à edição anterior do evento [47].

Em relação aos dispositivos utilizados, aproximadamente 47% dos clientes rodam dispositivos móveis, sendo que desses, 25% usam aplicativo proprietário e 22% navegadores *web*. O restante, 52%, assistiu por meio de dispositivos fixos como computadores e TVs inteligentes. Essa distribuição é diferente da observada em 2014, que registrou mais de 90% de sessões em dispositivos fixos [47].

A Figura 4.4 descreve em mais detalhes a distribuição de dispositivos em cada partida do conjunto de dados. A figura mostra que em transmissões veiculadas aos fins de semana (em negrito), há um aumento médio de 6% no uso de dispositivos móveis, se comparado a transmissões em dias úteis. Em alguns casos essa diferença é ainda mais acentuada. Por exemplo, a partida Costa Rica vs. Sérvia, transmitida num fim de semana,

[1] Os dados cobrem 46 das 64 partidas

Tabela 4.1: Visão geral dos dados: número de clientes distintos, sessões e volume de tráfego por partida (transmissões aos fins de semana marcadas com a letra *W*)

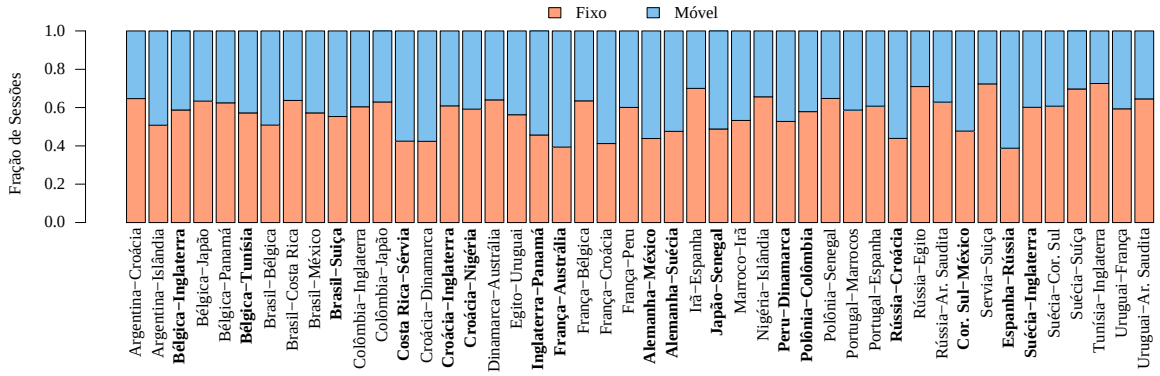
Fase de Grupos							
Partida	Clientes	Sessões	Carga(TB)	Partida	Clientes	Sessões	Carga(TB)
Russia x Arabia Saudita	1,0 M	1,59 M	872,14	Rússia x Egito	910,3 K	1,38 M	1051,52
Fança x Croácia	806,7 K	1,28 M	681,72	Irã x Espanha	893,2 K	1,39 M	1009,06
Egito x Uruguai	970,0 K	1,65 M	937,01	Portugal x Marrocos	1,2 M	2,06 M	1357,82
Marrocos x Irã	867,8 K	1,42 M	550,51	Uruguai x Arabia Saudita	880,6 K	1,50 M	784,42
Portugal x Espanha	1,2 M	2,14 M	1192,09	Argentina x Croácia	1,3 M	2,05 M	1317,17
Argentina x Islândia^{W}	800,5 K	1,29 M	733,91	Dinamarca x Austrália	881,8 K	1,40 M	939,52
Croácia x Nigéria^{W}	365,8 K	542,2 K	290,11	França x Peru	1,0 M	1,76 M	1034,40
França vs Austrália^{W}	539,3 K	805,3 K	287,55	Brasil x Costa Rica	837,1 K	1,24 M	862,40
Peru x Dinamarca^{W}	518,2 K	813,7 K	382,39	Nigéria x Islândia	735,1 K	1,12 M	459,05
Alemanha x México^{W}	767,1 K	1,24 M	502,09	Servia x Suiça	796,2 K	1,20 M	949,00
Brasil x Suíça^{W}	671,9 K	1,09 M	736,49	Alemanha x Suécia^{W}	645,3 K	993,91 K	497,09
Costa Rica x Sérvia^{W}	459,1 K	702,1 K	265,41	Bélgica x Tunísia^{W}	502,7 K	763,40 K	399,98
Bélgica x Panamá	858,0 K	1,36 M	265,41	Coreia do Sul x México^{W}	527,7 K	815,73 K	278,53
Suécia x Coreia do Sul	738,0 K	1,19 M	731,55	Inglaterra x Panamá^{W}	352,1 K	538,48 K	200,26
Tunísia x Inglaterra	787,4 K	1,22 M	923,25	Japão x Senegal^{W}	394,5 K	619,33 K	243,96
Colômbia x Japão	910,3 K	1,49 M	935,96	Polônia x Colômbia^{W}	338,7 K	502,28 K	264,55
Polônia x Senegal	850,1 K	1,36 M	802,63				
Fase Final							
Espanha x Rússia^{W}	790,6 K	1,30 M	573,53	Uruguai x França	1,5 M	2,36 M	1311,96
Bélgica x Japão	1,3 M	2,17 M	1256,93	Rússia x Croácia^{W}	603,6 K	915,95 K	457,06
Brasil x México	1,3 M	2,26 M	1224,75	Suécia x Inglaterra^{W}	394,5 K	565,16 K	279,49
Colômbia x Inglaterra	1,5 M	2,54 M	1651,23	França x Bélgica	1,4 M	2,38 M	1423,73
Suécia x Suíça	874,8 K	1,33 M	769,14	Croácia x Inglaterra	1,5 M	2,47 M	1740,78
Brasil x Bélgica	1,1 M	1,82 M	1062,24	Bélgica x Inglaterra^{W}	425,7 K	600,31 K	333,35

Fonte: Elaborado pelo autor.

registrou 48% mais clientes móveis que a partida Irã vs. Espanha, transmitida em um dia útil.

O dia de transmissão também interfere na audiência da transmissão. Jogos transmitidos aos fins de semana possuem menor audiência, se comparado às transmissões em dias úteis (média de 518 mil sessões em fins de semana vs. 910 mil em dias úteis), mesmo para conteúdos de grande interesse, como o jogo Brasil vs. Suíça. Ou seja, é possível observar que o comportamento do usuário, exemplificado pela sua escolha de dispositivo e interesse, pode depender do contexto, representado por sua vez pelo dia da semana.

Figura 4.4: Fração de sessões por dispositivo (fins de semana em negrito)



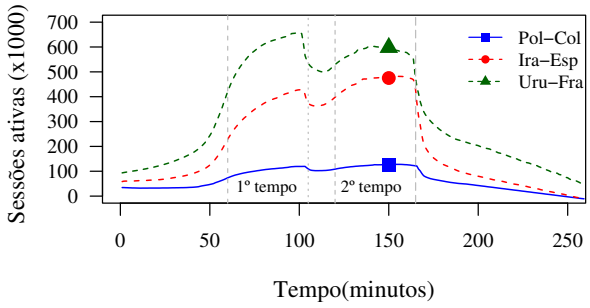
Fonte: Elaborado pelo autor.

4.3.1 Evolução da Carga de Trabalho

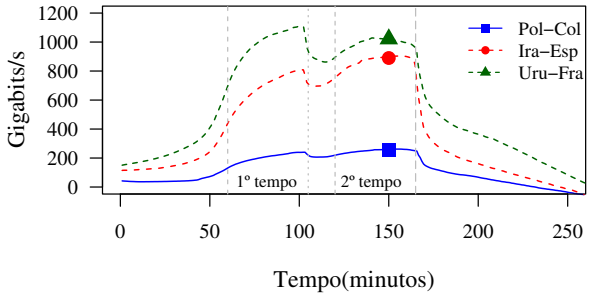
Para ilustrar o regime de chegada e permanência de usuários, foram utilizadas 3 partidas representativas: Polônia vs. Colômbia, com 502 mil sessões, como representante de uma transmissão de baixa carga, Irã vs. Espanha, com 1,39 milhões de sessões, para representar uma transmissão de carga média, e Uruguai vs. França, com 2,36 milhões de sessões, que ilustra uma transmissão de alta carga. Essas transmissões estão na Figura 4.5[2]. Observando a Figura, notam-se claras tendências em todas as partidas: **(1)** uma baixa audiência no período pré jogo; **(2)** um rápido crescimento no início da partida; **(3)** uma audiência relativamente estável durante o primeiro e segundo tempos; **(4)** uma pequena queda durante o intervalo da partida e **(5)** uma visível queda ao final da partida. Jogos com prorrogação prolongam e até aumentam a quantidade dos clientes (os jogos escolhidos não possuem prorrogação).

Figura 4.5: Audiência e largura de banda consumida ao longo da transmissão

(a) Número de clientes



(b) Largura de banda



Fonte: Elaborado pelo autor.

[2]As figuras foram calculadas assumindo intervalos de 1 s e contando o número de clientes e *bytes* de cada um desses intervalos.

4.4 Métricas de Desempenho e Engajamento

O objetivo desta tese é estudar o engajamento dos usuários, que é uma medida relacionada com Qualidade de Experiência (QoE). A tese parte da premissa de que o engajamento se relaciona com o desempenho de transmissão.

O engajamento nesse trabalho foi formalizado de múltiplas formas, todas relativas ao tempo de permanência do cliente na transmissão. Todas formas de engajamento consideradas neste trabalho estão descritas na Seção 4.4.1.

No que concerne a sua relação com desempenho, o engajamento pode ser descrito por combinações de valores em um conjunto de métricas, chamadas de *métricas de desempenho de transmissão* ou *métricas de QoS de aplicação*. Esse tipo de métrica tem sido adotada em oposição às métricas de QoS de rede, dada sua maior relação com aspectos visuais, que traduzem melhor a satisfação do usuário. Já as métricas de rede, como latência e *jitter*, visam medir o desempenho na perspectiva do sistema de transmissão e têm menor expressividade na tarefa de explicar engajamento [27, 103].

As métricas de desempenho de transmissão adotadas neste trabalho foram divididas em métricas relativas à adaptação de taxa de transmissão e métricas relativas a nível de *buffer*. Foram consideradas tanto as métricas mais utilizadas na indústria e academia [27, 103], quanto a proposição de novas medidas, com o objetivo de dar suporte aos modelos desenvolvidos ao longo da tese.

4.4.1 Métrica de Engajamento

O engajamento é uma métrica de aceitação de serviço e QoE, que nesta tese se refere ao tempo que o usuário permanece no sistema. No Capítulo 5, será mostrado que essa métrica tem correlação com as métricas de desempenho e adaptação mostradas na Seção 4.4.2. Ou seja, quanto maior o desempenho de transmissão, maior será o engajamento do usuário [27, 72].

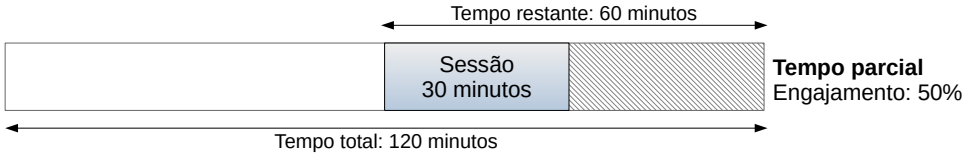
O engajamento como tempo do vídeo assistido pode ser modelado de diversas formas. Ao longo desta tese, diferentes noções de engajamento foram utilizadas, dependendo da análise efetuada. A Figura 4.6 mostra que o tempo de permanência não é utilizado em seu valor absoluto, mas sim normalizado. A parte sombreada indica que essa normalização pode ser pelo *tempo restante de transmissão* ou o t*empo total de transmissão*.

No Capítulo 5, foi utilizada como métrica de engajamento a *fração do tempo restante de transmissão*. Por exemplo, se o tempo restante de uma transmissão é 60 minutos

e o usuário assistiu 30, então seu engajamento é de 50%, como mostra a Figura 4.6a. Já no Capítulo 7, e na Seção 5.4 do Capítulo 5, que abordam algoritmos de classificação e regressão para o engajamento, optou-se por utilizar a *fração do tempo total*, ao invés do tempo restante, tendo em vista que algoritmos desta natureza na literatura utilizam esse tipo de normalização [8]. Nesse esquema, se o tempo total de uma transmissão é 120 minutos e o usuário assistiu 30, então seu engajamento é de 25%, como mostra a Figura 4.6b.

Figura 4.6: Tipos de normalização do engajamento do usuário

(a) Tempo restante



(b) Tempo Total

Fonte: Elaborado pelo autor.

Já no Capítulo 6, que trata do modelo de comportamento de clientes, foi utilizado uma forma expandida de engajamento, que é composta de 3 atributos, sendo eles o *tempo de sessão*, o *tempo entre sessões* e a *quantidade de sessões*.

É importante salientar que o engajamento tem forte influência de fatores contextuais subjetivos como interesse do usuário, e objetivos, como provedor e dispositivo. Sempre que possível, esses fatores serão levados em consideração.

4.4.2 Métricas de Desempenho e Adaptação

Esta seção apresenta as métricas de desempenho e adaptação e as suas respectivas heurísticas de estimação. Como já mencionado, essas métricas servem de base para o entendimento do engajamento do usuário e podem dar origem a modelos produzidos a fim de auxiliar no direcionamento de recursos para aspectos de desempenho que preservam melhor o engajamento dos usuários. Em primeiro lugar, serão mostradas as métricas

associadas ao *buffer* do cliente e, em seguida, as métricas relacionadas com a dinâmica de adaptação do cliente.

Métricas de Nível de Buffer

As métricas de nível de *buffer* se referem a dinâmica de preenchimento de *buffer*. A ocorrência de esvaziamento de *buffer* é um dos eventos que mais impacta negativamente o engajamento dos usuários [27, 72]. A informação sobre o nível de *buffer* do cliente não é armazenada no mesmo conjunto de *logs* que registra a requisição de segmentos. Por isso, as métricas relacionadas ao *buffer* precisam ser estimadas a partir de heurísticas que se baseiam nas informações contidas nos dados disponibilizados.

- **Latência de inicialização**. Ela mede o intervalo de tempo decorrido entre o acesso do cliente ao vídeo e o início de sua reprodução ou, de maneira mais específica, o tempo necessário para que o *buffer* do cliente seja preenchido pela primeira vez na sessão.

  A heurística para calcular a latência de inicialização foi monitorar a quantidade de segmentos que são baixados no estado de armazenamento de uma sessão, isto é, no período em que são transferidos na maior velocidade possível, conforme descrito na Seção 2.1.2. Foi assumido que o estado de armazenamento só é ativado para o preenchimento do *buffer* quando este se esvazia completamente. Durante este estado, foi observada uma média de 4 segmentos, o que leva a 16 segundos de *buffer* (assumindo 4 segundos por segmento). Como resultado, a latência de inicialização é o intervalo de tempo entre a primeira e a quarta requisições de segmentos de uma sessão.

- **Taxa de congelamentos**. Ela consiste no número de esvaziamentos de *buffer* dividido pela duração em minutos. O congelamento é um evento acessível somente por meio de instrumentação de clientes, que é um processo pouco escalável e pode gerar problemas de privacidade. Assim, foi utilizada uma heurística para simular o *buffer* do cliente com base nos logs de requisição de segmentos [68].

  Essa simulação utiliza uma fila *FIFO* (*First-in, First-out*) que é alimentada pela chegada dos segmentos nos clientes. É assumido por simplificação em que um cliente recebe os segmentos assim que suas requisições são atendidas (i.e., são registradas nos *logs* do servidor). Essa admissão é conservadora e pode superestimar o nível de *buffer* do cliente e subestimar a quantidade de congelamentos. Na prática, é esperado que o impacto na estimação da taxa de congelamentos seja baixo, pois a duração do segmento é pelo menos uma ordem de magnitude maior que média atual de RTT (*Round-Trip time*) de segmentos de rede.

  Durante a simulação, se o intervalo entre requisições de segmentos for longo, há uma diminuição da fila. Se esse intervalo se mantiver longo por muito tempo, há então o esvaziamento da fila e a sinalização de um congelamento. Vale salientar que o instante

em que o congelamento é detectado pela heurística pode não coincidir de forma exata com o instante de sua ocorrência real. Contudo, inspeções em uma amostra das sessões mostraram que essa detecção muitas vezes coincide com o disparo do estado de armazenamento, evento que geralmente ocorre quando o *buffer* está completamente vazio, ou seja, quando ocorreu de fato o congelamento.

Métricas de Adaptação de Taxa de Transmissão

As métricas de adaptação são baseadas na dinâmica de trocas de taxa de transmissão ao longo de uma sessão efetuadas pelo algoritmo de adaptação. Essas métricas começaram a ser estudadas com o advento das abordagens adaptativas de transmissão. Nesse sentido, procura-se estudar como os diversos aspectos relativos à variação de taxa de transmissão influenciam na percepção de qualidade por parte dos usuários. A seguir estão listadas as métricas derivadas da atuação da adaptação no cliente.

- **Taxa de transmissão média**. Obter uma alta taxa é fundamental em vídeos ao vivo [27]. Ela é calculada pela média aritmética da taxa de transmissão de todos os segmentos de uma sessão. Como descreve a Seção 4.1.2, a taxa de transmissão está apresentada na URL de cada segmento.
- **Distribuição de permanência nas taxas de transmissão**. Um dos problemas da taxa de transmissão é que ela oculta o tempo gasto pelo cliente em cada taxa de transmissão. Por essa razão, foi proposto como uma métrica adicional a distribuição de permanência nas taxas de transmissão. Isto é, um conjunto de 7 atributos que quantificam a porcentagem de segmentos em cada taxa. Essa métrica é mais expressiva porque uma mesma taxa de transmissão média pode ter como fonte diferentes distribuições de permanência, que por sua vez, oferecem desempenhos distintos. Essa métrica é utilizada nos modelos de engajamento e comportamento de clientes desta tese.
- **Taxa de adaptações**. Expressa o número de trocas de taxa durante uma sessão. O conhecimento do impacto da adaptação no engajamento do usuário é ainda hoje menos explorado, porque a taxa de transmissão de um segmento é uma informação que só pode ser obtida se o segmento for decodificado ou se essa informação constar de alguma forma em seus metadados, algo que em geral não ocorre. Foi considerada separadamente a taxa de adaptação negativa, que trata das reduções da taxa de transmissão, e a taxa de adaptação positiva, que conta os aumentos de taxa.
- **Matriz de transição de taxas de transmissão**. A distribuição de permanência nas taxas de transmissão mostra apenas uma visão estacionária (i.e., de longo prazo) da adaptação dos clientes. Essa matriz, por outro lado, é mais expressiva e descreve as probabilidades de transição entre nós de um grafo direcionado. Cada vértice representa uma taxa de transmissão disponível para um evento. Já as arestas determinam a pro-

babilidade de taxa de transmissão para o próximo segmento. Ou seja, assume-se que a taxa de transmissão de um segmento n depende apenas da taxa do segmento $n-1$. Essa nova métrica é abordada a partir do Capítulo 6.

4.5 Sumário

Esta Seção teve como objetivo apresentar um panorama a respeito da infraestrutura de transmissão e dos dados originados dessas transmissões. Em particular, foi apresentado o conceito de sessão, para delimitar a atuação do cliente durante sua permanência no sistema. Em seguida, detalhes iniciais a respeito dos dados foram apresentados no que tange a distribuição da quantidade de clientes e tipo de dispositivos.

Por fim foram apresentadas as métricas de desempenho e adaptação, que visam descrever em conjunto o engajamento do usuário, modelado por aspectos de permanência e padrão de retorno dos usuários.

Os conceitos apresentados servem de base para o entendimento das análises e mecanismos que serão apresentados nos próximos capítulos.

Capítulo 5

Caracterizando Desempenho de Transmissão e sua Relação com Engajamento

Este capítulo aborda o primeiro objetivo de pesquisa, relativo à caracterização de desempenho em transmissões adaptativas ao vivo e sua relação com o engajamento.

Diversos trabalhos da literatura têm sido desenvolvidos com a proposta de conhecer a relação entre desempenho e engajamento, uma vez que esse conhecimento ajuda na melhoria de mecanismos de disseminação de vídeo na Internet. No entanto, apesar dos avanços obtidos na compreensão dessa relação, existem ainda alguns pontos que carecem de melhor entendimento. Nota-se que ainda é comum a premissa de que o engajamento de todos os usuários é afetado de maneira similar, desconsiderando o efeito do contexto para a percepção pessoal de desempenho ou, em geral, é aplicada uma diferenciação básica, por tipo de dispositivo, por exemplo [72, 3, 85]. O problema desse tipo de diferenciação é que ela desconsidera o impacto combinado de métricas de desempenho no engajamento [8], mesmo em clientes com dispositivos similares. Por exemplo, dois usuários móveis, que experimentam um congelamento, podem reagir a este evento de maneiras diferentes, dependendo do valor da sua taxa de transmissão (Seção 5.4).

Outro fato relevante é que a maioria das análises de desempenho em transmissões de vídeo se concentra em métricas relativas ao *buffer* do cliente, como a latência de inicialização e a taxa de congelamentos, descritas no Capítulo 4. Por outro lado, o estudo da adaptação e seu impacto para o engajamento ainda carece de aprofundamento [94].

Assim, com base no cenário exposto, este capítulo visa contribuir para um melhor entendimento sobre o desempenho de transmissão em um evento recente ao vivo em larga escala. Esse estudo abarca diversos aspectos que podem influenciar no desempenho de uma transmissão, incluindo fatores contextuais como o tipo de dispositivo, provedor e nível de carga de trabalho. Além disso, a relação entre métricas de desempenho de transmissão e engajamento é explorada em detalhes, tanto a partir da perspectiva de cada métrica em particular, quanto em relação ao arranjo de múltiplas métricas, utilizando um conceito de integração de métricas que foi denominado de *cenários de desempenho*.

Vale recordar que o desempenho é medido por meio de métricas de QoS da aplicação cliente, também chamadas de métricas de desempenho de transmissão. Nesse sentido, as métricas para medição do desempenho de transmissão deste capítulo são um subconjunto das métricas apresentadas na Seção 4.4. Mais especificamente serão abordadas a *latência de inicialização*, a *taxa de adaptação*, a *taxa de congelamentos* e a *taxa de adaptação média*. Além disso, na Seção 5.4, também será utilizada a *distribuição de permanência por taxa de transmissão* no processo de construção de cenários de desempenho.

O estudo é conduzido da seguinte forma. Primeiro, é mostrado, na Seção 5.1, uma análise do impacto de fatores contextuais no engajamento e métricas de desempenho de transmissão. Já na Seção 5.2, é apresentada uma caracterização do desempenho de transmissão em clientes fixos e móveis, bem como o impacto da escala de transmissão e provedor no desempenho de transmissão. Para fechar a análise, na Seção 5.3 é apresentado um estudo do impacto no engajamento devido a variações nas métricas de desempenho de transmissão. Em seguida, na Seção 5.4 é proposta uma forma alternativa de análise por meio da construção de cenários de desempenho. Esses cenários têm por objetivo integrar o efeito de múltiplas métricas de desempenho de transmissão. Por fim, na Seção 5.5, são sumarizadas as contribuições do capítulo.

5.1 Impacto do Contexto no Engajamento e no Desempenho de Transmissão

Eventos contextuais não são percebidos visualmente pelo usuário, ao contrário de, por exemplo, uma troca de taxa de transmissão ou um congelamento. No entanto, fatores contextuais têm o potencial de interferir no desempenho de transmissão e no engajamento. Exemplos de tais fatores são o tipo de dispositivo, velocidade de conexão, localização geográfica, entre outros. Por isso é importante medir o impacto de fatores contextuais para que seja possível classificar melhor os usuários segundo seu contexto e empregar uma análise mais refinada do desempenho e sua relação com engajamento.

Para decidir quais são os fatores contextuais que mais interferem no desempenho e engajamento, foi utilizada a noção de *ganho de informação*. Esse processo é útil para a descoberta de relacionamentos ocultos entre fatores. A vantagem deste método é que ele não faz nenhum tipo de admissão a respeito da natureza das relações, isto é, se são estritamente crescentes ou decrescentes ou se são lineares, por exemplo. Em outras palavras, por meio do ganho de informação é possível avaliar quais fatores contextuais mais interferem nas métricas de desempenho de transmissão e no engajamento.

O ganho de informação é baseado na ideia de *entropia* de uma variável aleatória Y definida como $H(Y) = \sum_i P[Y = y_i] \log \frac{1}{P[Y=y_i]}$ onde $P[Y = y_i]$ é a probabilidade de $Y = y_i$. Ela representa o número de *bits* que precisam ser transmitidos para identificar Y a partir de n probabilidades equiprováveis. Quanto menor a entropia, mais uniforme será a distribuição dessas probabilidades.

Já a entropia condicional de Y dado outra variável aleatória X é $H(Y|X) = \sum_j P[X = X_j]H(Y|x_j)$. Essa entropia representa o número de bits que devem ser transmitidos a fim de identificar Y dado que ambos emissor e receptor conhecem X. Com isso, o ganho de informação é $H(Y) - H(Y|X)$, que consiste no número de bits economizados em média na transmissão de Y quando emissor e receptor conhecem X. Isto é, quanto mais correlacionadas duas variáveis, maior será a quantidade de bits economizada. Será utilizado nas medidas o ganho de informação relativo, que consiste em $RIG(Y|X) = \frac{H(Y)-H(Y|X)}{H(Y)}$ e pode assumir valores de 0 até 1.

5.1.1 Fatores Contextuais

Os fatores contextuais cuja relação com desempenho e engajamento é analisada neste trabalho estão listados a seguir:

– **Conteúdo:** este fator se refere à partida que está sendo transmitida. A ideia central é investigar se o apelo da partida é um fator capaz de explicar uma parcela do engajamento do usuário e desempenho de transmissão nos clientes.

– **Período do dia:** é um fator ternário, que divide o dia em {*manhã, tarde, noite*}. O objetivo é verificar se o engajamento e as métricas de desempenho de transmissão variam consistentemente ao longo de um dia. Esse fator foi importante em edições anteriores deste evento, onde foi registrado que partidas noturnas tinham menor engajamento. A informação sobre o período do dia foi retirada diretamente dos *logs* utilizados neste trabalho, a partir do *timestamp* do segmento.

– **Cidade:** é a lista de estados e cidades. Pretende-se verificar o impacto da localização geográfica dos usuários no engajamento e nas métricas de desempenho de transmissão. A informação de localização é determinada a partir do endereço IP do cliente com o auxílio de biblioteca externa [87].

– **Período da partida:** é um fator que descreve as etapas chave de uma partida, isto é, o {*pré-jogo, primeiro tempo, intervalo, segundo tempo, prorrogação/pós-jogo*}. O objetivo é verificar o impacto do interesse do usuário nas diversas etapas do jogo e sua

relação com o engajamento e métricas de desempenho de transmissão. Uma sessão é rotulada em uma etapa específica considerando o *timestamp* de seu primeiro segmento. Por exemplo, uma sessão é rotulada como de *primeiro tempo* se ela entre o minuto 1 e 45 da partida. A título de simplificação, os acréscimos foram ignorados.

– **Dispositivo e plataforma:** são fatores que descrevem o tipo de dispositivo {*móvel, fixo*} e o sistema operacional utilizado. Essa informação é extraída da decodificação do *user agent* do cliente, que é uma informação contida também nos *logs* de requisição de segmentos. É utilizada uma biblioteca externa para a decodificação [28].

– **Sistema autônomo:** A inclusão deste fator serve para verificar se há diferenças de desempenho dependendo de cada sistema autônomo e seu respectivo provedor de serviços de Internet de origem. A informação sobre o sistema autônomo é obtida por meio de biblioteca externa a partir do endereço IP do cliente [87].

– **Localização geográfica:** descreve a latitude e a longitude do cliente. A informação sobre localização geográfica aproximada é obtida por meio de biblioteca externa a partir do endereço IP do cliente [87].

5.1.2 Caracterização do Impacto do Contexto

A Tabela 5.1 apresenta a correlação calculada por meio do ganho de informação. As colunas representam os fatores contextuais extraídos dos dados. A primeira linha apresenta o impacto de fatores contextuais para o engajamento. Já as demais linhas tratam do impacto dos fatores contextuais para as métricas de desempenho de transmissão.

Como a Tabela 5.1 mostra, são quatro fatores contextuais mais relevantes, tanto para o engajamento quanto para as métricas de desempenho de transmissão. São eles o *tipo de dispositivo, plataforma, período do jogo e sistema autônomo*. Esses fatores apresentaram um ganho de informação acima de 1% para todas as métricas e com isso são responsáveis por caracterizar uma parte não desprezível do desempenho de transmissão e do engajamento.

A seguir, são descritos os impactos das métricas contextuais, primeiro para o engajamento, e depois para as métricas de desempenho de transmissão.

Contexto versus Engajamento

Começando pelo impacto do período do jogo no engajamento. Sua relevância vem do fato de que certos pontos da partida atraem mais o interesse do usuário, isto é, há momentos

Tabela 5.1: Ganho de informação relativo. Fatores com ganho $> 0,01$ em destaque

	Conteúdo	Período Dia	Cidade	Período jogo	Dispos.	Plataf.	Sist. Aut.	Loc. Geograf.
Engajamento	0,0032	0,0002	0,0056	0,0725	0,1252	0,0868	0,0142	0,0069/0,0071
Latência Inic.	0,0022	0,0001	0,0062	0,0447	0,0323	0,0283	0,0101	0,0076/0,0075
Tx. transm. média	0,0041	0,0002	0,0041	0,0444	0,0745	0,0583	0,0112	0,0046/0,0046
Tx. Congelamento	0,0022	0,0004	0,0080	0,0099	0,0297	0,0313	0,0130	0,0095/0,0094
adaptação +	0,0017	0,0068	0,0100	0,0068	0,0329	0,0428	0,0150	0,0113/0,0112
adaptação -	0,0016	0,0001	0,0085	0,0059	0,0261	0,0306	0,0127	0,0095/0,0094

Fonte: Elaborado pelo autor.

em que esse interesse é alto, como durante a partida, e baixo, como no intervalo e no final. Nota-se esse fato através da Figura 4.5. Eventos chave durante a partida também podem influenciar nesse interesse, como a ocorrência de um gol [122]. Esse resultado revela que, embora grande parte do engajamento esteja associada às métricas de desempenho de transmissão, há também uma fração que é induzida por efeitos coletivos, como a saída de grandes grupos de usuários.

O impacto do dispositivo e plataforma no engajamento se deve ao fato de que usuários em clientes móveis tendem a assistir o conteúdo por menos tempo, tendo em vista suas limitações de energia e conexão. Por fim, a influência do sistema autônomo é fruto de um impacto indireto do desempenho de transmissão. Na Seção 5.2.4, é mostrado que existem provedores que oferecem menor desempenho, o que por sua vez interfere negativamente no engajamento, como será mostrado também nesse capítulo.

Vale salientar que a comparação dos resultados obtidos aqui com os previamente computados para o conjunto de dados de 2014 revelou mudanças no impacto de fatores contextuais no engajamento. Por exemplo, o período do dia, que antes apresentava alto ganho de informação, já não influencia mais o engajamento. Isso sugere que os usuários passaram a usar dispositivos ligados à rede independentemente do horário, ao invés de recorrerem ao sinal de TV tradicional quando estão em casa.

Contexto versus Métricas de Desempenho de Transmissão.

No caso do impacto de fatores contextuais em métricas de desempenho de transmissão, constatou-se que o ganho de informação decorrente do tipo de dispositivo e plataforma se deve às características particulares de hardware e conexão em clientes móveis, que fazem com que sua dinâmica de adaptação seja diferente da dos clientes fixos e isso influencie diretamente nas métricas de desempenho de transmissão. Será visto no Capítulo 6 que os regimes de adaptação de clientes fixos e móveis são muito distintos entre si, algo que impacta diretamente na quantidade de ocorrências de adaptações e congelamentos.

Além das características do dispositivo, notou-se também o impacto do sistema autônomo nas métricas. Uma das razões pode ser associada à infraestrutura do provedor de conectividade, que nem sempre pode estar preparado para lidar com a quantidade de requisições a que é submetido. Com isso, é possível que haja uma variação no desempenho

dependendo do provedor ao qual o cliente está conectado. Na Seção 5.2.4, esse fato será explorado com mais detalhes, apresentando uma caracterização de desempenho de transmissão nos principais provedores de conectividade registrados.

Por fim, observou-se também que o período do jogo impacta no desempenho, em especial na taxa de transmissão média e latência de inicialização. Esse impacto pode ser explicado por meio de uma possível relação entre a quantidade de clientes servidos e desempenho. Isto é, as partidas possuem picos de audiência em momentos específicos. Nesses momentos, foram registradas reduções da taxa de transmissão, frutos de um possível esforço para manter a escalabilidade do serviço, o que pode provocar queda de desempenho geral. A Seção 5.2.1 trará mais evidências para embasar essa hipótese.

Em suma, com base na análise desta seção, é possível concluir que o período da partida, tipo de dispositivo e sistema autônomo foram os fatores contextuais que mais interferiram no conjunto das métricas de desempenho de transmissão e engajamento. Tomando como base essa conclusão, a caracterização de desempenho de transmissão e engajamento, mostrada a seguir, será dividida por tipo de dispositivo. Já na Seção 5.2.4, será abordado o desempenho de transmissão nos principais provedores de conectividade do conjunto de dados. O período da partida, por outro lado, se vincula ao interesse do usuário, que é um fator menos associado ao desempenho de transmissão. Por isso o período da partida não será abordado nessa caracterização.

5.2 Caracterização de Desempenho de Transmissão

Esta seção apresenta uma caracterização das métricas de desempenho associadas de transmissão no conjunto de dados considerado. O desempenho de transmissão é medido quantitativamente por meio das métricas mostradas na Seção 4.4.2. Na literatura essas métricas estão vinculadas principalmente à ocorrência de congelamentos. Adicionalmente, foram incluídas também métricas relativas ao regime de trocas de taxa de transmissão, menos exploradas na literatura. A seção inicia com uma avaliação do impacto da escala no desempenho de transmissão. A partir da Seção 5.2.2, a análise foi segmentada considerando os fatores contextuais mais relevantes, isto é, dispositivo e provedor de conectividade.

5.2.1 Desempenho de Transmissão versus Escala

Esta seção analisa a evolução da taxa de transmissão, como um indicador de desempenho de transmissão, ao longo de partidas e sua relação com o número de clientes atendidos. Os dados irão mostrar que existe uma tendência dos clientes experimentarem menores taxas em transmissões em partidas com maior número de clientes. Esse fato é ilustrado pela Figura 5.1, que apresenta a fração de segmentos de cada taxa ao longo de uma transmissão para clientes fixos. A figura divide a partida em janelas consecutivas de 10 minutos. É possível notar, sobretudo após o início da partida, que os jogos de alta carga (Uruguai vs. França e Brasil vs. México) possuem uma fração de segmentos de 3127 kbps, a maior taxa de transmissão, abaixo de 30% para a maior parte da partida. Já para as transmissões com carga baixa (Croácia vs. Nigéria e Polônia vs. Colômbia), registrou-se uma fração acima de 50% de segmentos referentes à taxa máxima durante toda a partida[1].

A aparente correlação negativa entre número de clientes e taxa de transmissão parece sugerir que em algum ponto da infraestrutura de transmissão foi preciso sacrificar a taxa de transmissão a fim de garantir a disponibilidade do serviço. Esse é um sinal de que as ferramentas de planejamento podem não ter sido capazes de prever apropriadamente a quantidade de recursos necessária para a transmissão. O estudo da taxa de transmissão e sua relação com a escala será expandido na Seção 5.2.2.

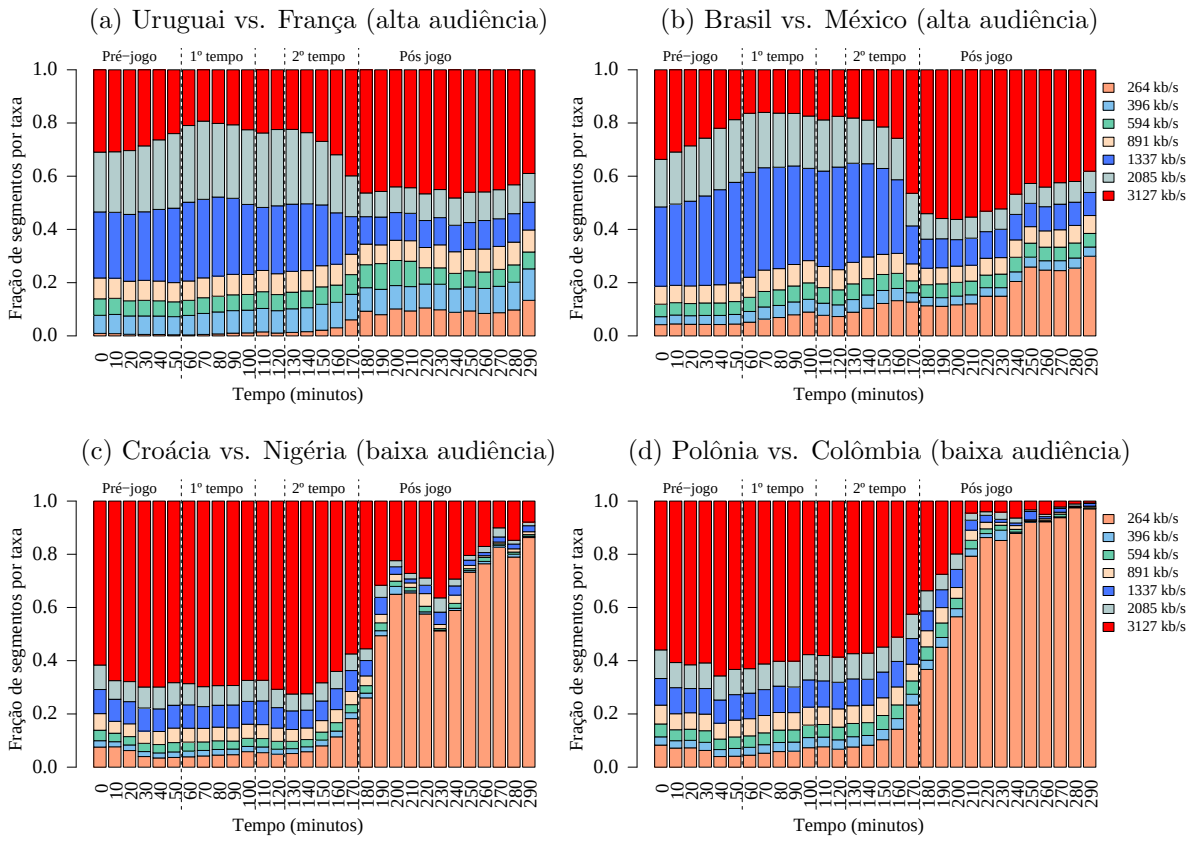
A Figura 5.2 apresenta a distribuição de taxas para as mesmas partidas, mas agora considerando apenas clientes móveis. Observações similares podem ser feitas. Embora as diferenças entre transmissões com alta e baixa audiência sejam menos evidentes, elas não são marginais. Por exemplo, considerando o período compreendido entre o início e fim da partida, é possível observar que a fração de segmentos da maior taxa para o jogo Polônia vs. Colômbia (baixa audiência) é 39% e no jogo Uruguai vs. França (alta audiência) é menor, de 29%.

Observando os resultados apresentados, têm-se evidências de que o desempenho, em termos de taxa de transmissão, diminui possivelmente em função do aumento do número de clientes. Essa diminuição de taxa tem impacto direto no engajamento dos usuários, como mostra a Seção 5.3.

Uma consideração importante é que a caracterização não foi dividida de acordo com o centro de distribuição ao qual cada cliente se conectou. Isso se deve à nossa análise de ganho de informação sobre os dados da edição de 2014 do evento, que revelou que considerar esse atributo não produziu mudanças substanciais no desempenho dos clientes.

[1]Os primeiros segmentos de uma sessão são da menor taxa disponível. Por isso sessões de curta duração acabam possuindo uma alta fração de segmentos de baixa taxa de transmissão. Isso explica o resultado da Figura 5.1c e 5.2d. O final dessas partidas é dominado por sessões de curta duração.

Figura 5.1: Distribuição de taxa de transmissão durante a exibição (sessões fixas)



Fonte: Elaborado pelo autor.

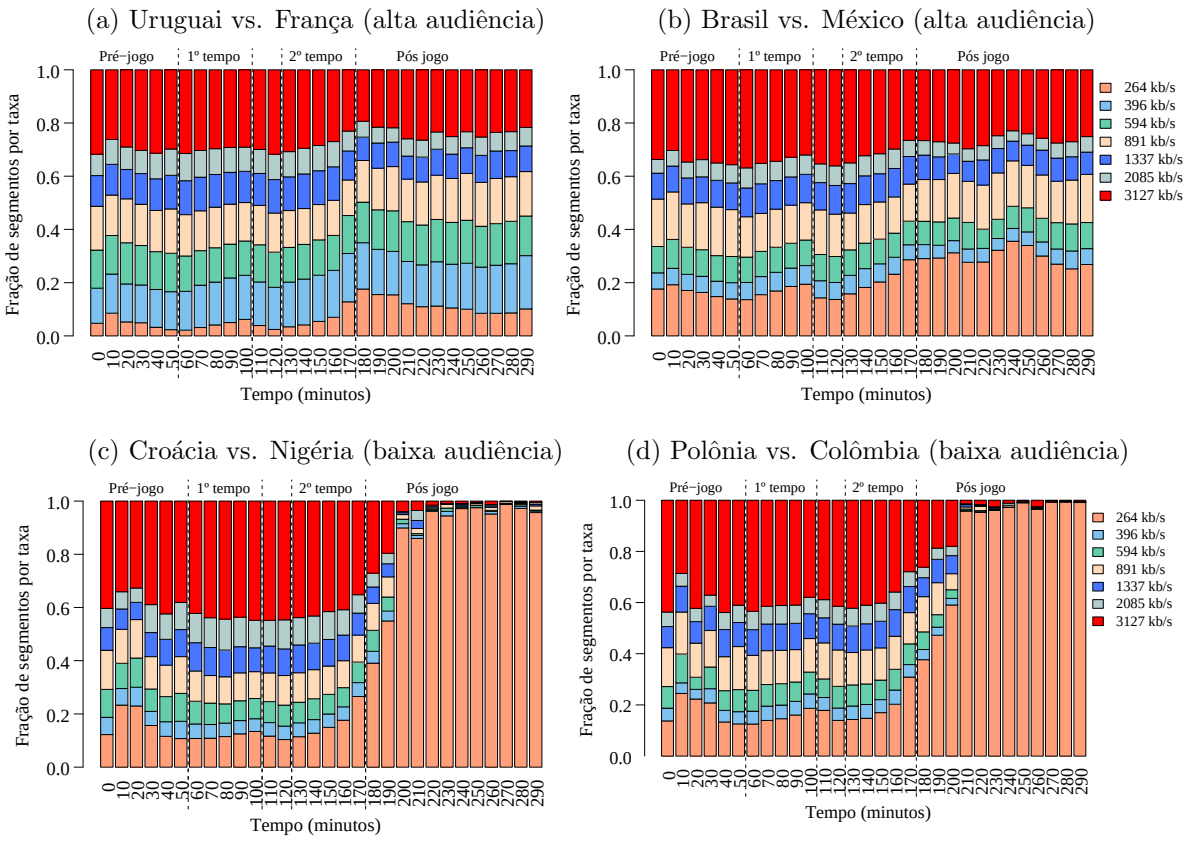
5.2.2 Desempenho de Transmissão em Clientes Fixos

Essa seção caracteriza o desempenho de transmissão para dispositivos fixos por meio das métricas mostradas na Seção 4.4. Para esta seção, foram escolhidos jogos que abrangem todos os tipos de carga registrados durante o evento (em termos de quantidade de clientes e padrão de chegada).

Tanto na Figura 5.3 quanto nas demais, que apresentam as CDF's das métricas de desempenho de transmissão, os eixos x foram discretizados em 50 intervalos (*bins*). Para a taxa de transmissão média, os intervalos têm tamanhos iguais. Para as outras métricas, a divisão foi logarítmica (como a escala), assim cada intervalo cresce exponencialmente.

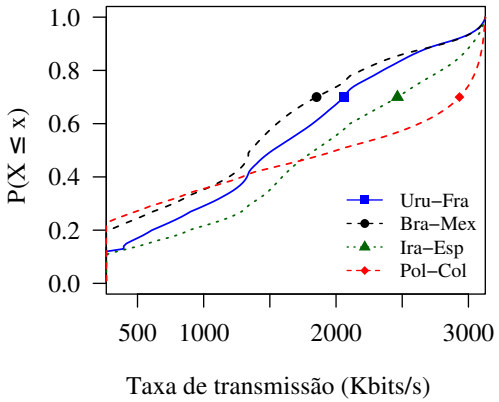
A Figura 5.3 mostra as distribuições acumuladas da taxa de transmissão média para as 4 partidas. Como pode ser visto, os clientes experimentam uma ampla faixa de taxas de transmissão. Nesse sentido, nota-se que os clientes da partida Brasil vs. México experimentam menores taxas de transmissão, se comparado aos da partida Polônia vs. Colômbia. Como a Tabela 4.1 descreve, Polônia vs. Colômbia foi um jogo com uma audiência mais baixa que a do jogo Brasil vs México. Sendo assim, o que a Figura 5.3 sugere é mais

Figura 5.2: Distribuição de taxa de transmissão durante a exibição (sessões móveis)



Fonte: Elaborado pelo autor.

Figura 5.3: Taxa transmissão média

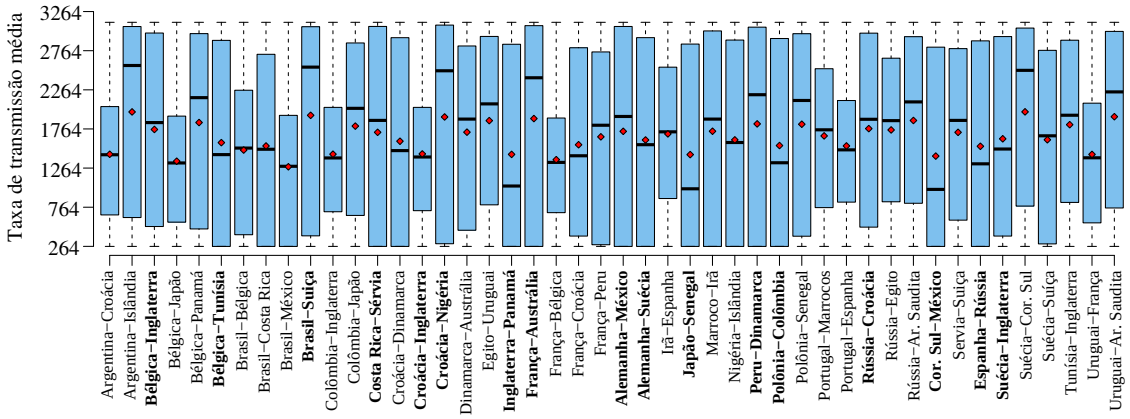


Fonte: Elaborado pelo autor.

uma evidência de que a taxa de transmissão média diminui à medida em que o número de clientes aumenta. Não é possível assegurar, por meio dos dados disponíveis, o que provoca a queda da taxa de transmissão. Ela pode ser fruto de uma redução tanto na infraestrutura de transmissão (provedores de mídia e conectividade), com o objetivo de atender a um número maior de clientes, quanto nos clientes, em resposta a um eventual

aumento no número de congelamentos por reduções repentinas em sua largura de banda, por exemplo.

Figura 5.4: Distribuição da taxa de transmissão média (clientes fixos)



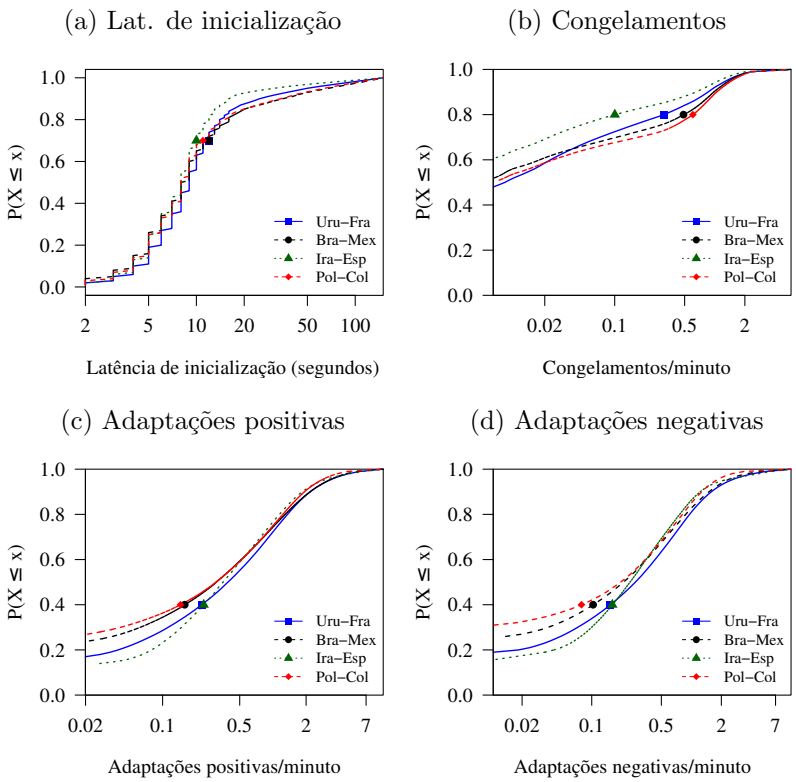
Fonte: Elaborado pelo autor.

A possível relação entre escala e taxa de transmissão é acentuada pela Figura 5.4, que apresenta o gráfico de caixa para a taxa de transmissão média em cada partida. Neste gráfico, a extremidade inferior e superior da caixa correspondem ao 25° e 75° percentil, respectivamente. O meio corresponde à mediana e os extremos denotam os limites inferior e superior. Os pontos vermelhos denotam a taxa de transmissão média de cada partida. Observando a figura e, tomando como base a audiência registrada nos jogos (Tabela 4.1), observou-se que, à medida em que a audiência de uma partida é maior, menor é a taxa de transmissão média disponível.

Para reforçar a tese da correlação entre escala e desempenho, foi calculado o coeficiente de *Spearman* entre a taxa de transmissão média e o número de clientes no pico de audiência de cada jogo. O resultado foi um coeficiente $\rho = -0.91$, que portanto traz mais indícios de que a taxa de transmissão média diminui à medida em que o número de clientes fixos aumenta. Esse resultado demonstra que o planejamento de alocação de recursos pode ter subestimado a demanda esperada para a transmissão. Como mostra a Seção 5.3, uma baixa taxa de transmissão pode levar a reduções no engajamento de usuários.

Dando continuidade à análise de desempenho de transmissão, a Figura 5.5a mostra que uma fração considerável de clientes tem uma latência de inicialização elevada, com 40% das sessões tendo latências acima de 10 segundos. Um valor similar foi registrado para a edição anterior do evento, ocorrida em 2014 [47]. Mesmo no passado, quando a largura de banda das conexões residenciais era menor do que hoje e os usuários estavam acostumados com longas esperas, esse tempo já era considerado longo. Trabalhos como [72] atestavam que latências acima de 2 segundos já tinham potencial para reduzir a aceitação de usuários.

Figura 5.5: Métricas associadas de desempenho em clientes fixos



Fonte: Elaborado pelo autor.

A Figura 5.5b, por sua vez, apresenta a taxa de congelamentos por minuto nas 4 transmissões. Essa figura mostra que cerca de 10% das sessões têm acima de 1 congelamento por minuto. Esse resultado revela uma melhoria se comparado com a edição de 2014, onde foi registrada uma fração de 20% de sessões com esta taxa [47]. A melhora desta métrica é muito importante, já que, como será mostrado na Seção 5.3, o congelamento é uma das principais causas para a redução do engajamento. Por outro lado, vale salientar a não extinção total dos eventos de congelamento, mesmo com a implementação da adaptação. Isso pode ter a ver com a natureza de transmissões ao vivo, que possuem um *buffer* de reprodução menor que o de transmissões sob demanda. Essa característica restringe o tempo para os algoritmos de adaptação reagirem a quedas de vazão.

Por fim, as figuras 5.5c e 5.5d mostram o número de adaptações positivas e negativas por minuto. A ocorrência de adaptações indica variações na largura de banda do cliente, isto é, podem ser indício de queda de desempenho de transmissão. Apesar disso, é importante ressaltar que uma adaptação é mais tolerável que um congelamento [103]. Nos dados desta tese, foi observado que cerca de 28% das sessões possui acima de 1 adaptação positiva por minuto e 19% tem 1 adaptação negativa por minuto. Em comparação, a edição de 2014 [47] registrou 1 ou mais adaptações positivas para 61% das sessões e 1 negativa ou mais para 28% das sessões. Com isso, nota-se uma evolução na infraestrutura que permitiu alcançar maior estabilidade na taxa de transmissão recebida nos clientes.

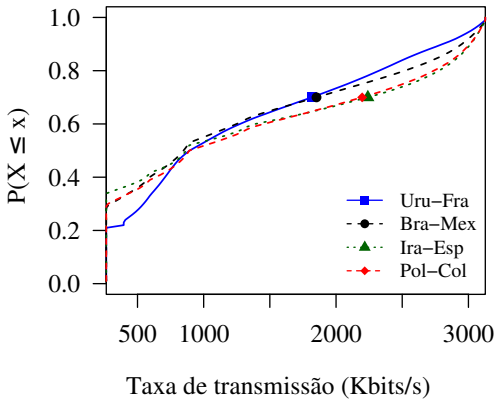
5.2.3 Desempenho de Transmissão em Clientes Móveis

Esta seção apresenta a caracterização das métricas de desempenho de transmissão em dispositivos móveis, começando pela Figura 5.6 que mostra a taxa de transmissão durante as sessões. Já a Figura 5.8 apresenta as demais métricas. Essas figuras incluem as mesmas métricas da seção anterior, calculadas de acordo com a mesma metodologia, assim como a discretização dos eixos, que segue o mesmo padrão.

A Figura 5.6 mostra a taxa de transmissão média das 4 partidas selecionadas para análise. Ao traçar um paralelo com os dispositivos fixos, percebe-se que a taxa de transmissão é menos afetada pela popularidade. Esse fato é reforçado pela Figura 5.7, que mostra uma taxa de transmissão mais similar em todas as partidas. Além disso, ao se calcular o coeficiente de *Spearman* para a taxa de transmissão média no pico de acessos, chegou-se a uma correlação de -0.38, que também indica que os clientes móveis experimentam uma taxa de transmissão mais uniforme, independentemente do número de clientes ativos.

Uma das razões que pode explicar a taxa de transmissão mais homogênea é que clientes móveis naturalmente requisitam segmentos em taxas de transmissão menores, muito em virtude de suas capacidades mais restritas de *hardware* e energia. Uma evidência para essa afirmação é que 60% das sessões móveis possui uma taxa de até 1294 kbps, enquanto que para sessões fixas esse valor é de 1840 kbps.

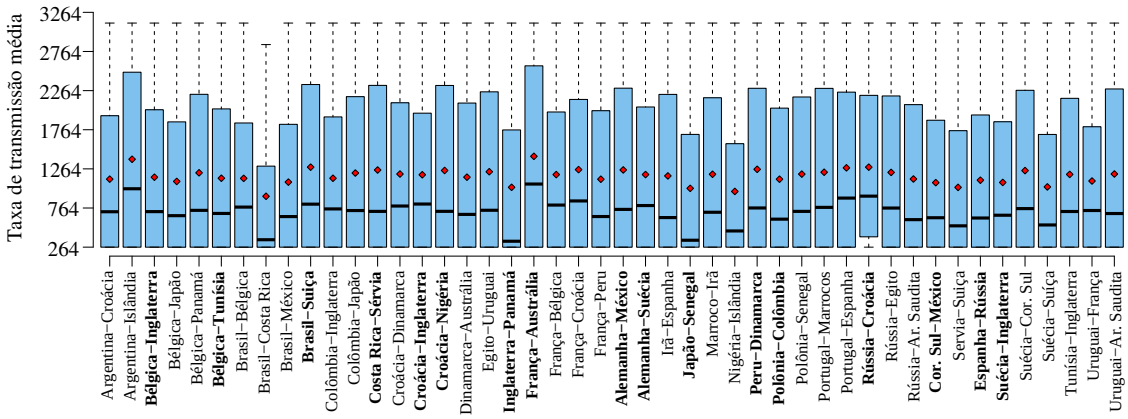
Figura 5.6: Taxa transmissão média



Fonte: Elaborado pelo autor.

A Figura 5.8 dá continuidade a análise apresentando as demais métricas de desempenho de transmissão para dispositivos móveis. Para a latência de inicialização (Figura 5.8a), foi possível observar um desempenho melhor que os dos clientes fixos. Mais especificamente, 60% das sessões móveis apresentaram uma latência de até 6 segundos contra 10 segundos em clientes fixos. Essa diferença pode ser devido à preferência por

Figura 5.7: Distribuição da tx. de transmissão média para clientes móveis.



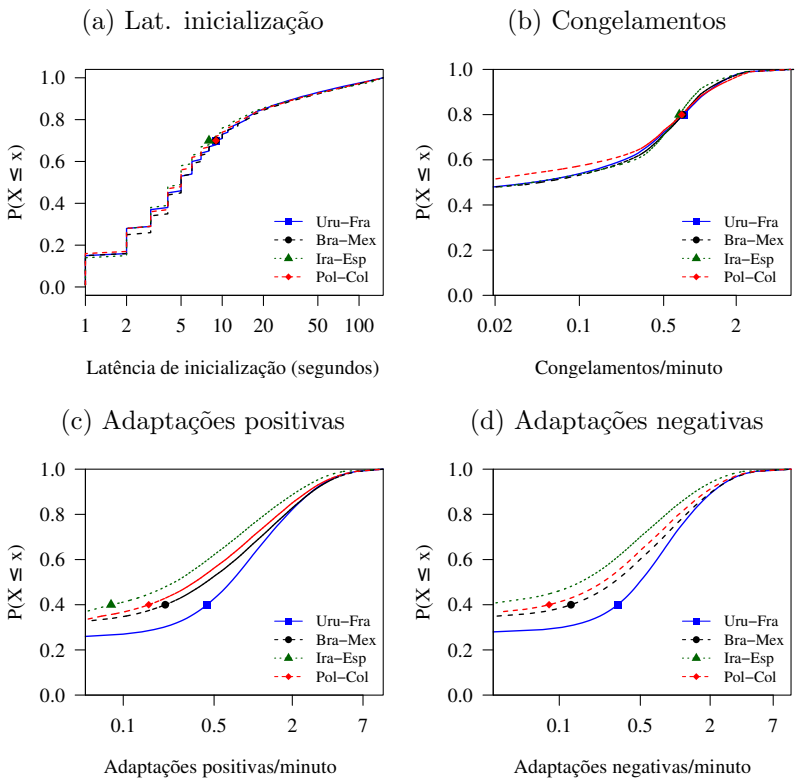
Fonte: Elaborado pelo autor.

segmentos de menor taxa de transmissão. Esse comportamento reduz o tempo de preenchimento do *buffer*. Já no caso da comparação desta métrica entre as partidas, foram constatados valores similares.

A Figura 5.8b mostra a taxa de congelamentos para os dados considerados. Para essa métrica, foi observado que os clientes móveis experimentaram um desempenho inferior, com 60% das sessões móveis apresentando 0.26 congelamentos por minuto contra 0.02 em sessões de clientes fixos. Não é possível determinar com precisão as razões para essa diferença entre dispositivos, mas aspectos como conexões móveis de menor estabilidade e algoritmos de adaptação simplificados, mais adequados ao hardware de dispositivos móveis, podem ter contribuído para uma maior ocorrência de congelamentos. Já na comparação entre partidas, a diferença no desempenho dessa métrica também é similar.

As Figuras 5.8c e 5.8d apresentam a taxa de adaptações positivas e negativas, respectivamente. Comparando estas figuras com as apresentadas para clientes fixos, nota-se que, no caso dos clientes móveis, houve uma diferença mais acentuada das taxas entre as partidas. No jogo Uruguai vs. França, a partida de maior audiência, foi registrado até 0.95 adaptações negativas para 70% das sessões. Já na partida Irã vs. Espanha, para a mesma parcela de clientes, o valor foi de 0.49 ou menos. Esse resultado sugere que transmissões com cargas de trabalho mais altas registraram uma adaptação mais frequente. No entanto, o cálculo da correlação de *spearman* entre a quantidade de sessões ativas e taxa de adaptações não revelou um valor significativo. Comparando o desempenho desta métrica entre clientes fixos e móveis, constatou-se que os clientes móveis registraram uma quantidade ligeiramente maior de adaptações em suas sessões, o que pode ser reflexo de uma conexão sujeita a mais instabilidades.

Figura 5.8: Métricas de Desempenho de Transmissão em clientes móveis.



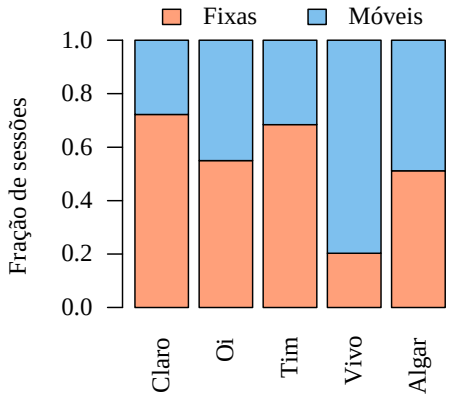
Fonte: Elaborado pelo autor.

5.2.4 Taxa de Transmissão em Provedores

Conforme mostra a Seção 5.1, foi observado um alto ganho de informação em métricas de desempenho de transmissão e engajamento produzido pela correlação com o provedor de conectividade do cliente. Com base nessa constatação, foi empregada uma investigação mais sistemática de como se dá a diferença de desempenho de transmissão nos clientes dos provedores utilizados pelos usuários. Foram selecionados os 5 provedores mais populares em número de sessões. São eles *Vivo*, *Claro*, *Oi*, *Tim*, e *Algar*, que geraram 32%, 21%, 10%, 5%, 2% de sessões, respectivamente. Os provedores de Internet foram analisados separadamente, sempre considerando todas as partidas contidas nos dados.

Em primeiro lugar, são mostradas as distribuições de tipos de dispositivos nos provedores escolhidos. A Figura 5.9 mostra que a fração de sessões para cada tipo de dispositivo. É possível observar uma grande diversidade de distribuições entre os provedores. *Claro* e *Tim* têm mais clientes fixos, com 72% e 68% de sessões deste tipo, respectivamente. Apesar disso, sabe-se que o provedor *Tim* em particular, é especializado em soluções de conectividade móvel [6]. Já o provedor *Vivo* tem quase 80% de clientes móveis. Por fim, *Oi* e *Algar* possuem frações similares de ambos os tipos de dispositivos. Vale ressaltar que no caso do provedor *Algar*, a probabilidade de um cliente, seja fixo ou

Figura 5.9: Fração de sessões de cada dispositivo por provedor de conectividade



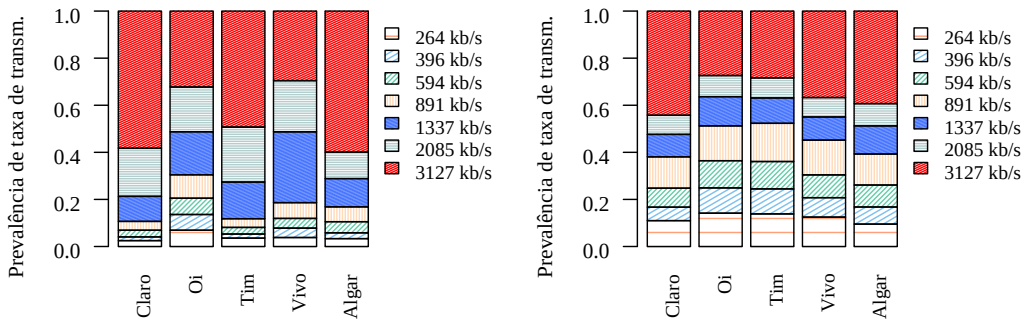
Fonte: Elaborado pelo autor.

móvel, ter usado uma conexão móvel é menor, tendo em vista que a *Algar*, na época da transmissão do evento considerado, era especializada principalmente no provimento de conexões residenciais fixas de banda larga [6, 7].

Em relação ao desempenho de transmissão em cada provedor, a Figura 5.10 mostra a fração de segmentos requisitados de cada taxa disponível, tanto para dispositivos móveis quanto para fixos. A Figura 5.10a mostra uma significativa diferença nessas distribuições no caso dos dispositivos fixos. Por exemplo, a fração de segmentos da maior taxa é 30% e 32% para *Vivo* e *Oi*, respectivamente. Por outro lado, a mesma fração chega a 49%, 58% e 60% para *Tim*, *Claro* e *Algar*, respectivamente. Nota-se também que, para as sessões servidas pelo provedor *Oi*, aproximadamente 20% dos segmentos têm taxa de 594 kbps ou menos, o que é insuficiente para para garantir resolução otimizada para telas de alta definição. Por meio desses números, notam-se possíveis gargalos de desempenho que impactam na taxa de transmissão para clientes de alguns provedores.

Figura 5.10: Distribuição de taxa de transmissão por provedor de conectividade

(a) Sessões fixas (b) Sessões móveis



Fonte: Elaborado pelo autor.

As sessões móveis, por sua vez, apresentam diferenças menores de desempenho

entre provedores (Figura 5.10b). Nesse sentido, o provedor *Claro* apresentou a maior fração de segmentos da maior taxa de transmissão. Também nota-se de forma geral uma fração maior de segmentos das taxas mais baixas (891 kbps ou menos) se comparado com sessões fixas. Isso pode ser devido tanto ao uso de conexões móveis, que são mais lentas se comparado às conexões residenciais, quanto às limitações de *hardware* que naturalmente restringem a taxa de transmissão a um patamar inferior ao de clientes fixos, que dispõem de telas maiores.

5.3 Impacto do Desempenho no Engajamento

A literatura tem atestado que o engajamento tem correlação com as métricas de desempenho de transmissão. Nesse sentido, destaca-se como principal fator de redução de engajamento os eventos relacionados com o *buffer* do cliente, particularmente a latência de inicialização e os congelamentos [27, 72]. No entanto, com o surgimento das transmissões adaptativas, nasceu também um novo tipo evento dentro de uma sessão de vídeo: a troca de taxa de transmissão ou adaptação.

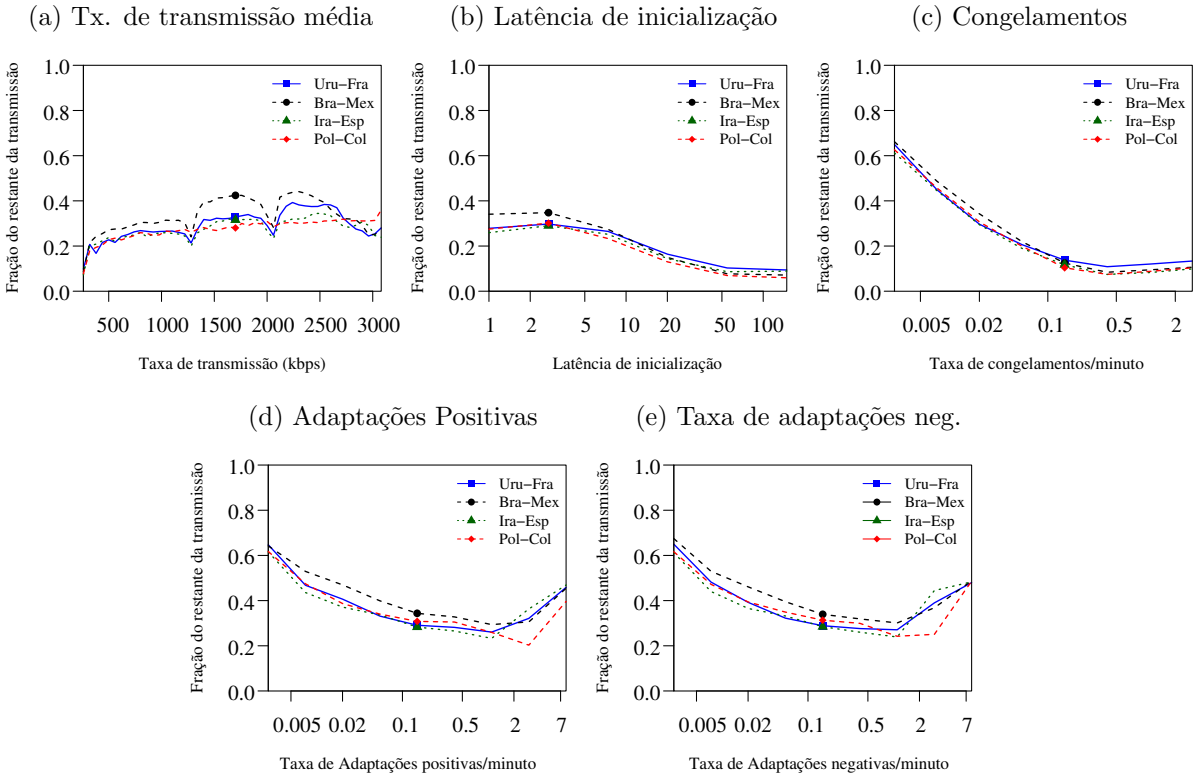
O impacto da adaptação no engajamento ainda é pouco conhecido se comparado com o que já se sabe sobre os congelamentos, sobretudo em transmissões ao vivo em larga escala, que se tornaram populares nos últimos anos. Nesse sentido, o principal problema que dificulta o estudo da adaptação é a falta de conjuntos de dados que permitam determinar a taxa de transmissão de cada segmento requisitado. Os dados utilizados nesta tese possuem essa informação e, por isso, propiciam um estudo mais sistemático desse fator.

Com base no exposto, esta seção apresenta um estudo do impacto da variação das métricas de desempenho de Transmissão no engajamento de usuários. Seu objetivo é estudar tanto as métricas relacionadas ao *buffer* do cliente, quanto também o impacto da adaptação que, como dito, é menos conhecido. Foram utilizadas transmissões das partidas que englobam todos os tipos de carga observada, tanto em termos de regime de chegadas, quanto número de clientes.

5.3.1 Clientes Fixos

As figuras desta seção e da seção seguinte apresentam a correlação entre as métricas de desempenho, mostradas no eixo x de cada figura, e o engajamento normalizado pelo tempo restante, conforme explicado na Seção 4.4.1.

Figura 5.11: Impacto de métricas de desempenho no tempo de sessão (clientes fixos)



Fonte: Elaborado pelo autor.

Os dados das transmissões mostram uma clara correlação não-linear entre as métricas de desempenho de transmissão e o engajamento. A Figura 5.11a mostra que a fração assistida aumenta quando os clientes recebem uma taxa de transmissão média próxima da máxima oferecida pelo provedor. Como mostrado na figura, foi registrado um engajamento abaixo de 20% para sessões com a menor taxa. Sessões com a maior taxa de transmissão, por outro lado, podem atingir uma fração de tempo assistido que pode chegar a 40% do tempo restante da transmissão.

Impactos similares podem ser vistos para as outras métricas, em diferentes intensidades. A Figura 5.11b apresenta a relação entre a latência de inicialização e o engajamento. Usuários em sessões com longa latência deixam a transmissão mais cedo (na média) se comparado àqueles cujas sessões têm um início mais rápido. Como exemplo, observa-se uma acentuação da queda de engajamento a partir de 7 segundos de latência

inicialização.

Correlações mais evidentes (negativas) podem ser observadas para as métricas restantes. Por exemplo, na figura 5.11c, é possível notar como a duração diminui com o aumento da taxa de congelamentos. A figura é interrompida em 2 congelamentos por minuto porque valores maiores ocorrem mais raramente. É evidenciada a baixa tolerância do usuário a congelamentos, visto que se observa um engajamento médio de mais de 60% para sessões sem congelamento contra menos de 20% em sessões com 0, 1 congelamento por minuto. Vale ressaltar que também foram registradas sessões sem congelamento mas com curta duração. Neste caso, o abandono pode ter outras razões como falta de interesse no conteúdo. Esse tipo de sessão foi removida para não enviesar os resultados apresentados.

A taxa de adaptações também tem forte correlação com a duração da sessão (Figura 5.11d e 5.11e). É possível observar uma significativa queda no engajamento quando são comparadas sessões com duas adaptações por minuto e sessões com uma taxa menor de adaptações. A estreita relação entre congelamentos e adaptações pode ser uma explicação para o resultado. Isto é, usuários com uma alta taxa de adaptações têm maior probabilidade de estarem sob condições ruins de rede, o que pode levar a congelamentos.

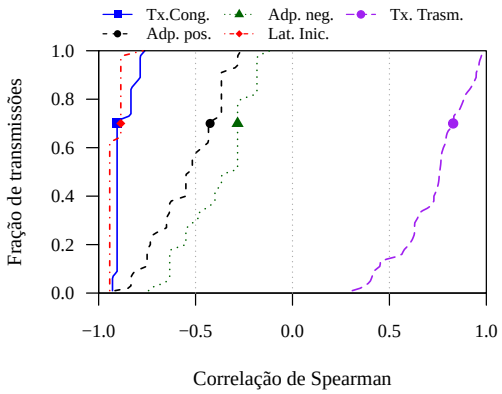
Além disso, como é possível observar, foi constatado um aumento na duração das sessões que estão sinalizadas com 5 a 10 adaptações por minuto. Na verdade, esse fato ocorre porque a tolerância a adaptações não é a mesma em cada usuário e depende da relação com outras métricas. Por exemplo, usuários em clientes com uma baixa taxa de transmissão – que pode ser evidência de conectividade ruim – são mais tolerantes às adaptações. Na Seção 5.4, foram mapeadas essas relações por meio da construção de cenários integrados de desempenho. Esse processo permite uma análise mais precisa do engajamento, considerando o efeito da co-dependência entre métricas de engajamento. Será visto que a variação dos valores das métricas impactam de formas diferentes o engajamento de acordo com cenários específicos de desempenho.

Os coeficientes de correlação de *Spearman* corroboram a intuição visual obtida das figuras. A Figura 5.12 sumariza os coeficientes calculados para os jogos do conjunto de dados para clientes fixos. Cada curva corresponde a uma métrica de desempenho. Quanto mais próxima está a curva das extremidades, maior é a correlação entre a métrica correspondente e o engajamento.

É possível observar que a latência inicial do vídeo e a taxa de congelamentos são os fatores que mais impactam negativamente o engajamento do usuário fixo. Ou seja, é esperado que usuários em clientes fixos tenham baixa tolerância a uma espera inicial muito longa e também por congelamentos ao longo de todo o vídeo. Isso evidencia uma mudança de comportamento, se comparado com a edição anterior do evento [47]. Em 2014, a tolerância à latência de inicialização era significativamente maior. Já em 2018, é esperado que o vídeo seja iniciado de maneira mais imediata.

Ainda em relação à Figura 5.12, é possível observar que as adaptações também

Figura 5.12: Correlação entre as métricas de desempenho e o engajamento (clientes fixos).



Fonte: Elaborado pelo autor.

produzem redução no engajamento. Essa correlação pode variar em cada partida, mas sempre se mantém negativa. Curiosamente, adaptações negativas impactam menos do que adaptações positivas. Uma hipótese para esse resultado é que o crescimento da taxa de transmissão pode levar a um aumento da probabilidade de congelamentos que, como mostrado, exercem grande impacto negativo no engajamento. Uma queda de taxa, por outro lado, pode prevenir um congelamento iminente, sendo então mais tolerada pelo usuário. Comparado com 2014 [47], a tolerância a adaptações aumentou, possivelmente em virtude de ela ser uma forma eficiente de evitar congelamentos.

Por fim é analisada a taxa de transmissão média. Como esperado, ela tem correlação positiva com o engajamento, isto é, sessões com maior taxa têm maior duração. Essa correlação também é variável, mas sempre se mantém positiva podendo chegar a próximo de 1, dependendo da transmissão.

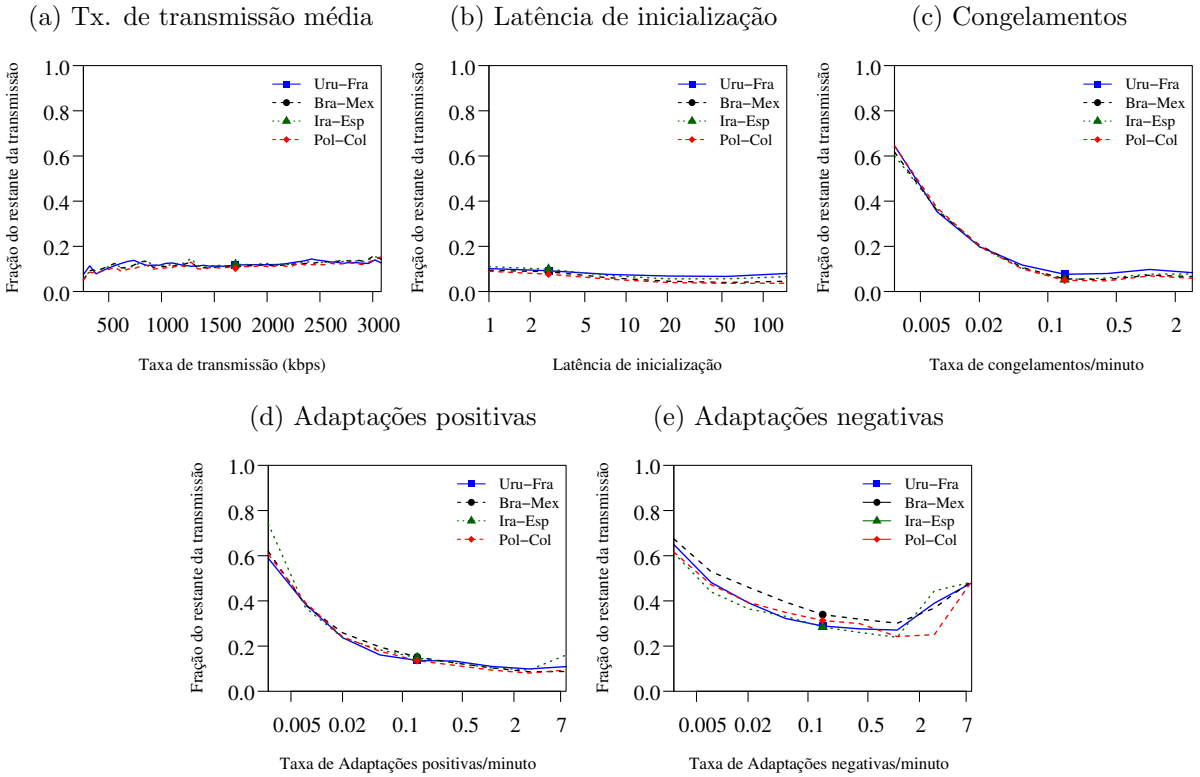
5.3.2 Clientes Móveis

Esta seção dá continuidade a análise se concentrando em sessões de clientes móveis. A Figura 5.13 mostra as correlações das métricas de desempenho de transmissão em sessões de clientes móveis com duração da sessão. As figuras incluem as mesmas métricas da seção anterior, calculadas de acordo com a mesma metodologia. Além disso, a discretização dos eixos segue o mesmo padrão.

A intensidade do impacto do desempenho de transmissão sobre a duração de sessões móveis apresenta algumas diferenças se comparado com clientes fixos. Nota-se o impacto reduzido que tanto a taxa de transmissão quanto a latência de inicialização têm para o engajamento. Isso significa que usuários em clientes móveis podem ter uma tolerância

um pouco maior em relação a uma longa espera para início de uma sessão e são menos exigentes em relação a baixa taxa de transmissão. Essa tolerância pode se dever ao uso de conexões móveis, ou mesmo conexões *wi-fi* públicas, que podem sofrer mais instabilidades comparadas a conexões residenciais de banda larga.

Figura 5.13: Impacto das métricas de desempenho no tempo de sessão (clientes móveis).



Fonte: Elaborado pelo autor.

O cenário se inverte ao analisar a taxa de interrupções. Comparando as figuras 5.11c e 5.13c é possível notar que a queda de engajamento é mais acentuada em dispositivos móveis. Por exemplo, usuários em clientes móveis com $0,1$ congelamentos por minuto possuem engajamento similar ao de usuários em clientes fixos com $0,5$ congelamentos por minuto. Isso demonstra que esses usuários priorizam um fluxo contínuo de exibição em detrimento de uma alta taxa de transmissão.

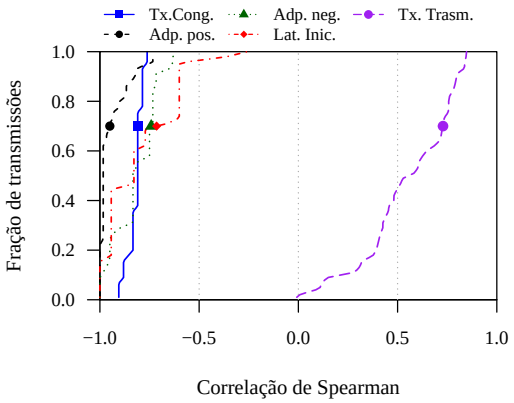
Uma outra diferença pode ser observada na taxa de adaptações. Para clientes fixos, foi visto que a queda de engajamento é similar para adaptações positivas e negativas. Esse resultado é menos evidente em clientes móveis. As Figuras 5.13d e 5.13e mostram que, neste caso, a queda de engajamento por uma adaptação positiva é maior do que a observada em uma adaptação negativa. Assim como para clientes fixos, especula-se que um aumento de taxa eleve a probabilidade de ocorrência de um congelamento que, no caso de clientes móveis, é um fator ainda mais impactante na redução de engajamento.

A Figura 5.14 apresenta os coeficientes de *Spearman* para a correlação entre engajamento e as métricas de desempenho de transmissão. A figura corrobora as observações

já apresentadas. É possível notar a menor tolerância de usuários móveis em relação a adaptações e, em particular, a adaptações positivas. Como dito, o aumento da taxa beneficia o engajamento, mas pode levar a um aumento de congelamentos. Reitera-se que este problema afeta de modo especial as conexões móveis, que possuem uma largura de banda mais variável.

Considerando a latência de inicialização e congelamentos, nota-se que seus valores estão em patamares similares aos dos dispositivos fixos. No entanto, comparado com 2014 [47], foi observado também uma pequena redução na tolerância a latência de inicialização, fruto de uma possível mudança nos requisitos de desempenho dos usuários, que passaram a exigir um início mais rápido da exibição da mídia, assim como em usuários em clientes fixos. Por fim, no caso da taxa de transmissão, observou-se valores de correlação menores, confirmando que a taxa de transmissão tem menos impacto no engajamento de usuários móveis.

Figura 5.14: Correlação entre as métricas de desempenho e engajamento (clientes móveis)



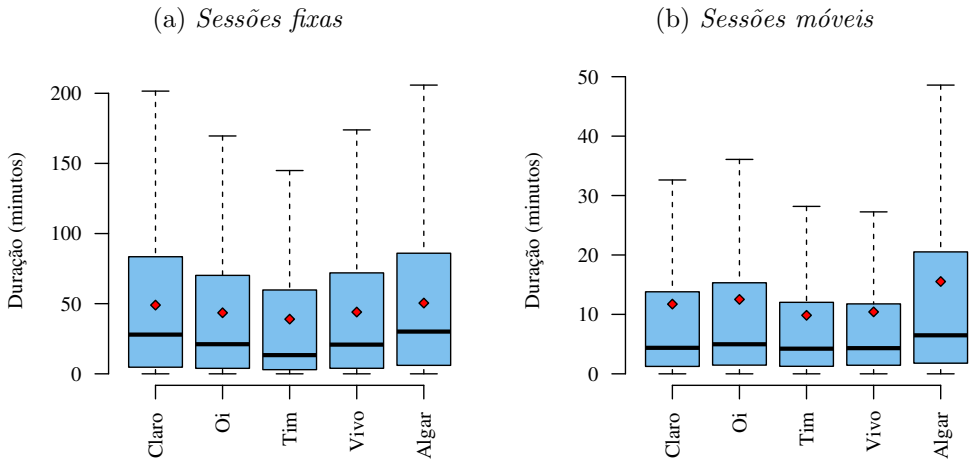
Fonte: Elaborado pelo autor.

5.3.3 Provedores de Conexão com a Internet

Na Seção 5.2.4 foram apresentadas evidências de que o engajamento de um usuário varia consistentemente dependendo do provedor de conectividade considerado. Esta seção caracteriza tal variação. A Figura 5.15 inicia a análise apresentando as distribuições de duração de sessão para dispositivos fixos e móveis nos mesmos provedores mostrados na Seção 5.2.4

A Figura 5.15a mostra que, no caso dos dispositivos fixos, foi observado que os provedores que ofereceram maior taxa de transmissão – *Algar* e *Claro* –também registraram as sessões mais duradouras, que alcançam até 100 minutos para 80% das sessões.

Figura 5.15: Distribuição da duração de sessão em cada provedor de conectividade



Fonte: Elaborado pelo autor.

Essa observação sugere, como esperado, mas uma evidência de correlação positiva entre o engajamento do usuário e as métricas de desempenho de transmissão no provedor.

Apesar do achado apresentar um indício da correlação positiva entre desempenho e engajamento, vale salientar que a taxa de transmissão sozinha não explica o engajamento. Isso é claro no caso da *Tim*, que ofereceu altas taxas de transmissão, comparáveis às dos provedores *Claro* e *Algar*, mas obteve as sessões com o menor tempo de duração. Uma das razões é que esse provedor oferece conexões móveis como principal produto. Como mencionado, tais conexões em geral possuem franquias de dados limitadas, que restringe a duração das sessões. De fato, medidas dão conta de que o serviço de banda larga residencial da Tim estava presente em apenas 0.7% das cidades brasileiras [6]. Esse fato é uma evidência do impacto de um fator contextual, no caso o tipo de conexão, no engajamento de usuários.

A Figura 5.15b apresenta a distribuição de durações para sessões móveis. Como esperado, clientes móveis possuem sessões mais curtas do que as de clientes fixos, possivelmente por conta de suas limitações de bateria e conexão. Outra hipótese para as menores sessões, levantada por [80], é de que as interfaces de sistemas móveis dificultam a utilização de mais de um aplicativo ao mesmo tempo, o que força os usuários a saírem da transmissão quando eles desejam acessar outros aplicativos.

Também é possível notar a similaridade nas distribuições das durações entre os cinco provedores, algo oposto ao encontrado para dispositivos fixos. A exceção é o provedor *Algar*, cuja distribuição apresenta sessões mais longas. Esse é mais um caso no qual as particularidades do provedor exercem um papel importante, pois 98% dos clientes do provedor *Algar* usam conexões de banda larga residenciais [7, 6], a despeito do tipo de dispositivo. Em contraste com o cenário apresentado para o provedor *Tim*, neste caso há uma maior probabilidade de os clientes terem se conectado via *Wi-Fi* particular, onde

não existem a limitações e franquias presentes nas conexões móveis. Como resultado, esse cenário estimula uma permanência mais longa do cliente.

5.4 Co-dependência entre Métricas de Desempenho

Este capítulo mostrou que o engajamento está ligado a variações nas métricas de desempenho de transmissão. No entanto, a avaliação individual dessas métricas pode não ser capaz de explicar integralmente essa relação. Isso fica claro quando se observa, por exemplo, a taxa de adaptações em clientes fixos (figuras 5.11d e 5.11e). Essas figuras mostram uma redução de engajamento em frações até 2 adaptações por minuto e um posterior crescimento a partir dessa taxa, o que é contra intuitivo, uma vez que é esperada uma queda constante de engajamento com o aumento de adaptações.

Uma hipótese para esse comportamento é de que os usuários percebem de formas diferentes uma variação em uma métrica de desempenho de transmissão e que essa percepção está ligada a uma visão mais ampla de desempenho de transmissão, que considera a co-dependência com outras métricas. Em outras palavras, não se pode medir o impacto de uma métrica particular no engajamento sem integrar nesse impacto a influência de outras métricas de desempenho de transmissão.

Com o objetivo de integrar o efeito de múltiplas métricas na variação de engajamento, foi desenvolvida a noção de Cenários de Desempenho. Um cenário de Desempenho é uma determinada configuração de valores das métricas de desempenho de transmissão. Esses cenários foram rotulados de acordo com nível de desempenho produzido, isto é, baixo, médio e alto.

Em cada cenário, é possível variar uma métrica particular e observar o impacto dessa variação no engajamento. Com isso, o objetivo é investigar se um usuário, com seu cliente inserido em determinado cenário, reage a uma variação em uma métrica particular da mesma forma que usuários com clientes em outros cenários de desempenho. O engajamento de um usuário em cada cenário de desempenho é aprendido por meio de um algoritmo de regressão, conforme será descrito na Seção 5.4.3.

5.4.1 Construção dos Cenários de Desempenho

Para a construção dos cenários de desempenho, foi utilizado o algoritmo de agrupamento *k-means* [36]. O *k-means* usa distância euclidiana entre as sessões num espaço n-dimensional. Com isso, espera-se que sessões próximas neste espaço tenham valores similares nas métricas de desempenho de transmissão. Cada grupo de sessões similares dá origem a um cenário de desempenho, que é expresso pelo representante médio ou centróide do grupo.

Para garantir a mesma contribuição de todos os fatores, foi aplicada uma normalização para o intervalo $[0 - 1]$ utilizando a normalização *max-min*[2]. Já os fatores categóricos (e.g. *dispositivo*) foram convertidos em variáveis binárias. Em uma variável binária, sessões de uma mesma categoria têm distância mínima (i.e., valores iguais) e aquelas com categorias distintas terão distância máxima (i.e., valores 0 e 1).

Os atributos de agrupamento são originados de um subconjunto das métricas de desempenho descritas no Capítulo 4, composto pela *taxa de congelamentos e adaptações, tanto positivas quanto negativas*, a *latência inicial* e o *dispositivo*, além da *fração de permanência em cada taxa de transmissão*.

Um aspecto crítico para o *k-means* é a definição de seu parâmetro k, i.e., a quantidade de grupos. Existem diversas maneiras de estimar o valor de k. Neste trabalho, executou-se o algoritmo com k variando de 2 a 30. Para cada instância de execução, foi avaliado o erro de agrupamento, que consiste na soma das distâncias de todos os elementos para seus centroides. Em geral, o incremento de k reduz esse erro. No entanto, a redução se torna marginal a partir de certo k e o custo do incremento passa a ser maior do que redução do erro [117]. O k é então escolhido neste ponto. A escolha do limite de 30 centroides é conservadora, sendo que em geral a estabilização da redução de erro se deu abaixo de $k = 10$ para todas as tarefas de agrupamento de todos os testes efetuados.

5.4.2 Caracterização dos Cenários de Desempenho

O algoritmo de agrupamento identificou a existência de 6 cenários de desempenho. A caracterização dos centroides de cada um desses cenários está presente na Tabela 5.2. A descrição dos centroides é importante para entender como é o engajamento médio em cada cenário. Como a tabela mostra, os cenários estão rotulados por níveis crescentes de

[2] $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$

desempenho/qualidade – *baixo*, *médio* e *alto* – sendo 3 cenários para clientes móveis e 3 para clientes fixos.

Tabela 5.2: Descrição dos cenários de desempenho a partir dos centróides

Cenário	Lat. inicialização	Congelamentos	Adaptações +	Adaptações -	Tx. Transm. média	Engajamento
Dispositivos fixos						
Baixo (BQ)	66,73 s.	0.88/min.	0,18/min.	0.12/min.	304 kbps	3%
Médio (MQ)	51,11 s.	0.78/min.	1,33/min.	0.93/min.	1478 kbps	20%
Alto (AQ)	9,67 s.	0.04/min.	0.68/min.	0.43/min.	2880 kbps	20%
Dispositivos móveis						
Baixo (BQ)	43,30 s.	0.66/min.	0,10/min.	0.08/min.	287 kbps	1,6%
Médio (MQ)	8,88 s.	0.20/min.	1,95/min.	1,34/min.	1023 kbps	5,2%
Alto (AQ)	5,11 s.	0.10/min.	1,05/min.	0,66/min.	2647 kbps	7%

Fonte: Elaborado pelo autor.

A Tabela 5.2 reforça a tese de que o engajamento está associado ao desempenho, com o nível *alto* registrando maiores engajamentos (durações médias relativas de sessão). Analisando os outros cenários, nota-se que o nível de desempenho é inversamente proporcional à taxa de congelamento, que alcança seu maior valor no cenário de nível *baixo*. Já a taxa de adaptação alcança seu valor mais elevado no nível de desempenho *médio*. Comparando os cenários de níveis *baixo* e *médio*, percebe-se que uma alta taxa de adaptações é mais tolerável pelo usuário do que um elevado número de congelamentos, tendo em vista que usuários com clientes no nível *médio* engajam-se mais do que os do nível *baixo*.

5.4.3 Relacionando Desempenho e Engajamento por Cenário

Como já mencionado, o objetivo desta seção é avaliar se variações em métricas particulares impactam o engajamento de maneiras distintas, dependendo do cenário de desempenho considerado. Para permitir essa análise, é necessária a utilização de um modelo de regressão. Esse modelo é treinado nas sessões de cada cenário de desempenho para aprender suas relações específicas entre métricas de desempenho de transmissão e engajamento. A seguir, os modelos treinados são invocados nos centróides de cada grupo, com uma métrica de interesse sendo variada dentro da faixa de valores contida nos dados. Como resultado, é possível registrar a evolução do engajamento, em resposta a uma variação de desempenho, levando em conta também o efeito das demais métricas do conjunto.

O modelo utilizado para o aprendizado das relações em cada cenário é conhecido como Máquina de Reforço de Gradiente (*Gradient Boosting Machine – GBM*) [37]. Esse modelo funciona treinando sucessivas árvores de decisão. Cada uma das árvores são

treinadas para minimizar uma função de perda (e.g., o erro quadrático médio) produzido pelas árvores geradas em iterações anteriores. A opção por esse modelo se deu pelo seu bom compromisso entre utilização de recursos computacionais e precisão. Além disso, foi possível atingir precisão superior a de modelos computacionalmente mais custosos como, por exemplo, as florestas aleatórias [54].

5.4.4 Análise da Variação de Desempenho em Cada Cenário

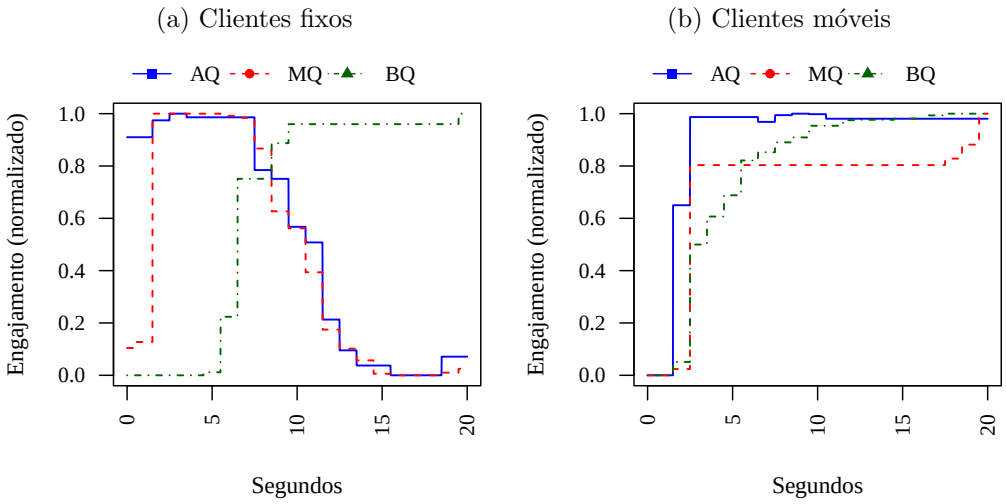
Esta seção mostra como a variação de uma métrica de desempenho individual impacta no engajamento de um usuário, dado um determinado cenário de desempenho. Nos gráficos desta seção, os cenários foram apresentados em conjunto para melhor comparação. Para isso os valores de engajamento foram normalizados pelo tempo máximo respectivo de cada cenário de desempenho.

Latência de Inicialização

A Figura 5.16a apresenta o impacto da variação da latência de inicialização no engajamento dos usuários de clientes fixos. É possível notar como o engajamento evolui de formas distintas dependendo do cenário de desempenho considerado, sendo que aqueles com nível médio e alto de desempenho esperam início imediato da reprodução, uma vez que seu engajamento cai com o aumento da latência. Um efeito inverso é observado em clientes com nível de desempenho baixo. Seus usuários percebem uma longa latência de inicialização de forma positiva, ao contrário dos demais cenários. Com isso, tamanhos de *buffer* reduzidos podem ser usados em clientes com alta vazão/desempenho, para que seu requisito de reprodução imediata seja atendido, em contraste com clientes em cenário de baixo desempenho, que podem, por exemplo, admitir *buffers* maiores e carregamento mais demorado.

A Figura 5.16b mostra a evolução da latência de inicialização, agora no caso de clientes móveis. Nota-se que latências longas não afetam negativamente o engajamento dos usuários da mesma maneira que nos clientes fixos. Uma hipótese é a natureza mais instável das conexões móveis, que pode fazer com que os usuários sejam mais tolerantes a longas esperas de inicialização. Ainda assim, parece haver uma diferença de tolerância em cada cenário. Clientes de nível baixo e médio toleram latências mais longas – 13 e 20 segundos, respectivamente – para alcançar o engajamento máximo se comparado a clientes com nível de desempenho alto, que alcançam o máximo engajamento do grupo com 3 segundos.

Figura 5.16: Latência de inicialização versus engajamento

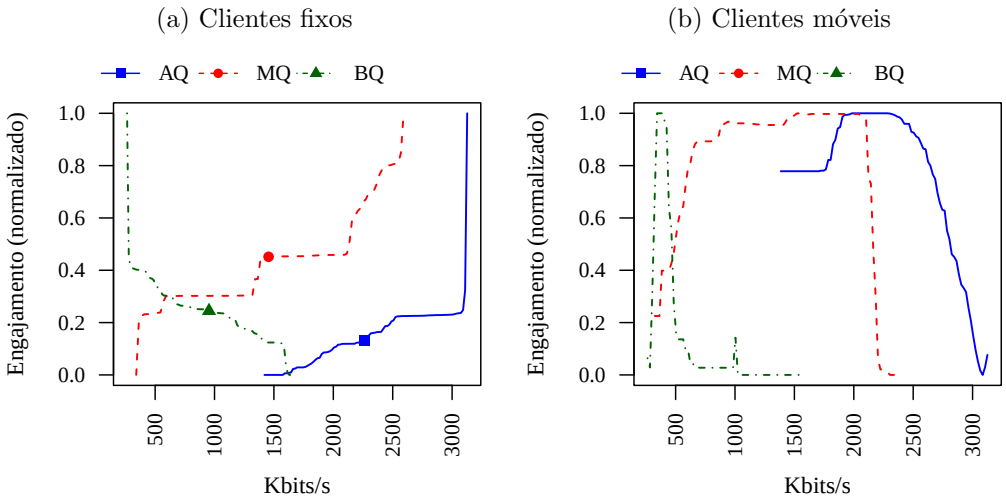


Fonte: Elaborado pelo autor.

Taxa de Transmissão

Continuando a análise, a Figura 5.17a mostra a relação entre a taxa de transmissão e o engajamento para clientes fixos. É possível observar que usuários em clientes com nível de desempenho baixo perdem engajamento com o aumento da taxa de transmissão. Isso revela que, ao contrário do que se acreditava [103], um aumento de taxa de transmissão nem sempre é desejável, tendo em vista que esse aumento pode prejudicar o desempenho das outras métricas. Já os demais clientes experimentam um crescimento de engajamento proporcional ao aumento da taxa de transmissão. Nesse sentido, os usuários em clientes de alto desempenho são os mais exigentes, uma vez que foi observado um engajamento de no máximo 20% em taxas diferentes da máxima possível.

Figura 5.17: Taxa de transmissão média versus engajamento



Fonte: Elaborado pelo autor.

A Figura 5.17b mostra a evolução do engajamento versus a taxa de transmissão para clientes móveis. É possível notar que o comportamento dos usuários desses clientes difere de seus equivalentes em dispositivos fixos. Nestes clientes, o máximo engajamento foi registrado para uma faixa de valores, ao invés de apenas o valor máximo de cada cenário. Por exemplo, clientes com alto nível de desempenho alcançam engajamento máximo entre 2000 e 2500 kbps. Outra diferença marcante é que o engajamento é mínimo na taxa máxima de transmissão de cada cenário, evidenciando que a conectividade desses clientes não dá suporte às mesmas taxas disponíveis para clientes fixos.

Em suma, é possível notar que diferentes cenários de desempenho induzem formas distintas de reação em relação a mudanças na taxa de transmissão. Essas diferenças podem ser exploradas para a criação de regras de alocação que possam atingir um melhor compromisso entre o engajamento dos usuários e utilização dos recursos disponíveis, uma vez que muitos usuários engajam-se mesmo em taxas relativamente baixas de transmissão.

Taxa de Adaptações Positivas

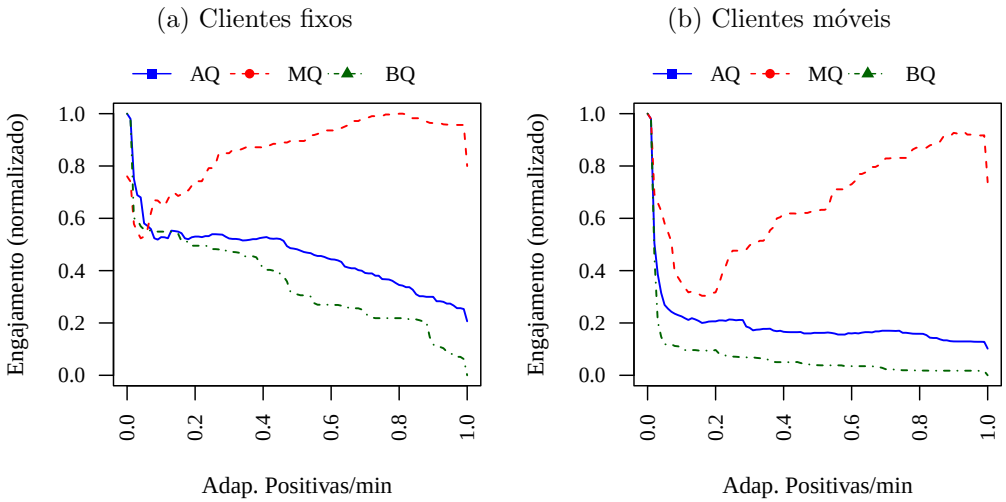
A Figura 5.18 mostra o impacto da variação da taxa de adaptação positiva no engajamento. No caso dessa métrica, foi observado que usuários com clientes em cenários similares seguiram tendências de reação parecidas, independentemente do dispositivo. Usuários com clientes nos níveis de desempenho baixo e alto reagem negativamente ao aumento de adaptações, enquanto usuários em clientes com nível médio experimentam o efeito contrário. Essa é mais uma evidência da influência do cenário de desempenho na percepção de variação de uma métrica. Isto é, usuários em clientes com nível de desempenho alto possivelmente esperam que a mídia não sofra variação de taxas ao longo de sua reprodução ou não desejam que a taxa ultrapasse a sua largura de banda (no caso de clientes móveis). Já usuários em clientes com nível baixo não desejam um aumento de taxa porque a vazão disponível em suas conexões de Internet pode não suportar tais aumentos.

Vale ressaltar também a intensidade da queda de engajamento entre diferentes dispositivos. Clientes móveis perderam mais rápido o engajamento que clientes fixos. Isso é nítido ao comparar as Figuras 5.18a e 5.18b. Esse resultado está em consonância com o observado na análise geral da Seção 5.3, que indica uma correlação negativa maior desta métrica com engajamento para clientes móveis.

Taxa de Adaptações Negativas

A Figura 5.19a apresenta a relação entre adaptações negativas e engajamento. É possível observar que, neste caso, o aumento de adaptações negativas também é tolerado por usuários em clientes com nível alto de desempenho, em adição aos usuários em clientes de nível médio, que já aceitavam o crescimento de adaptações positivas. Esse comportamento

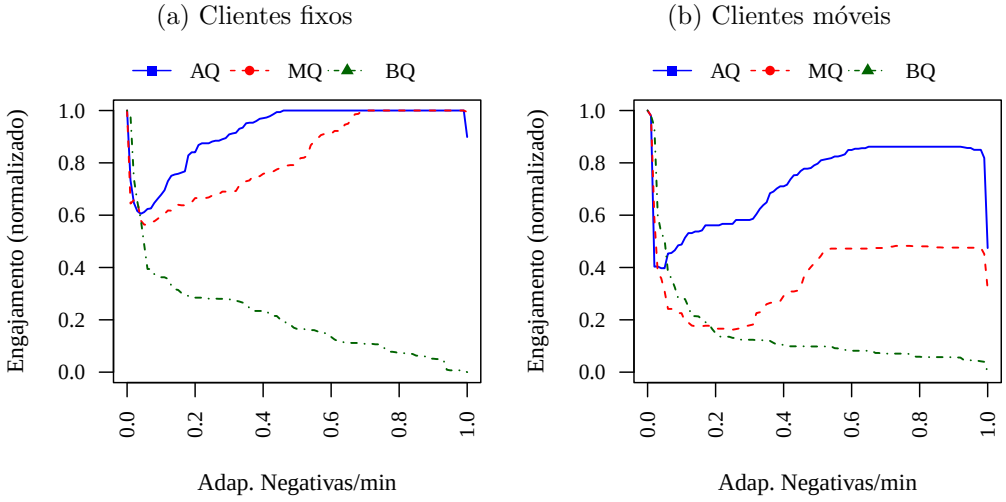
Figura 5.18: Taxa de adaptações positivas versus engajamento



Fonte: Elaborado pelo autor.

pode se dever ao fato de que a redução na taxa de transmissão tem o potencial de reduzir interrupções. No entanto, vale ressaltar que não é uma relação monotônica, uma vez que há uma queda de engajamento próximo do valor de zero adaptações por minuto, indicando que o engajamento é alcançado na ausência de adaptações ou numa quantidade mínima que a partir da qual é possível preservar o *buffer* e evitar congelamentos. Por exemplo, clientes fixos de nível de desempenho alto registraram engajamento máximo com zero adaptações negativas ou com mais de $0,5$ adaptações por minuto. Esse mesmo tipo de fenômeno também foi registrado para adaptações positivas em clientes com nível médio.

Figura 5.19: Taxa de adaptações negativas versus engajamento



Fonte: Elaborado pelo autor.

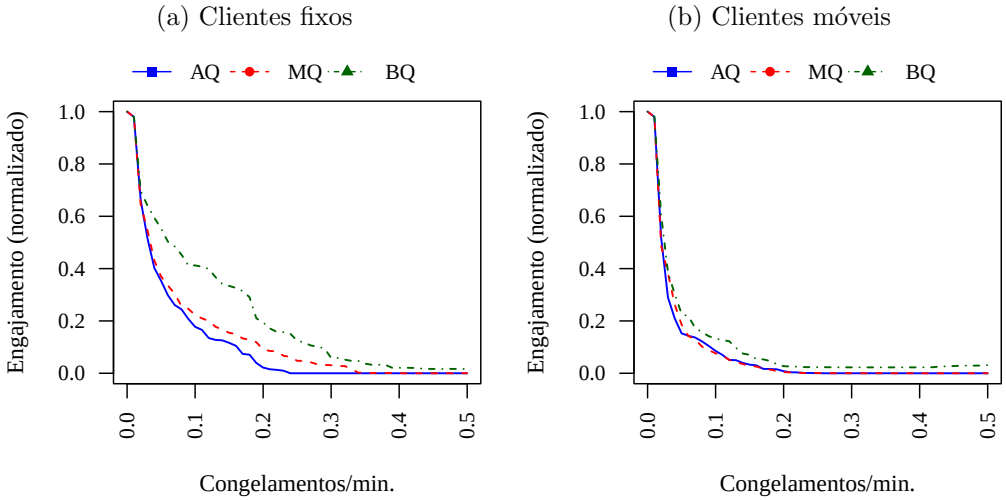
A Figura 5.19 também pode fornecer uma explicação para o aumento de engajamento mostrado na Figura 5.13d a partir de 2 adaptações por minuto. Observando as

duas figuras, é possível concluir que esse aumento de engajamento ocorre apenas para os grupos de usuários submetidos a um cenário de desempenho médio e alto.

Taxa de Congelamentos

Para finalizar o estudo, a Figura 5.20 mostra a relação entre taxa de congelamentos e engajamento. Esta figura mostra que o congelamento é um evento de grande impacto negativo e sua interferência é similar nos cenários, com tolerância um pouco maior em usuários de clientes fixos. Vale ressaltar também que usuários em clientes fixos com baixo nível de desempenho registraram uma tolerância ainda maior à congelamentos. Essa tolerância pode estar relacionada à percepção de baixo desempenho geral por parte do usuário, o que o estimula a aguardar mais tempo pela retomada da exibição do vídeo.

Figura 5.20: Taxa de congelamentos versus engajamento



Fonte: Elaborado pelo autor.

5.4.5 Resumo das Observações

Com base nas relações entre engajamento e desempenho ilustradas na Seção 5.4.4, é possível constatar que diferentes cenários de desempenho podem induzir diferentes requisitos em cada métrica. Com essa constatação em mente, a Tabela 5.3 apresenta um resumo das condições gerais nas quais cada cenário alcança engajamento máximo.

Essa tabela pode ser usada como ponto de partida em novas transmissões, junto com o monitoramento de desempenho de sessões, para refinar as regras de alocação e melhorar o compromisso entre engajamento médio e gasto de recursos. Porém, mais

Tabela 5.3: Diretrizes para estimular engajamento em cada cenário de desempenho

Dispositivos fixos	
Baixo Desempenho (BQ)	Latência de inicialização pode ser alta, baixa taxa de transmissão (500 kbps ou menos). Evitar adaptações.
Médio Desempenho (MQ)	Latência de inicialização pode ser alta, taxa de transmissão entre 1000 e 2000 kbps. Maior liberdade de variação de taxa de transmissão.
Alto Desempenho (AQ)	Latência de inicialização pode ser alta, taxa de transmissão entre 2000 e 2500 kbps. Manter taxa até 2500 kbps.
Dispositivos móveis	
Baixo Desempenho (BQ)	Latência de inicialização pode ser alta, baixa taxa de transmissão (500 kbps ou menos). Evitar adaptações.
Médio Desempenho (MQ)	Latência de inicialização deve ser pequena (<5s.), Taxa de transmissão próxima de 2000 kbps. Maior liberdade de variação de taxa.
Alto Desempenho (AQ)	Latência de inicialização deve ser pequena (<5s.), Taxa de transmissão próxima de 3127 kbps. Usar adaptação em caso de congelamento iminente.

Fonte: Elaborado pelo autor.

importante do que os resultados mostrados, é também o processo de descoberta e análise dos cenários de desempenho, que pode ser adaptado para qualquer tipo de conteúdo e estrutura de transmissão adaptativa.

5.5 Sumário das Contribuições

Este capítulo apresentou uma caracterização de métricas de desempenho de transmissão em um conjunto de vídeos ao vivo em larga escala. Também foi mostrado o impacto desse desempenho no engajamento de usuários. Acredita-se que esse estudo atualiza o entendimento do nível de desempenho proporcionado por infraestruturas de transmissão atuais e pode ser usado para melhorar o desempenho de transmissão em eventos de larga escala.

Mais especificamente, o capítulo apresenta, na Seção 5.1, um estudo que visa descobrir quais fatores contextuais mais impactam em métricas de desempenho de transmissão e engajamento. Foi constatado que o tipo de dispositivo, período da transmissão e provedor de conectividade interferem tanto no desempenho de transmissão quanto no engajamento.

Na Seção 5.2, por sua vez, foi apresentado o panorama geral de desempenho de transmissão para clientes fixos e móveis, onde foi possível também observar o papel da escala no desempenho de transmissão e o desempenho a nível de provedor de conectividade.

Já na Seção 5.3 foi detalhado um estudo do impacto do desempenho de transmissão para o engajamento. Foi mostrado que as perdas de desempenho levam a uma redução

na duração de sessões. Nesse sentido, vale destacar o estudo da influência das métricas de adaptação para o engajamento, que é um aspecto pouco explorado na literatura.

Por fim, na Seção 5.4, é proposta uma ampliação desse estudo, demonstrando que integrar os efeitos de todas as métricas de desempenho, sob a forma de cenários, pode levar a novas descobertas quanto a relação entre desempenho e engajamento. O estudo demonstrou que certas variações de desempenho são toleradas em cenários específicos. Por exemplo, mostrou-se que um aumento da taxa de transmissão não é desejável se o cenário for de baixo desempenho.

O próximo capítulo utiliza os conhecimentos aprendidos neste estudo para orientar a construção de um modelo de comportamento de clientes. Esse modelo foi projetado para considerar a influência do desempenho de transmissão no comportamento dos clientes e o impacto das trocas de taxa de transmissão na quantidade carga de trabalho gerada.

Capítulo 6

Modelando o Comportamento de Clientes de Vídeo Adaptativo ao Vivo

Este capítulo discute o segundo objetivo de pesquisa, que consiste na caracterização e modelagem de comportamento de clientes em sistemas de transmissão de vídeo adaptativo ao vivo.

Conforme mostrado na Seção 3.2, os serviços de vídeo na Internet têm passado por mudanças de tecnologia e escala ao longo dos últimos anos. Com isso, mesmo modelos de comportamento relativamente recentes não representam apropriadamente o contexto das transmissões atuais. Um dos motivos para esse fato é o surgimento de tecnologias como a transmissão adaptativa via *HTTP*. Sua dinâmica de funcionamento introduz novas variáveis no comportamento da aplicação cliente que ainda carecem de estudo mais aprofundado, principalmente em relação ao impacto dessas variáveis no engajamento do usuário e consumo de recursos.

A dinâmica de adaptação da aplicação cliente interfere na taxa de transmissão dos clientes, que como mostrado no Capítulo 5, tem relação com o engajamento dos usuários. No entanto, observa-se que modelos de comportamento têm desconsiderado essa relação, estabelecendo parâmetros de comportamentos gerais, independentes do desempenho experimentado. Isso pode levar a modelos que estimam de forma imprecisa aspectos da atividade do cliente e do usuário dentro do sistema.

Com base nessa premissa, este capítulo tem como foco a criação de um modelo de comportamento de clientes capaz de **(1)** descrever a adaptação de taxa de transmissão da aplicação cliente e **(2)** modelar o impacto do desempenho de transmissão no padrão de permanência e reingresso de usuários nesses clientes. A partir da correlação entre métricas de desempenho de transmissão e engajamento, foi desenvolvido um esquema para a identificação de grupos, similar ao processo de criação de cenários de desempenho, abordado na Seção 5.4, que permitiu identificar as nuances de comportamento de cada grupo e construir seus modelos especializados, mais precisos se comparados ao modelo único.

Este capítulo está organizado da seguinte forma. Na Seção 6.1 são discutidas as características do modelo proposto neste capítulo, assim como o processo de parametrização de seus componentes. Já na Seção 6.2, é mostrado o processo de especialização do modelo em sub-modelos originados de diferentes níveis de desempenho detectados por meio de agrupamento. Em seguida, na Seção 6.3, são avaliados os parâmetros dos modelos especializados, e na Seção 6.4, é apresentada uma comparação de desempenho entre o modelo único de comportamento versus o modelo especializado, utilizando cargas sintéticas produzidas por geradores construídos usando os parâmetros de ambos os modelos. Por fim, na Seção 6.5, é apresentado um sumário das contribuições do capítulo.

6.1 Um Modelo de Comportamento Hierárquico

Esta seção apresenta o modelo proposto para descrever o comportamento de clientes em transmissões adaptativas ao vivo. Seus componentes são apresentados na Seção 6.1.1 e a discussão dos parâmetros que descrevem o comportamento médio de um cliente é mostrada na Seção 6.1.2. O modelo desenvolvido nesta seção, único e geral para todos os clientes, é em seguida estendido para capturar padrões de comportamento específicos por meio de modelos especializados. Os modelos orientados por comportamento serão apresentados nas Seções 6.2 e 6.3.

No modelo proposto, foi assumido um conjunto de condições com o intuito de reduzir a probabilidade de interferências no comportamento padrão de um usuário. Em primeiro lugar, foram descartadas as sessões que começam ou terminam fora do intervalo de coleta, isto é, que começam antes do segundo minuto e/ou terminam após o penúltimo minuto. Esse tempo foi escolhido com base no intervalo de 2 minutos sem recepção de segmentos, que caracteriza fim de sessão, como descrito na Seção 4.1.2.

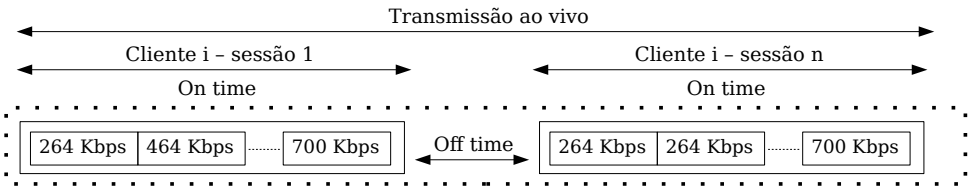
Em segundo lugar, considerou-se apenas sessões com inicialização bem sucedida, isto é, aquelas cujo cliente consegue receber todos os segmentos necessários para preencher seu *buffer* e iniciar a reprodução do vídeo. Admitindo essas restrições, foi mantido 68.2% das sessões para análise.

6.1.1 Componentes do Modelo de Comportamento

A proposta para este capítulo se trata de um modelo hierárquico que captura o comportamento de um cliente durante uma transmissão ao vivo. Este modelo é composto de duas camadas que são, respectivamente, a camada de *sessão* e a camada de *segmento de vídeo*.

A camada de sessão, no topo, é responsável por modelar os aspectos relativos à *permanência* (i.e. engajamento) do usuário do respectivo cliente na transmissão, como ilustrado na Figura 6.1. Ao longo da transmissão, é permitido ao cliente iniciar (e interromper) uma sequência não sobreposta de sessões. Durante uma sessão, são recebidos pelo cliente segmentos de vídeo assim que são gerados, sendo que a taxa de transmissão de cada segmento é determinada pelo cliente. Um cliente pode interromper uma sessão a qualquer momento e retornar posteriormente. Ou seja, um cliente pode possuir *múltiplas sessões* não sobrepostas em uma mesma transmissão, sendo estas sessões separadas por um período de inatividade.

Figura 6.1: A camada de sessão de um cliente



Fonte: Elaborado pelo autor.

Por meio deste modelo, conhecido como *"on/off"*, é possível capturar aspectos de alto nível de abstração do comportamento do cliente e seu usuário. Denota-se de *on-times* os intervalos nos quais o cliente está recebendo segmentos, e de *off-times* os períodos entre dois *on-times* consecutivos.

Uma vez que o cliente se junta a uma transmissão, ele imediatamente inicia sua sessão (estágio *on-time*). O cliente permanece nesse estágio por um tempo finito, eventualmente interrompendo a sessão e passando para o estágio de *off-time* ou deixando a transmissão permanentemente. A camada de sessão, portanto, é caracterizada pelo tempo de permanência em cada estágio, bem como o número de sessões de um cliente durante a transmissão. A modelagem de comportamento por meio de *on-times* e *off-times* é uma prática comum em diversos domínios [35, 40, 13] e, em particular, no contexto de transmissões ao vivo [119, 11].

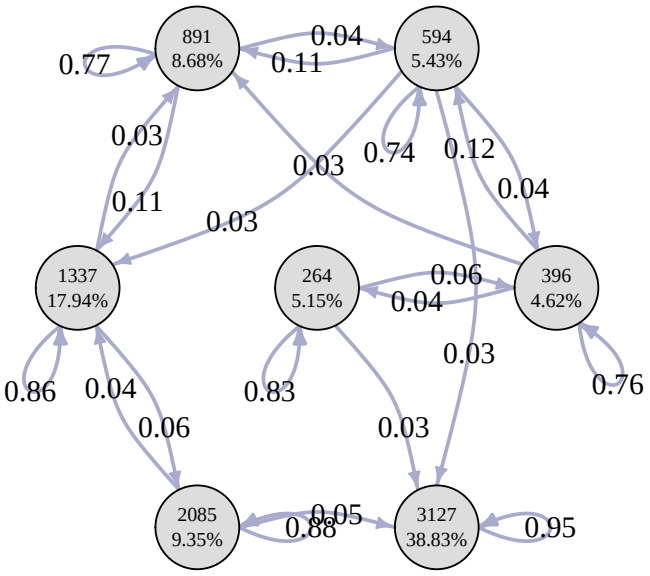
O modelo desta tese se diferencia dos anteriormente citados pela inclusão de uma segunda camada que captura explicitamente o desempenho de transmissão durante a

sessão, descrito especificamente por métricas associadas à adaptação de taxa de transmissão.

A nova camada, menos abstrata, denominada de camada de *segmento*, é responsável por descrever os aspectos relativos à dinâmica de adaptação durante as sessões do cliente (i.e., *on-times*). Como mostra o Capítulo 5, existem evidências de que a taxa de transmissão escolhida para cada segmento tem relação com o cenário de desempenho de um cliente. Por exemplo, clientes com baixo desempenho tenderão a escolher segmentos em uma taxa de transmissão menor para evitar congelamentos. Já clientes com alto desempenho de transmissão irão escolher taxas mais elevadas, o que impacta positivamente em seu engajamento.

Para capturar a adaptação de um cliente, é utilizada uma matriz de transições. Essa matriz descreve as probabilidades de transição entre nós de um grafo direcionado. Cada vértice representa uma taxa de transmissão disponível para um evento. Já as arestas determinam a probabilidade de transição da taxa de transmissão atual para a taxa do próximo segmento. Ou seja, assume-se que a taxa de transmissão de um segmento n depende apenas da taxa do segmento $n-1$.

Figura 6.2: Camada de segmento de vídeo



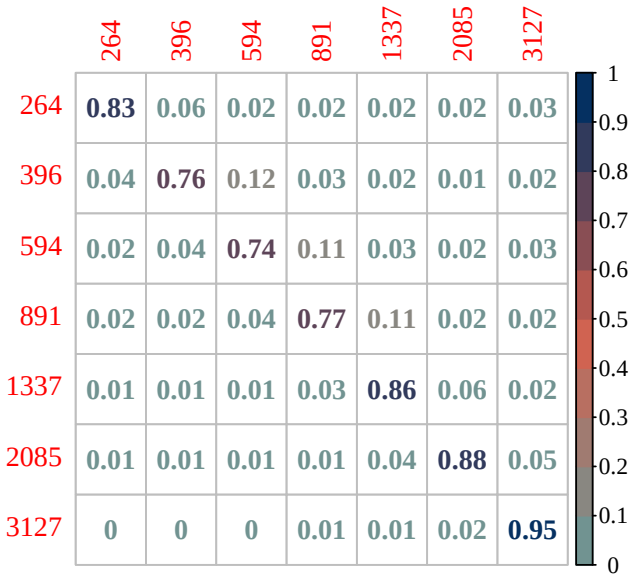
Fonte: Elaborado pelo autor.

A Figura 6.2 apresenta uma visão parcial da estrutura, obtida a partir de todo o conjunto de dados, onde as arestas com probabilidade abaixo de 2% foram removidas para simplificar a visualização. Nota-se na figura as 7 taxas de transmissão utilizadas nas transmissões. Nota-se também que, para cada vértice, a soma de todas as arestas de saída de um vértice (incluindo as omitidas) deve ser igual a 1. A figura também mostra, em cada vértice, a probabilidade estacionária de cada estado, isto é, a probabilidade

de permanência em cada estado caso essa esse grafo fosse "percorrido" por um tempo suficientemente longo usando as probabilidades apontadas nas transições.

A Figura 6.3 oferece uma perspectiva alternativa das probabilidades entre estados, agora incluindo todas as transições. Pela análise do conjunto de dados e desta figura, é possível observar a lógica geral de funcionamento do algoritmo de adaptação adotado no provedor. As sessões iniciam na menor taxa de transmissão (264 kbps) e sobem gradualmente para a maior taxa suportada por cada cliente, passando (e eventualmente permanecendo por um tempo) em taxas intermediárias. Uma evidência para essa subida gradual de taxa é que, para cada estado (representado pelas linhas da matriz), a segunda maior probabilidade é a que se refere à passagem para a taxa imediatamente superior. Nesse sentido, a maior probabilidade se refere aos laços (diagonal da matriz), o que evidencia a permanência na taxa corrente por um período de tempo antes da passagem para a próxima taxa.

Figura 6.3: Matriz de transição para as taxas de transmissão do evento



Fonte: Elaborado pelo autor.

Ainda em relação à permanência em uma dada taxa, a Figura 6.2, que apresenta a visão parcial da adaptação, mostra que a probabilidade estacionária de permanecer em 1337 kbps e 3127 kbps é de aproximadamente 18% e 39%, respectivamente, o que representa uma probabilidade muito maior se comparado com as demais taxas. Em outras palavras, nota-se que os clientes tendem a permanecer por mais tempo nessas referidas taxas. Além disso nota-se, agora pela Figura 6.3, que as probabilidades de saída de taxas mais baixas, notadamente 396, 594, e 891 kbps, são mais altas comparado aos estados associados com taxas mais altas. Por exemplo, a probabilidade de deixar a taxa

de 594 kbps, incluindo as transições omitidas, é de 26%. Por outro lado, um cliente transferindo a 3127 kbps tem apenas 5% de chance de deixar esta taxa. Ou seja, este padrão reforça a tendência de permanência em altas taxas de transmissão.

Vale ressaltar que as Figuras 6.2 e 6.3 representam o comportamento médio dos clientes no sistema. Essa representação pode ocultar padrões de adaptação secundários nos dados. Isso será explorado na Seção 6.3.4, que mostra que os clientes podem ser agrupados segundo seus cenários de desempenho de transmissão, revelando regimes de adaptação específicos. Mas antes de mostrar esses padrões, uma discussão a respeito da parametrização do modelo será apresentada.

6.1.2 Parametrização do Modelo Único

Esta seção descreve como são determinados os parâmetros do modelo que descreve o comportamento médio de clientes. O procedimento descrito nessa seção também será usado nas seções subsequentes para parametrizar os modelos especializados em cada grupo de comportamento encontrado.

A descrição começa pelos parâmetros da camada de sessão, nominalmente o *on-time*, o *off-time*, e o número de sessões. Para cada parâmetro, houve o ajuste de uma variável aleatória (descrita por distribuições bem conhecidas na literatura) em relação aos dados empíricos. A estimação dos parâmetros usou estimativa por máxima verossimilhança (*Maximum-Likelihood Estimator* (MLE) [120]) no caso de variáveis contínuas (*on* e *off-times*) e minimização do erro quadrático médio no caso de variáveis discretas (número de sessões). Para variáveis contínuas, foram consideradas as seguintes distribuições candidatas: Normal, Log-Normal, Exponencial, Cauchy, Chi, Chi-Quadrada, Gamma, e suas variações (Gamma Generalizada, Erlang), Logística, Beta, Uniforme, Weibull e Pareto. Para a variável discreta, foram consideradas as distribuições Poisson, Binomial, Binomial Negativa, Geométrica, e Hipergeométrica.

Foi utilizada a biblioteca `scipy`[1] para estimação dos parâmetros. Ela dá como saída as variáveis de posição e fator de escala das curvas. Além disso, foram utilizadas duas abordagens complementares para decidir qual o melhor ajuste entre os candidatos. Em primeiro lugar, foi empregada uma inspeção visual do corpo e da cauda das curvas, em busca das regiões com maior discrepância. Em segundo lugar, usou-se testes estatísticos. Para variáveis contínuas, foi utilizado o teste de Kolmogorov-Smirnov (KS) [15]. Este teste quantifica a diferença entre a distribuição empírica e uma função de distribuição de referência. Ele testa a hipótese nula de que ambas as distribuições são equivalentes.

[1] https://scipy.org/

Valores pequenos nessa estatística indicam que não se pode rejeitar a hipótese nula, e portanto existem evidências de que os dados empíricos seguem a distribuição de referência em um certo nível de significância α. Esse método também exibe um valor p-value bilateral que representa a probabilidade de se cometer um erro ao rejeitar a hipótese nula. Todos os ajustes consideram um nível de significância de $\alpha = 0.05$. Assim, pode-se rejeitar a hipótese nula quando o valor está abaixo de α.

No caso da variável discreta (número de sessões), foi empregado o teste de Anderson-Darling (AD) [110]. Assim como no teste KS, o teste AD produz, como saída, uma estatística computada para os dados empíricos: se a estatística é maior do que um valor de referência – calculado para um valor de significância α e um tamanho de amostra – é possível rejeitar a hipótese nula de que a amostra empírica e a distribuição teórica são equivalentes. Foi adotado o mesmo nível de significância de $\alpha = 0.05$ considerado para as variáveis contínuas.

Os valores dos melhores ajustes para os parâmetros da camada de sessão do modelo são mostrados na Tabela A.1 (veja Apêndice A), onde parâmetros do conjunto de dados completo estão na parte inferior da tabela. As duas colunas mais à direta mostram os testes estatísticos que indicam que as distribuições escolhidas se adequam bem aos dados. Foi determinado que a distribuição Weibull produziu o melhor ajuste para o *on-time*, enquanto que a Weibull Exponencial melhor se ajustou ao *off-time*. Para o número de sessões, o melhor ajuste se deu pela Binomial Negativa.

Em relação aos parâmetros da camada de segmento de vídeo, vale relembrar que eles se referem às probabilidades de cada transição entre taxas. Nesse caso, os parâmetros do modelo único já foram mostrados na Figura 6.3. Essas probabilidades foram estimadas tomando as adaptações de todos os clientes e calculando as probabilidades médias correspondentes.

6.2 Modelos Especializados por Nível de Desempenho

O Capítulo 5 mostrou que o engajamento de um usuário depende do nível de desempenho a que seu cliente está submetido. Analogamente, essa seção utiliza tal motivação para investigar a possível existência de múltiplos comportamentos de clientes induzidos por diferentes níveis de desempenho de transmissão.

O primeiro passo para a identificação das nuances de cada comportamento é o agrupamento dos clientes segundo suas similaridades de desempenho e comportamento.

Se o processo de agrupamento automatizado identificar mais de um grupo, isso significa que pode haver mais de um perfil de desempenho dentro do sistema de transmissão. Em seguida, de posse desses grupos, é feito o processo de parametrização do comportamento, que vai determinar se existem diferenças significativas nos valores dos parâmetros que definem o modelo proposto. Diferenças nas distribuições estatísticas de um parâmetro P para grupos diferentes, implicam que o desempenho de transmissão afetou o componente de comportamento dos clientes descrito por P. Nesse sentido, a hipótese é que a utilização de uma parametrização personalizada por grupo melhora a representação dos múltiplos comportamentos dos clientes.

A organização da seção é a seguinte. A Seção 6.2.1 descreve a metodologia para agrupamento de clientes, e a Seção 6.2.2 caracteriza as propriedades dos grupos identificados. Por fim, a Seção 6.3 irá discutir os ajustes dos parâmetros dos modelos especializados, e fará a comparação deste com o modelo único.

6.2.1 Processo de Agrupamento de Clientes

Os clientes foram agrupados com base na sua similaridade de desempenho e comportamento. Foi utilizada para isso uma abordagem não supervisionada clássica em aprendizado de máquina. Nela, cada cliente é representado por um vetor de atributos relacionados ao modelo de comportamento apresentado na Seção 6.1.

Admite-se que um mesmo usuário pode exibir diferentes padrões de comportamento em diferentes transmissões em virtude de, por exemplo, o uso de diferentes dispositivos ou outros aspectos contextuais. Por isso, considerou-se cada par $\{cliente, partida\}$ como *clientes diferentes* em essência para fins de agrupamento. Um usuário que assiste a múltiplas transmissões contribui com múltiplos vetores de atributos durante o agrupamento. Com isso, o mesmo usuário pode estar em diferentes grupos caso participe de múltiplas transmissões. Por questões de simplicidade, será usado o termo *cliente* para se referir aos vetores de atributos. Com base nos estudos apresentados no Capítulo 5, foram selecionados os seguintes atributos para representar cada cliente[2]:

- *Número de sessões* que o cliente teve durante uma transmissão.
- *Duração média de sessão* (*on-time*) de um cliente durante uma sessão.
- *Off-time médio* entre sessões consecutivas de um cliente durante uma transmissão.

[2]Atributos relacionados com o *buffer*, isto é, a latência de inicialização e taxa de congelamentos, não fazem parte do comportamento do cliente e, por isso, não foram incluídos. No entanto, tais eventos induzem mudanças no comportamento do cliente. Por exemplo, uma redução no nível de *buffer* pode forçar o cliente a requisitar segmentos em uma taxa de transmissão menor.

- *Prevalência em taxa de transmissão*, definido como a fração de requisições de segmentos de cada taxa de transmissão, considerando todas as as sessões de um cliente (um atributo por taxa)
- *Probabilidades de transição entre taxas* que captura a matriz de transições que modela o comportamento do algoritmo de adaptação do cliente (ver Figura 6.3)
- *Tipo de dispositivo* utilizado pelo cliente.
- *Múltiplas sessões*, um atributo binário que indica se um cliente possui uma única sessão ou múltiplas sessões durante a transmissão. Como será discutido a seguir, a computação do *off-time* médio leva este atributo em consideração.

Os atributos numéricos foram normalizados para o intervalo $[0-1]$ para assegurar que estivessem na mesma escala. Utilizou-se a normalização *min-max*, formalizada como $x_{norm} = (x - x_{min})/(x_{max} - x_{min})$, onde x_{norm} é o valor normalizado e x_{min} e x_{max} são o valores mínimo e máximo de um dado atributo dos dados de entrada, respectivamente. O tipo de dispositivo, segmentado em duas variáveis binárias (móvel e fixo), indica se o dispositivo usado pelo cliente é móvel ou fixo.

Finalmente, foi adicionado um atributo binário que indica se um cliente tem uma ou várias sessões durante uma transmissão. Esse atributo é necessário porque, no caso de uma sessão única, não existem *off-times* e tal atributo deve ser ajustado de maneira a manter a semântica dos dados. Se um cliente tem apenas uma sessão, então o atributo de "múltiplas sessões" é ajustado para 0 e o *off-time* médio para -1; caso contrário o atributo é ajustado para 1 e o *off-time* é calculado usando os tempos entre sessões consecutivas. Empregando essa abordagem, é possível separar os clientes com múltiplas sessões dos que possuem somente uma.

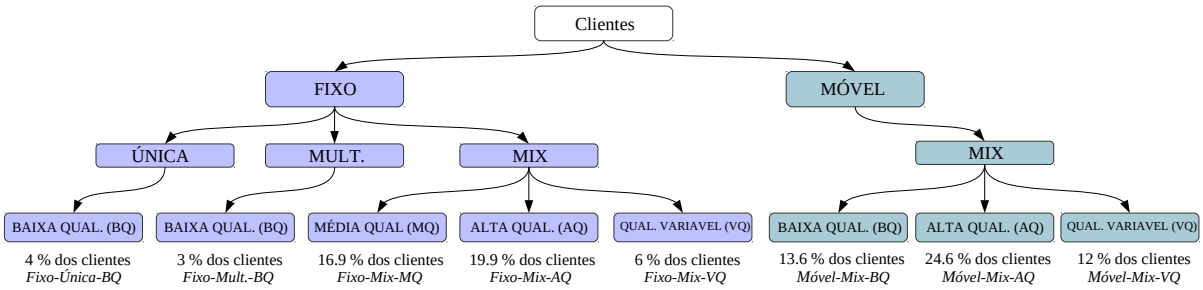
Assim como na Seção 5.4, utilizou-se o algoritmo *k-means* [36] para fazer o agrupamento de clientes. Foi estipulado o uso da distância euclidiana num espaço n-dimensional, onde n corresponde ao número de atributos, como critério de semelhança entre clientes. Também foram avaliadas técnicas de agrupamento por densidade e por hierarquia [31, 93], que levaram a resultados similares.

A forma de determinar o número de grupos é análoga à utilizada no Capítulo 5 [117]. Usando esta metodologia, foi determinado um $k = 8$, isto é, oito grupos ou tipos de comportamentos de clientes. Na próxima seção, será mostrada uma descrição desses grupos, em termos dos valores dos atributos, com foco na análise dos centroides, que oferecem uma visão do comportamento médio dentro dos grupos.

6.2.2 Comportamento Médio de Cada Perfil de Desempenho

A Figura 6.4 apresenta graficamente os 8 grupos de comportamento identificados nos dados, com os nomes sendo apresentados na parte inferior da mesma. Uma inspeção manual revela que o comportamento dentro de cada grupo é relativamente homogêneo e ditado por um pequeno subconjunto de atributos, representados hierarquicamente na figura. Cada grupo é rotulado com sua fração de clientes, assim como uma sigla para referência. Os grupos foram nomeados seguindo a hierarquia apresentada na figura. Por exemplo, o nome *Fixo-única-BQ* é usado para se referir ao grupo caracterizado por clientes fixos, com uma única sessão cada, que requisitam a maior parte dos segmentos em baixa taxa de transmissão (o grupo mais à esquerda da Figura 6.4).

Figura 6.4: Representação hierárquica dos grupos de comportamento

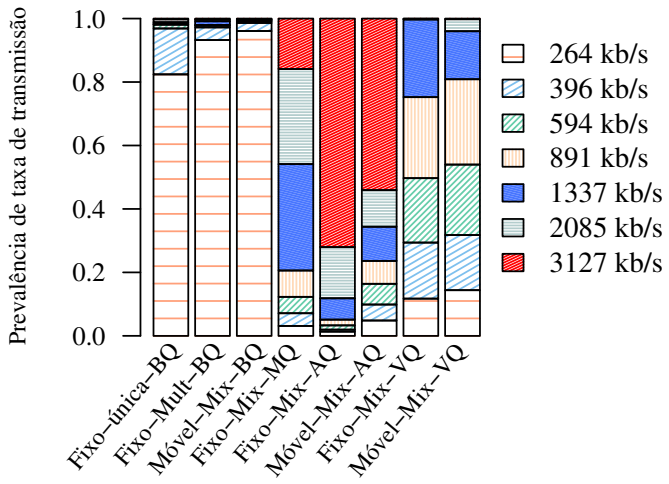


Fonte: Elaborado pelo autor.

O tipo de dispositivo exerce um papel preponderante na formação dos perfis de comportamento: 5 dos 8 grupos são caracterizados por clientes usando dispositivos fixos como *smart-TV's* e computadores *desktop* (rotulados como *Fixo*) e os 3 restantes são compostos de dispositivos móveis como *smartphones* e *tablets* (rotulados como *Móvel*). Além disso, os grupos são também caracterizados pelo número de sessões que possuem durante uma transmissão. Pela inspeção dos centróides dos grupos, foram encontradas três categorias: clientes com uma ou mais sessões (rotulados como *Mix*), clientes com duas ou mais sessões (rotulados como *Mult*) e clientes com sessões únicas (rotulados como *Única*). Vale notar que todos os grupos de clientes móveis estão na categoria *Mix*, enquanto que os clientes fixos possuem três grupos na categoria *Mix*, um na categoria *Única* e um na categoria *Mult*.

Finalmente, nota-se que os grupos também se distinguem pelo desempenho de transmissão, representado por suas taxas de transmissão. Foram identificados 4 diferentes padrões, aos quais se rotulou de *baixa qualidade (BQ)*, *média qualidade (MQ)*, *alta qualidade (AQ)* e *qualidade variável (VQ)*. Vale enfatizar que o termo *qualidade* é usado como sinônimo para a taxa de transmissão devido a relação direta desta métrica com aspectos de qualidade de imagem como, a resolução e taxa de quadros, por exemplo.

Figura 6.5: Fração média de taxa de transmissão em cada grupo/perfil



Fonte: Elaborado pelo autor.

Os padrões de qualidade são caracterizados na Figura 6.5, que ilustra a média de prevalência de taxas de transmissão em cada grupo. Cada barra representa um grupo/perfil de comportamento e as frações por taxas são suas médias considerando todos os clientes dos respectivos grupos. Na figura as três barras mais à esquerda se referem aos grupos de baixa qualidade (BQ), cuja vasta maioria dos segmentos (mais de 80%) é da menor taxa de transmissão disponível, i.e., 264 kbps. Nota-se também que a fração destes segmentos é maior no grupo de clientes móveis (*Móvel-Mix-BQ*), e alcança quase 100%. A menor taxa experimentada por clientes móveis também foi constatada na Seção 5.2.

O grupo *Fixo-Mix-MQ*, por sua vez, é caracterizado por mais de 60% dos segmentos em taxas intermediárias, entre 1337 e 2085 kbps. Dois outros grupos – *Fixo-Mix-AQ* e *Móvel-Mix-AQ* – são caracterizados por receber segmentos na maior taxa disponível, isto é, 3127 kbps. Esses dois grupos diferem entre si em termos do dispositivo usado e, novamente, nota-se que o grupo de clientes fixos tem uma fração maior de segmentos na maior taxa (acima de 70% de segmentos). Por fim, dois grupos são caracterizados por uma taxa variável – *Fixo-Mix-VQ* e *Móvel-Mix-VQ* – com uma fração não desprezível de segmentos sendo requisitados em taxas diferentes, de 264 a 1337 kbps.

Uma vez definidos os oito grupos, será dado enfoque à distribuição dos clientes entre esses grupos. Como mostrado na parte inferior da Figura 6.4, é possível notar que a fração de clientes móveis e fixos é similar, de por volta de 50% cada um. Além disso, nota-se que a fração de clientes no grupo *Móvel-Mix-BQ* (13%), caracterizado por segmentos de baixa taxa de transmissão, é mais de duas vezes maior que a fração de clientes fixos com a mesma característica – 4% e 3% de clientes no grupo *Fixo-Única-BQ* e *Fixo-Mult-BQ*, respectivamente. Estas figuras refletem um padrão já mencionado de que os clientes móveis experimentam baixa qualidade mais frequentemente do que os clientes fixos. De

modo similar, foi registrado que a fração de clientes móveis que experimentou qualidade variável, i.e., clientes no grupo *Móvel-Mix-VQ* (12%), é duas vezes maior do que fração de clientes fixos no grupo correspondente *Fixo-Mix-VQ* (6%). A maior variação de vazão em conexões móveis pode ser uma explicação para esses números.

Nota-se também que quase metade dos clientes móveis estão no grupo *Móvel-Mix-AQ*, caracterizado pela requisição de segmentos de alta qualidade. Essa fração é maior que a do grupo correspondente de clientes fixos (*Fixo-Mix-AQ* – em torno de 20%). Por outro lado, uma larga fração de clientes fixos (quase 17%) é caracterizada por segmentos de qualidade média, estando no grupo *Fixo-Mix-MQ* (o segundo maior grupo entre os grupos de clientes fixos), o que não é visto nas categorias móveis.

A seguir, serão discutidas as propriedades de cada grupo, com foco nos componentes que compõem o modelo de comportamento de clientes.

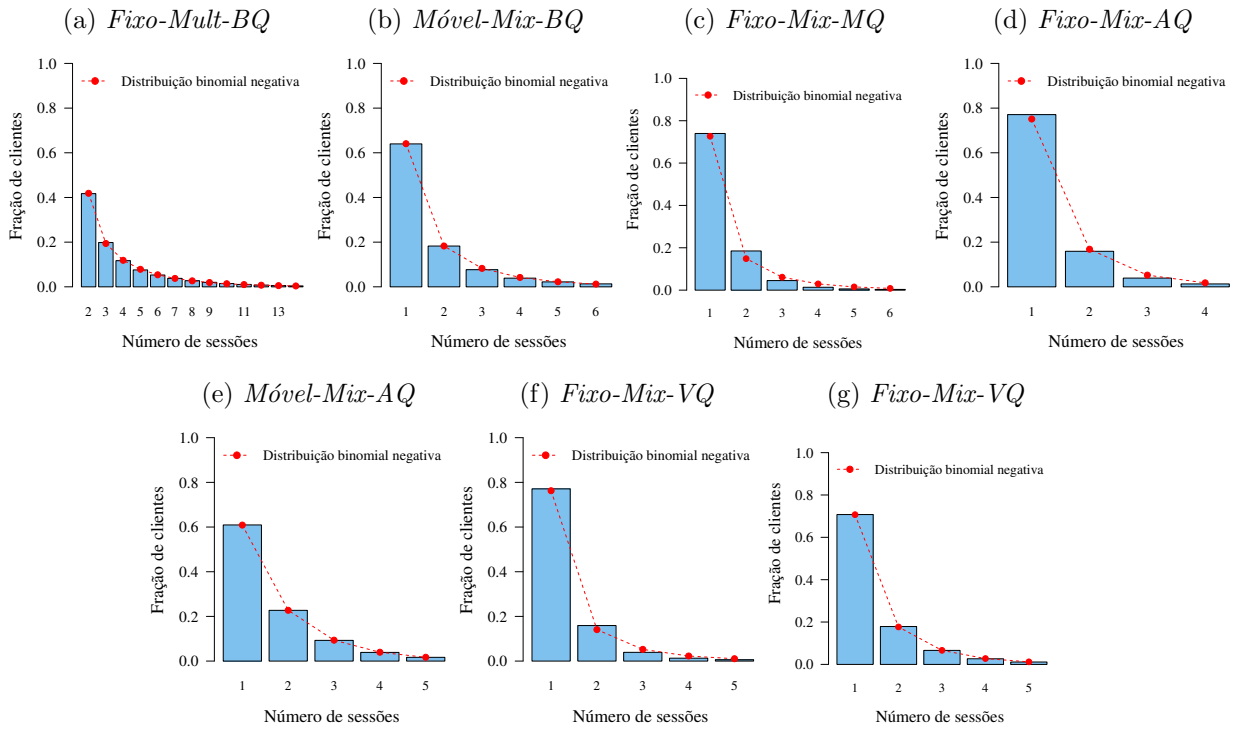
6.3 Parametrização dos Modelos Especializados

Nesta seção, será apresentada a parametrização dos componentes de comportamento no nível dos modelos especializados. Foi empregada a mesma metodologia discutido na Seção 6.1.2. Em paralelo a essa discussão, serão também comparados os erros produzidos nos ajustes de curvas destes modelos em relação ao modelo único. A métrica utilizada será o erro quadrático médio (*Mean Squared Error* (MSE) [121]), definido como $\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$ onde y_i é o i-ésimo percentil da CDF com os dados reais e o $\tilde{y}_i$ o i-ésimo percentil da distribuição que melhor se ajustou aos dados. Será então comparado o MSE obtido pelo modelo único e o MSE médio ponderado ($wMSE$) computado para todos os grupos[3].

As próximas seções discutem os ajustes de cada parâmetro do modelo separadamente, salientando os padrões particulares. As descrições completas dos melhores ajustes, incluindo resultados dos testes de aderência e MSE são apresentados no Apêndice A.

[3]O MSE médio ponderado, referido como $wMSE$, é definido como $\sum_i p_i \times \text{MSE}_i$, onde p_i é a fração dos clientes no grupo i e MSE_i é o erro quadrático médio da distribuição que produziu o melhor ajuste para o grupo/perfil i.

Figura 6.6: Melhores ajustes para o número de sessões em diferentes grupos



Fonte: Elaborado pelo autor.

6.3.1 Número de Sessões

A Figura 6.6 apresenta as distribuições do número de sessões por cliente durante a transmissão. O grupo *Fixo-Única-BQ* não é mostrado, uma vez que seus clientes possuem apenas uma sessão por transmissão. Notar que, para todos os grupos, a maior parte dos clientes têm poucas sessões por transmissão. Além disso, com exceção do grupo *Fixo-Mult-BQ* que, por definição, consiste de clientes com pelo menos duas sessões, todos os outros grupos têm a maior parte dos clientes (pelo menos 60%) com apenas uma sessão por transmissão.

Curiosamente, os grupos que experimentaram altas qualidades, notadamente *Fixo-Mix-AQ*, *Fixo-Mix-AQ*, *Fixo-Mix-VQ* e, em uma menor extensão, o grupo Movel-Mix-VQ, têm uma grande concentração de clientes com apenas uma sessão por transmissão. Em contraste, os grupos de clientes com mais sessões por transmissão foram os caracterizados por clientes com baixa taxa de transmissão, isto é, os grupos *Fixo-Mult-BQ* e *Fixo-Mix-BQ* (Especialmente *Fixo-Mult-BQ*). Como será discutido na Seção 6.3.4, e também como foi mostrado no Capítulo 5, existem evidências de que esses clientes experimentam mais congelamentos, que são conhecidos por provocar um impacto negativo significativo no engajamento do usuário [91, 39]. A partir dessa perspectiva, especula-se que os grupos

Fixo-Mult-BQ e *Movel-Mix-BQ* são de clientes que estão interessados no conteúdo e permanecem no sistema, iniciando novas sessões, interrompidas por um baixo desempenho persistente. Em contraste, o grupo *Fixo-Única-BQ* pode consistir de clientes que escolhem abandonar o sistema tanto por não estarem interessados no conteúdo quanto por simplesmente não tolerarem o baixo desempenho de transmissão.

No caso dos parâmetros dos modelos, foi registrado que o número de sessões por transmissão é melhor descrito por uma distribuição Binomial Negativa para todos os grupos, assim como para o modelo único (Seção 6.1.2). O teste Anderson-Darling mostrou evidências de que os dados empíricos vêm das distribuições teóricas escolhidas. Entretanto, os parâmetros das distribuições são muito diferentes entre os grupos e também em relação ao ajuste do modelo único, o que evidencia grande diversidade de comportamento entre os grupos. A média ponderada do MSE de todos os grupos ($wMSE$=0,00008) é uma ordem de magnitude menor do que a encontrada para o modelo único (MSE=0,00022), indicando que a especialização de fato melhora a acurácia da representação deste componente do comportamento dos clientes.

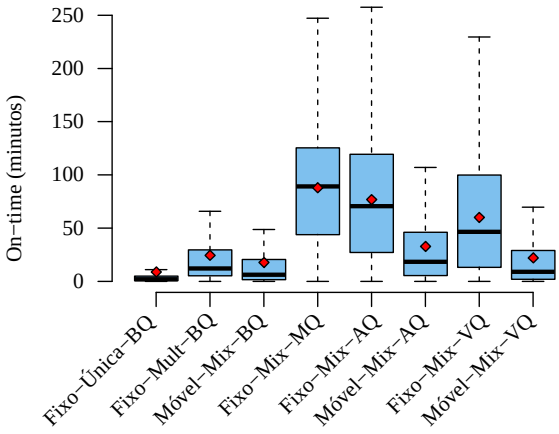
6.3.2 Duração de Sessão (*On-times*)

A Figura 6.7 sumariza as distribuições para as durações de sessão (i.e., *on-times*) dos clientes para todos os grupos. Cada distribuição é representada por um um gráfico de caixa, onde as caixas azuis vão do 1º ao 3º quartis, e os traços representam o 5º e o 95º percentis. O traço horizontal no centro das caixas representa a mediana e os pontos vermelhos representam as médias.

A Figura 6.7 mostra que a duração de sessão varia significativamente entre os grupos, mesmo considerando as médias e medianas. A título de comparação, a média geral de *on-time* é de 30 minutos, valor significativamente diferente das médias por grupo mostradas na figura. Nota-se também que grupos de clientes móveis têm durações de sessões distintas às dos grupos correspondentes em clientes fixos. Por exemplo, o par mediana/média do grupo *Móvel-Mix-AQ* é $32,85/18,37$ minutos, já o valor para o grupo *Fixo-Mix-AQ* é 76.89/70.65 minutos. A qualidade das conexões 3G e 4G e a autonomia restrita de energia dos dispositivos móveis, assim como o contexto em que são usados, pode restringir o tempo que esses usuários permanecem conectados à transmissão.

Por outro lado, é possível perceber que o desempenho de transmissão, representado pela taxa de transmissão, também desempenha um papel importante para a permanência. Clientes dos grupos de desempenho (**-*-AQ*, **-*-MQ* e mesmo **-*-VQ*) tendem a assistir o vídeo por mais tempo do que clientes em baixa qualidade (**-*-BQ*). Por exemplo, a

Figura 6.7: Distribuição das durações de sessão (*on-times*) para diferentes grupos



Fonte: Elaborado pelo autor.

duração média de permanência no grupo *Fixo-Mix-AQ* e *Fixo-Mix-AQ* é de 76,89 e 87,99 minutos, respectivamente, contra 8.84 e 24.49 minutos para clientes em *Fixo-Única-BQ* e *Fixo-Mult-BQ*. Essa observação está em linha com os trabalhos relacionados, que reportam uma correlação positiva entre a duração de sessão e taxa de transmissão [27, 72, 8, 47], assim como na discussão da Seção 5.

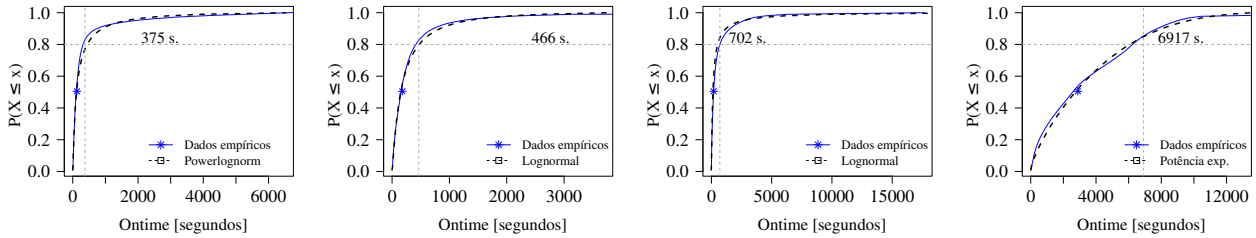
A Figura 6.7 também ilustra um benefício do mecanismo adaptativo já reportado em trabalhos anteriores [128]. Os usuários parecem tolerar melhor mudanças frequentes de taxa de transmissão do que congelamentos (grupos *BQ* possuem as maiores taxas de congelamentos). Esse comportamento é visível quando se compara o grupo *Fixo-Mix-VQ* com os grupos *Fixo-Única-BQ* e *Fixo-Mult-BQ*. A duração média de sessão do grupo *VQ* é 43 minutos contra 8.84 e 24.49 minutos dos grupos *BQ*.

A Figura 6.8 mostra os ajustes das curvas relativas à duração de sessão nos grupos. Para facilitar a interpretação das curvas, foi destacado o 80^{o} percentil de cada distribuição, por meio de linhas tracejadas nos gráficos. A distribuição que proporcionou o melhor ajuste varia significativamente entre os grupos, tanto em termos do tipo, quanto dos parâmetros. Por exemplo, como apresentado no Apêndice A, Lognormal Potência, Weibull Exponencial e Erlang são alguns dos modelos de distribuição usados para ajustar o *on-time* de alguns grupos. Nota-se que, para o modelo único, constatou-se que a distribuição Weibull foi a que ofereceu o melhor ajuste para o *on-time*. Adicionalmente, a metodologia mostrada na Seção 6.1.2 foi usada para validar a qualidade desses ajustes usando o teste de Kolmogorov-Smirnov. Durante o processo, foi possível observar que o MSE ponderado ($wMSE = 97.332$) é quase três vezes menor que o MSE computado para o melhor ajuste do modelo único ($MSE = 275.164$). Mais uma vez, os modelos especializados foram capazes de descrever com mais acurácia a diversidade de padrões, bem como distinguir as propriedades que as caracterizam. Os modelos especializados são particularmente úteis

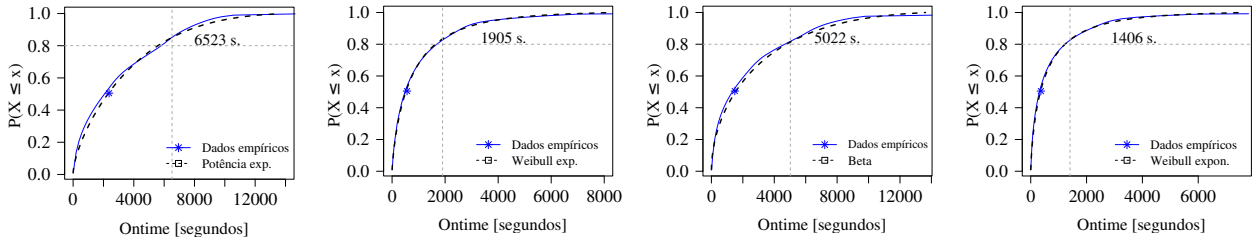
para capturar a cauda das distribuições de forma mais precisa, reduzindo a chance de ocorrência de sessões cujas durações excedem a duração de uma transmissão.

Figura 6.8: Ajustes das distribuições de *on-time* para diferentes grupos

(a) *Fixo-Única-BQ* (b) *Fixo-Mult-BQ* (c) *Móvel-Mix-BQ* (d) *Fixo-Mix-MQ*



(e) *Fixo-Mix-AQ* (f) *Móvel-Mix-AQ* (g) *Fixo-Mix-VQ* (h) *Móvel-Mix-VQ*



Fonte: Elaborado pelo autor.

6.3.3 Intervalo Entre Sessões (*Off-times*)

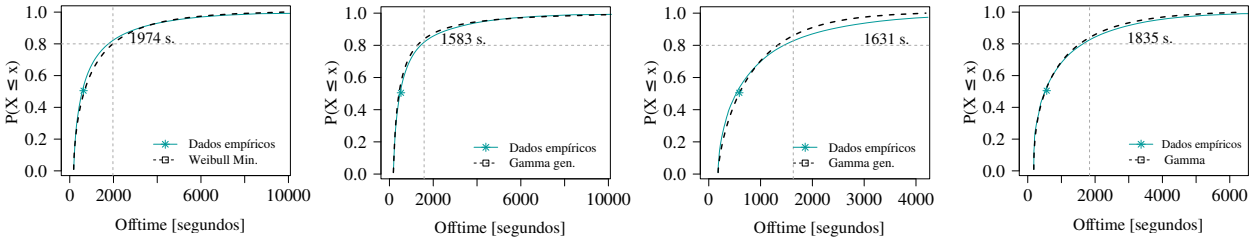
A Figura 6.9 apresenta as distribuições dos *off-times*, isto é, o intervalo entre duas sessões consecutivas de um cliente durante uma transmissão, assim como os melhores ajustes. Novamente, o tipo de dispositivo parece desempenhar um importante papel: clientes fixos tendem a ter *off-times* mais curtos, retornando ao sistema mais rapidamente quando comparado com clientes móveis. Esse comportamento pode ser percebido pela comparação do 80º percentil das distribuições. A única exceção é o grupo *Móvel-Mix-BQ*. Nesse caso, vale relembrar que o grupo *Móvel-Mix-BQ* tem também as menores durações de sessão. clientes fixos com qualidade baixa (*Fixo-Mult-BQ*) possuem *off-times* mais longos. Conjectura-se que, no caso de clientes móveis, há um maior interesse que faz com que o retorno seja mais rápido. Em contraste, clientes fixos esperam boa qualidade e são menos tolerantes a grandes degradações. Vale lembrar também da existência de um grupo *BQ* com sessões únicas (e curtas) apenas para clientes fixos, que reforça a hipótese de que clientes fixos são menos susceptíveis a retornar em caso de baixa qualidade.

Com relação às distribuições, novamente foi encontrada uma grande diversidade

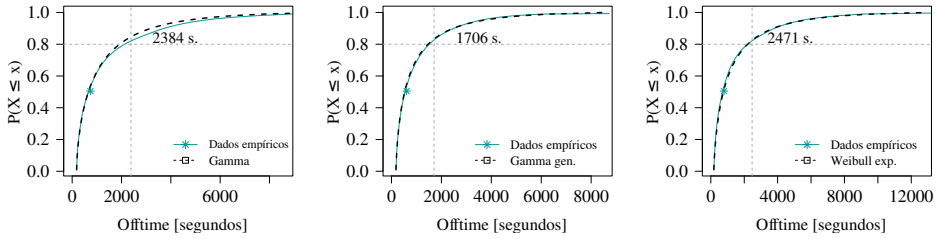
nos modelos, tais como Gamma e Weibull, assim como nos parâmetros para diferentes grupos. Isso enfatiza a heterogeneidade do comportamento de clientes. Além disso, como já foi observado para os outros parâmetros, registrou-se que o modelo especializado oferece uma descrição mais precisa dos *off-times* ($wMSE = 63.384$) do que o modelo único ($MSE = 119.985$).

Figura 6.9: Ajustes das distribuições de *off-time* para diferentes grupos

(a) *Fixo-Mult-BQ* (b) *Móvel-Mix-BQ* (c) *Fixo-Mix-MQ* (d) *Fixo-Mix-AQ*



(e) *Móvel-Mix-AQ* (f) *Fixo-Mix-VQ* (g) *Móvel-Mix-VQ*



Fonte: Elaborado pelo autor.

6.3.4 Matrizes de Transição de Taxa de Transmissão

Finalmente, é apresentado o conjunto final de parâmetros do modelo proposto, correspondente às matrizes de probabilidades de transição, que capturam os diferentes regimes de adaptação seguidos pelos clientes. A Figura 6.10 mostra as cadeias de probabilidades para cada grupo.

Figura 6.10: Matrizes de transição entre taxas de transmissão para cada grupo

(a) *Fixo-Única-BQ*

	264	396	594	891	1337	2085	3127
264	**0.98**	0.02	0	0	0	0	0
396	0.05	**0.95**	0	0	0	0	0
594	0.13	0.03	**0.83**	0.01	0	0	0
891	0.28	0.01	0.01	**0.65**	0.04	0.01	0
1337	0.24	0.01	0.01	0.01	**0.71**	0.02	0
2085	0.14	0	0.01	0	0.01	**0.83**	0.01
3127	0.44	0.01	0.01	0.01	0.01	0.01	**0.51**

(b) *Fixo-Mult-BQ*

	264	396	594	891	1337	2085	3127
264	**0.99**	0.1	0	0	0	0	0
396	0.09	**0.9**	0.01	0	0	0	0
594	0.12	0.05	**0.8**	0.01	0	0.01	0.01
891	0.22	0.02	0.02	**0.67**	0.04	0.02	0.01
1337	0.05	0	0	0	**0.93**	0.02	0
2085	0.07	0.01	0.01	0.01	0.01	**0.89**	0
3127	0.27	0.02	0.02	0.01	0.01	0.01	**0.66**

(c) *Móvel-Mix-BQ*

	264	396	594	891	1337	2085	3127
264	**0.99**	0.1	0	0	0	0	0
396	0.13	**0.86**	0.01	0	0	0	0
594	0.18	0.05	**0.68**	0.01	0.01	0.01	0.06
891	0.11	0	0.01	**0.87**	0.01	0	0
1337	0.14	0.01	0.01	0.02	**0.74**	0.06	0.02
2085	0.05	0	0	0	0.01	**0.92**	0.02
3127	0.25	0.02	0.08	0.01	0.03	0.04	**0.57**

(d) *Fixo-Mix-MQ*

	264	396	594	891	1337	2085	3127
264	**0.67**	0.14	0.04	0.04	0.06	0.03	0.02
396	0.05	**0.63**	0.2	0.04	0.05	0.02	0.01
594	0.03	0.06	**0.62**	0.20	0.06	0.02	0.01
891	0.02	0.02	0.04	**0.68**	0.2	0.03	0.01
1337	0.01	0.01	0.01	0.03	**0.88**	0.05	0.01
2085	0	0.01	0.01	0.01	0.04	**0.91**	0.02
3127	0	0	0	0.01	0.02	0.02	**0.94**

(e) *Fixo-Mix-AQ*

	264	396	594	891	1337	2085	3127
264	**0.51**	0.05	0.03	0.03	0.06	0.08	0.24
396	0.05	**0.38**	0.18	0.05	0.09	0.1	0.15
594	0.03	0.02	**0.44**	0.14	0.09	0.08	0.2
891	0.02	0.01	0.03	**0.44**	0.2	0.1	0.2
1337	0.01	0.01	0.01	0.03	**0.71**	0.12	0.11
2085	0.01	0	0.01	0.01	0.03	**0.86**	0.08
3127	0	0	0	0.01	0.01	0.01	**0.97**

(f) *Móvel-Mix-AQ*

	264	396	594	891	1337	2085	3127
264	**0.76**	0.04	0.05	0.03	0.03	0.03	0.06
396	0.02	**0.79**	0.07	0.03	0.03	0.02	0.04
594	0.02	0.03	**0.77**	0.07	0.03	0.02	0.06
891	0.02	0.02	0.04	**0.76**	0.09	0.03	0.04
1337	0.02	0.01	0.01	0.04	**0.78**	0.08	0.06
2085	0.01	0.01	0.01	0.02	0.06	**0.76**	0.13
3127	0.01	0	0	0.01	0.01	0.02	**0.95**

(g) *Fixo-Mix-VQ*

	264	396	594	891	1337	2085	3127
264	**0.86**	0.1	0.02	0.01	0.01	0	0
396	0.04	**0.84**	0.1	0.01	0.01	0	0
594	0.02	0.05	**0.85**	0.07	0.01	0	0
891	0.01	0.02	0.03	**0.91**	0.03	0	0
1337	0.01	0.01	0.01	0.02	**0.95**	0	0
2085	0.07	0.08	0.12	0.13	0.13	**0.46**	0.01
3127	0.06	0.07	0.19	0.2	0.13	0.01	**0.34**

(h) *Móvel-Mix-VQ*

	264	396	594	891	1337	2085	3127
264	**0.87**	0.05	0.04	0.03	0.01	0	0
396	0.03	**0.88**	0.06	0.02	0.01	0	0
594	0.02	0.04	**0.88**	0.05	0.01	0	0
891	0.02	0.01	0.03	**0.9**	0.04	0	0
1337	0.01	0.01	0.02	0.05	**0.89**	0.02	0
2085	0.01	0.01	0.01	0.03	0.09	**0.84**	0.01
3127	0.08	0.11	0.11	0.11	0.14	0.17	**0.28**

Fonte: Elaborado pelo autor.

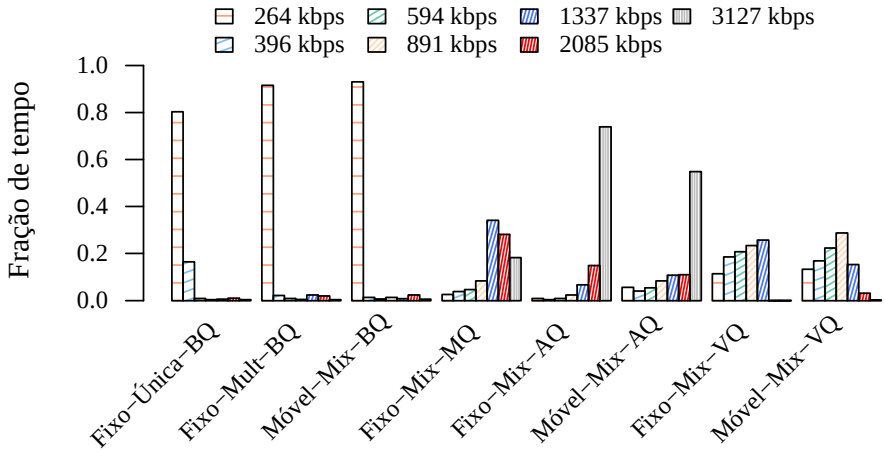
Por meio da análise das primeiras colunas das matrizes de transição da Figura 6.10, é possível notar, para grupos de baixa qualidade (**-*-BQ*), **(1)** altas probabilidades de transicionar diretamente para a menor taxa de transmissão (independente da taxa corrente); **(2)** altas probabilidades de permanecer na menor taxa. Por exemplo, um cliente *Fixo-Única-BQ* assistindo com a taxa de 3127 kbps tem 44% de chance de redução direta até 264 kbps. O mesmo tipo de transição ocorre com até 8% de probabilidade para grupos de alta qualidade (i.e., **-*-MQ*, **-*-AQ* e **-*-VQ*), incluindo grupos de clientes

móveis. Transições diretamente para a menor taxa de transmissão usualmente indicam a ocorrência de congelamentos, uma vez que o comportamento do algoritmo de adaptação é, após a ocorrência de congelamentos, requisitar segmentos da menor taxa para acelerar a recomposição do *buffer* do cliente.

Ao comparar as probabilidades de transição dos grupos **-*-BQ* e **-*-VQ*, nota-se novamente o benefício da adaptação de taxa: há uma chance maior de transições entre taxas vizinhas, especialmente quando iniciado em taxas mais altas (e.g., 2085 e 3127 kbps) em grupos **-*-VQ*. Por exemplo, os clientes do grupo *Fixo-Mix-VQ* recebendo segmentos a 3127 kbps têm 13% de chance de transicionar para 1337 kbps, e 20% de ir para 891 kbps, mas apenas 6% de chance de ir para a menor taxa. Uma observação similar pode ser feita para clientes iniciando a 2085 kbps. Essas reduções mais suaves quando comparado com os clientes **-*-BQ* parecem se dever ao algoritmo de adaptação reagindo a flutuações na largura de banda com o intuito de evitar congelamentos. Essa abordagem se justifica uma vez que os grupos **-*-VQ* registraram um engajamento maior que os grupos **-*-BQ* (i.e., tiveram sessões mais duradouras), como mostra a Figura 6.7.

Como uma última observação em relação à Figura 6.10, nota-se que em grupos de alta e média qualidade, notadamente os grupos *Fixo-Mix-MQ, Fixo-Mix-AQ* e *Móvel-Mix-AQ*, uma elevada chance de crescimento para as maiores taxas de transmissão e posterior permanência nas mesmas.

Figura 6.11: Distribuições estacionárias de cada perfil de desempenho



Fonte: Elaborado pelo autor.

Por fim, a Figura 6.11 mostra as probabilidades de longa duração (i.e., estado estacionário) dos clientes requisitando segmentos em cada taxa de transmissão. Novamente, nota-se que essas probabilidades são diferentes entre os grupos e refletem diretamente a qualidade média experimentada nos grupos: **(1)** as probabilidades de longa duração são majoritariamente concentradas em taxas mais baixas para os grupos **-*-BQ* (especialmente para clientes móveis); **(2)** em taxas elevadas para os grupos **-*-AQ* (especialmente

para clientes fixos); e **(3)** são distribuídas de maneira uniforme ao longo de diferentes taxas para grupos **-*-VQ*.

Ao comparar as Figuras 6.10 e 6.11 com o modelo único das Figuras 6.3 e 6.2, é possível verificar que os modelos especializados podem oferecer uma descrição mais acurada da dinâmica de adaptação de taxa dos clientes. Por outro lado, o modelo único captura um comportamento médio e é menos eficaz na tarefa de descrever comportamentos específicos encontrados nos clientes. Pelo fato da dinâmica de adaptação exercer um papel central no engajamento dos usuários, acredita-se que os modelos especializados podem contribuir para a geração de cargas de trabalho mais realísticas, que podem portanto dar melhores subsídios às decisões de provisionamento de recursos dos serviços de vídeo.

6.4 AdpGen: um Gerador Cargas Sintéticas para Vídeos Adaptativos Ao Vivo

Esta seção apresenta uma avaliação do desempenho dos modelos de comportamento de clientes propostos neste capítulo. Seu objetivo é avaliar se, de fato, os modelos especializados são capazes de produzir cargas de trabalho mais realísticas. Foram desenvolvidos dois geradores de cargas sintéticas com base nos parâmetros apresentados nas seções anteriores. Ou seja, um dos geradores usa os parâmetros do *modelo único* enquanto o outro usa os do *modelo especializado em comportamento*. O modelo especializado foi nomeado como *AdpGen*.

Para avaliar a acurácia de geração dos dois modelos, foram geradas 20 cargas de trabalho sintéticas (10 de cada modelo), além de 10 amostras da carga real. Foi utilizado o mesmo conjunto de sementes aleatórias para a geração dos três conjuntos e cada amostra possui ≈ 2 milhões de clientes. O objetivo dos três conjuntos de cargas é a delimitação dos intervalos de confiança das métricas usadas para comparação[4] de precisão dos modelos avaliados em relação aos dados reais. A avaliação de precisão dos modelos se deu em dois níveis de granularidade:

- **Granularidade de sistema/provedor**: neste nível, as cargas são comparadas com base no somatório do *número de sessões*, *on-time*, *off-time* dos clientes. O critério de comparação é a proximidade do intervalo de confiança com o intervalo das amostras

[4]Confiança de (95%) de dois lados, usando a distribuição de *Student t*. O intervalo estimado para a soma Y é $\pm t(1-\alpha/2, N-1)\frac{s}{\sqrt{n}}$, onde s é o desvio padrão, N é o tamanho da amostra (números de execuções do gerador e $t(1-\alpha/2, N-1)$ é o $100(1-\alpha/2)$-percentil da distribuição t com $N-1$ graus de liberdade.

dos dados reais, isto é, quanto mais próximo o intervalo de uma métrica das cargas sintéticas está em relação a mesma métrica no intervalo para as cargas reais, mais realística são os dados sintéticos. Neste nível, é possível ter uma visão global da precisão de cada gerador.

- **Granularidade de cliente**: neste nível, é comparada a função de massa de probabilidade (PMF) do *número de sessões*, e as funções de densidade de probabilidade (PDF) do *on-time total*, *off-time total* e *taxa de transmissão média* dos clientes.

  A métrica utilizada para comparar a similaridade das distribuições da carga sintética e a carga real é a divergência de Kullback-Leibler (ou entropia relativa). Ela quantifica a quantidade de informação perdida quando se usa uma distribuição Q para aproximar os fenômenos descritos por P. Sua faixa de valores vai de 0 (idêntico) a 1 (diferença máxima), o que permite obter diretamente a porcentagem de similaridade entre cadeias. A formula para a divergência de Kullback-Leibler é $KL(P||Q) = \sum_x P(x) \times \log(\frac{P(x)}{Q(x)})$. Vale notar que $KL(P||Q) \neq KL(Q||P)$, isto é, há uma assimetria no cálculo da divergência de Kullback-Leibler. Por essa razão, foi utilizada uma versão simétrica dessa medida, chamada de distância de Jensen-Shannon. Sua fórmula é $JS(P||Q) = \sqrt{0.5 \times KL(P||M) + 0.5 \times KL(Q||M)}$ onde $M = 0.5 \times (P + Q)$. Neste nível, é possível verificar a precisão do gerador de maneira mais individualizada, para cada cliente. As curvas foram geradas usando todas as amostras.

6.4.1 Resultados de Execução dos Simuladores

A avaliação no nível de sistema procura estudar a precisão da carga do ponto de vista de sistema, isto é, em relação ao somatório do tempo de permanência e ausência dos clientes, bem como seus números de sessões.

Tabela 6.1: Intervalos de confiança das amostras reais e dos modelos apresentados

	Carga Real	Modelo Único	Modelo AdpGen
On-time (hs.)	1.569.242 a 1,571,549	1.764.415 a 1.768.368	1.633.965 a 1.638.081
Off-time (hs.)	438.328.8 a 439.884,4	597.855,1 a 600.206,7	436.892 a 439.339,6
Qtd. sessões	3.126.898 a 3.130.622	3.472.495 a 3.476.087	3.150.291 a 3.154.286

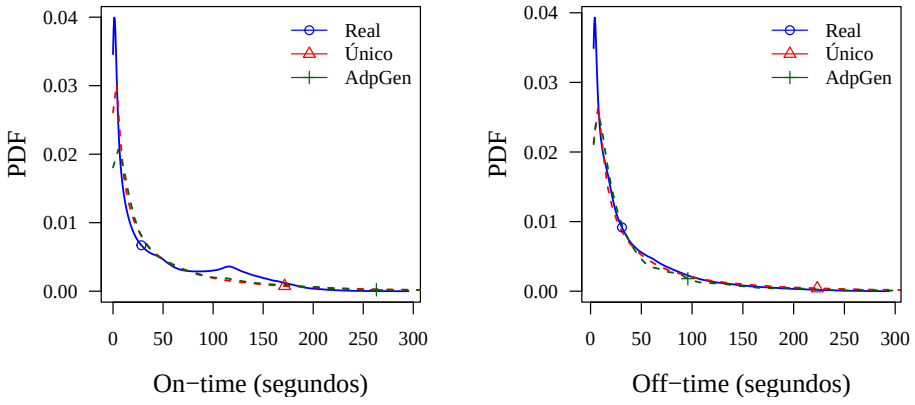
Fonte: Elaborado pelo autor.

A Tabela 6.1 mostra os intervalos de confiança das amostras sintéticas e das amostras dos dados reais. É possível notar que, para as três métricas, há uma aproximação maior dos intervalos do *AdpGen* em relação aos intervalos reais, se comparado com o

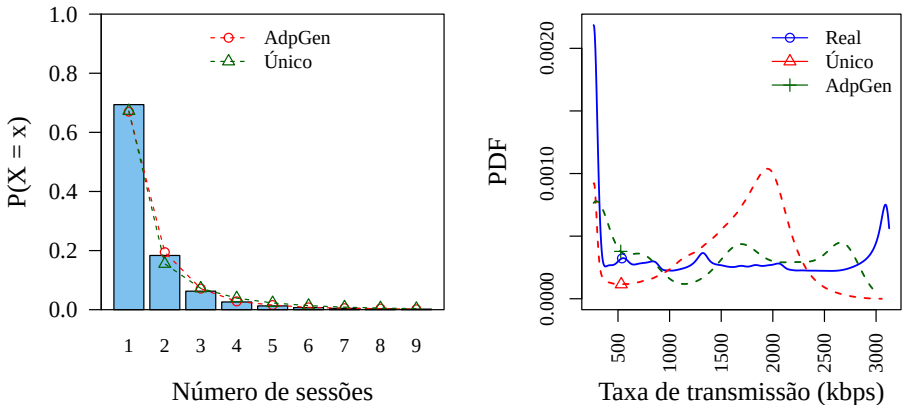
modelo único. Há inclusive uma interseção de intervalos, como no caso do *off-time*, o que evidencia uma chance de que a média do tempo total de ausência seja igual para os dois conjuntos de amostras. Já o modelo único, por outro lado, apresenta intervalos mais distantes, com maior superestimação em todas as métricas. Com isso, percebe-se que o uso de modelos especializados é capaz de produzir uma carga total mais similar ao conjunto de dados real.

Figura 6.12: Distribuições relativas à carga real e geradores (nível de clientes)

(a) *On-time* (b) *Off-time*



(c) *Quantidade de sessões* (d) *Tx. transmissão média*



Fonte: Elaborado pelo autor.

A avaliação no nível de cada cliente permite investigar como os modelos especializados melhoram a representação do comportamento dos clientes individualmente.

A Figura 6.12 apresenta as distribuições referentes às métricas escolhidas. Novamente é possível constatar evidências de que o *AdpGen* foi capaz de produzir cargas com valores mais próximos aos observados na amostra de carga real. Essa conclusão é particularmente evidente para o *off-time*, onde a medida de Jensen-Shannon foi de $0,7451$ e $0,5861$ para o os geradores baseados no modelo único e *AdpGen*, respectivamente. Ou seja, o *AdpGen* tem uma similaridade maior com a carga real, pelo fato de sua medida

estar mais próxima de zero. Isso também acontece para a quantidade de sessões ($0,0699$ no modelo geral e 0.0267 no *AdpGen*) e taxa de adaptação média (0.181 no modelo geral contra 0.0992 no *AdpGen*). No entanto, foi registrado um desempenho melhor para o modelo único no caso do *on-time* (0.7029 contra 0.767 do *AdpGen*). Contudo, na nossa avaliação, acreditamos que esse ganho é compensado pelo ganho do *AdpGen* nas demais métricas, que foi mais significativo.

6.5 Sumário das Contribuições

Este capítulo explorou o objetivo de pesquisa 2, referente a modelagem de comportamento de clientes em sistemas de vídeo adaptativo ao vivo na Internet. O capítulo contribui nesse sentido apresentando a proposta de um novo modelo de comportamento multicamadas, que abarca as decisões de adaptação dos clientes ao longo de suas sessões, além dos seus tempos de permanência e ausência e número de sessões.

Outra contribuição foi a caracterização do impacto do desempenho de transmissão no comportamento dos clientes. Foi mostrado que clientes com níveis de desempenho diferentes têm perfis de engajamento diferentes (*on-time*, *off-time* e número de sessões) e adotam diferentes estratégias de adaptação para preservar a qualidade de reprodução. Essa caracterização deu origem à modelos especializados de comportamento que, por meio de avaliação de desempenho, se mostraram mais precisos e produziram cargas de trabalho sintéticas mais realistas, se comparado com o modelo único de comportamento, que desconsidera o impacto do desempenho no comportamento.

As contribuições produzidas neste capítulo, embora baseadas num conjunto de dados específico, podem ser estendidas a qualquer transmissão adaptativa, bastando fazer um novo ajuste de parâmetros que se adeque ao contexto de utilização. Mesmo em cenários onde os parâmetros não foram especificados, é possível usar o modelo proposto como ponto de partida, a fim de obter uma compreensão básica de comportamento. Assim, com base no exposto, acredita-se que este estudo avança na compreensão do comportamento de clientes em vídeos adaptativos, o que pode auxiliar o provedor a criar abordagens de alocação de recursos mais compatíveis com os requisitos de desempenho de seus usuários.

Capítulo 7

Modelos de Engajamento para Vídeos Adaptativos Ao Vivo

Este capítulo aborda o objetivo de pesquisa 3 (OP3), que trata do desenvolvimento de modelos descritivos e preditivos de engajamento para transmissões de vídeo adaptativo ao vivo.

Entender as relações entre desempenho de transmissão e engajamento pode ajudar a embasar decisões de restrição de recursos que afetem menos o engajamento dos usuários. É possível, por exemplo, empregar restrições mais severas em usuários mais tolerantes, enquanto destina mais recursos àqueles que exigem desempenho mais elevado.

Apesar dessa potencial vantagem, adquirir uma compreensão abrangente da correlação entre desempenho e engajamento exige a consideração de uma série de questões. Em primeiro lugar, muitas métricas de desempenho importantes para explicar o engajamento são de difícil aquisição. Isso ocorre porque estão associadas ao funcionamento interno do cliente. Um exemplo é a latência de inicialização e a taxa de congelamentos, que estão relacionadas ao *buffer* da aplicação cliente. Receber esses dados em massa de todos os clientes, a fim de treinar modelos de engajamento, pode agravar cenários de congestionamento na rede. Além disso, coletar dados do cliente pode suscitar problemas de privacidade.

Em segundo lugar, apesar de os Capítulos 5 e 6 terem mostrado uma caracterização bastante detalhada da relação entre desempenho de transmissão e engajamento, na prática não é possível enumerar todos os aspectos dessa relação a partir de uma exploração manual. Isso só é possível por meio de modelos automatizados de regressão e classificação. No entanto, ainda não existem modelos e/ou um conjunto de métricas que combinadas são capazes de descrever ou prever com a alta acurácia o engajamento do usuário [9].

Assim, tendo em mente essas limitações, e observando a relação entre adaptação e o engajamento descrita nos Capítulos 5 e 6, projetou-se um novo tipo de abordagem para modelos descritivos e preditivos. Essa abordagem utiliza uma combinação de métricas derivadas da mecânica de adaptação do cliente para explicar o engajamento do usuário. Esses aspectos estão disponíveis a partir dos *logs* do provedor e diminuem a necessidade de extração de dados dos clientes.

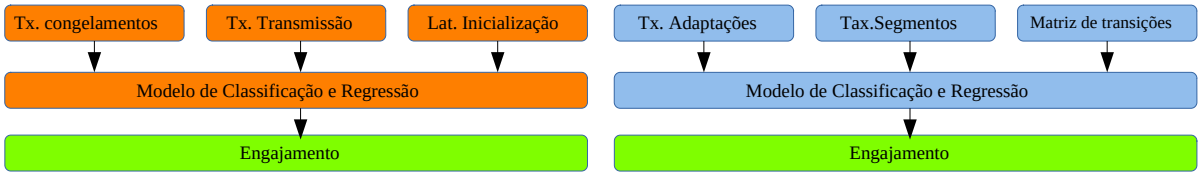
O capítulo está dividido da seguinte forma. Na Seção 7.1, é apresentada a proposta do novo conjunto de métricas de desempenho, associadas a adaptação do cliente, e a motivação para seu uso. Na Seção 7.2, por sua vez, é descrito o ferramental teórico utilizado para construção e treinamento do modelo proposto. Já na Seção 7.3, é apresentado um comparativo de precisão entre o modelo descritivo com métricas clássicas de desempenho e o modelo descritivo com métricas associadas a adaptação, enquanto a Seção 7.4, a partir da constatação da melhor performance produzida pelo modelo descritivo associado a adaptação, passa a discutir seu benefício para a tarefa preditiva, e avalia as vantagens da construção de sub-modelos especializados em cenários de desempenho, um conceito proposto no Capítulo 5. Por fim, na Seção 7.5, são sumarizadas as contribuições do capítulo.

7.1 Modelando Engajamento com base na Adaptação

Como mostra a Figura 7.1, modelos tradicionais de engajamento são baseados principalmente em métricas relacionadas ao *buffer* do cliente, como sua taxa de congelamentos e latência de inicialização [8, 114]. Como já destacado, a aquisição dessas métricas é dificultada pelo fato de precisarem de dados dos clientes.

Por outro lado, no caso do modelo proposto, que é baseado em adaptação, o único requisito é conhecer a taxa de transmissão dos segmentos requisitados pelos clientes. A partir desta informação, é possível gerar todas as métricas necessárias, que são: (1) a *taxa de adaptações negativa e positiva*, (2) *fração de segmentos em cada taxa de transmissão* e (3) a *matriz de transições que descreve o regime de adaptação dos clientes*.

Figura 7.1: Diferença nas métricas do modelo tradicional e baseado em adaptação



Fonte: Elaborado pelo autor.

A motivação para o uso da adaptação, e em particular a matriz de transição como métrica para o modelo de engajamento, surgiu a partir das observações apresentadas no Capítulo 6, que mostram que usuários com diferentes níveis de engajamento possuem também diferentes regimes de adaptação.

7.2 Aspectos Teóricos do Processo de Modelagem

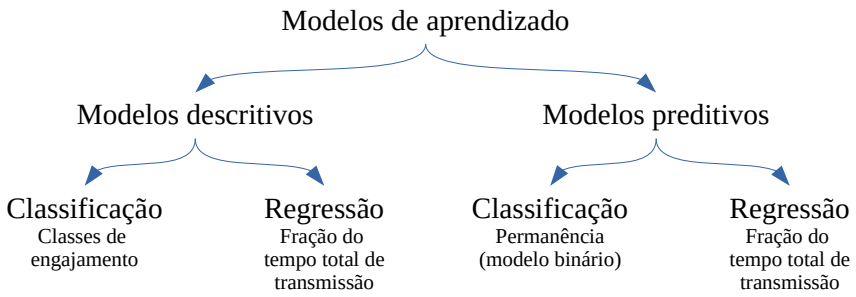
O modelo de engajamento foi implementado nesta tese em duas formas distintas, que são a forma descritiva e preditiva, conforme detalhado a seguir:

- **Modelo de descrição de engajamento**: um modelo descritivo é um modelo basilar, que permite inferir a viabilidade de se determinar o engajamento a partir de atributos que descrevem uma sessão. Ele toma variáveis, como por exemplo métricas de desempenho de transmissão de sessões já finalizadas e constrói uma combinação dessas métricas que permite descrever o engajamento em uma sessão. Um exemplo pode ser encontrado em [8].
- **Modelo de previsão de engajamento**: esse modelo tem a capacidade de estimar o tempo restante de engajamento com base no desempenho histórico de sessão. Ou seja, ao contrário do modelo descritivo, são usados os dados até o momento da previsão, e não os dados de uma sessão finalizada. Assim, trata-se de um modelo mais complexo, em que se espera uma acurácia menor, se comparado à contraparte descritiva.

7.2.1 Classificação vs. Regressão

O modelo de engajamento, implementado seja por meio de uma previsão ou descrição, pode ser abordado como uma *classificação* ou como uma *regressão*. Nesse sentido, a Figura 7.2 apresenta as variações do modelo e as formas como o engajamento foi considerado. Após a figura, é apresentada uma descrição detalhada de cada tipo de modelo.

Figura 7.2: Tipos de modelos de engajamento



Fonte: Elaborado pelo autor.

Classificação

Para o caso do modelo de engajamento sendo formalizado por meio de uma tarefa de *classificação*, o objetivo é obter uma função f que mapeie um conjunto de variáveis independentes X em uma variável dependente Y *discreta*. Em outras palavras, a variável dependente (engajamento) é discretizada em *classes* e objetiva-se determinar a qual classe uma sessão pertence. Neste caso, a eficácia do modelo está relacionada ao *número de vezes em que o previsor determina corretamente a classe de engajamento de uma sessão*.

Foram considerados dois conjuntos de classes diferentes, uma para o modelo descritivo e outra para o preditivo. Para os modelos descritivos, as classes são definidas considerando o engajamento normalizado pelo tempo total de transmissão, observando outros trabalhos como [8]. Por exemplo, em uma classificação considerando 4 classes, o engajamento de uma sessão pode estar nos intervalos {*0-25%, 26-50%, 51-75% e 76-100%*}. Ou seja, uma sessão com permanência de 30 minutos em uma transmissão de 2 horas tem engajamento de 25%.

Para o modelo de previsão, foi pensado um problema binário, no qual o objetivo é prever se um cliente vai permanecer na transmissão pelos próximos n minutos com n variando de 1 a 5. Nesse caso as classes são {*sim, não*} para cada n considerado.

Regressão

A segunda forma de solução é por meio de uma *regressão*, onde o objetivo é obter uma função f que mapeie um conjunto de variáveis independentes X em uma variável dependente Y *contínua*. Ou seja, a previsão é feita sobre os dados contínuos do engajamento normalizados pelo tempo total de transmissão, tanto para o modelo descritivo quanto para o preditivo. Em outras palavras, deseja-se prever o tempo total (modelo descritivo) ou restante (modelo preditivo), que o cliente permanecerá no sistema na sua sessão atual. Neste caso, um bom modelo é aquele que consegue prever o engajamento com o *menor erro possível entre valores previsto e real*.

7.2.2 Forma de Aprendizado dos Modelos

Os modelos descritivos e preditivos desta tese são construídos com base em aprendizado supervisionado, isto é, aprendem a relação entre as métricas de entrada e engajamento a partir dos dados de sessões que lhe são submetidas. Mais especificamente, o aprendizado é estabelecido explorando um conjunto de atributos monitorados e coletados durante a sessão do cliente, sejam eles as métricas de desempenho de transmissão (como

em técnicas da literatura [8, 114]) ou atributos do regime de adaptação (proposta deste capítulo – ver Figura 7.1). Em ambos os casos, um *algoritmo de previsão* é empregado para que, por meio de um treinamento, aprenda como combinar as contribuições de cada atributo, de cada sessão, para o engajamento em um modelo $\mathcal{M}$. O algoritmo utilizado nesta tese é mostrado na Seção 7.2.4.

Esse aprendizado é feito em uma parcela dos dados denominada de *conjunto de treinamento*. Esse conjunto reúne instâncias (e.g.; sessões) que são usadas para que o modelo aprenda as correlações entre as *variáveis independentes* (atributos de entrada – no caso métricas de desempenho e adaptação) e a *variável dependente* (variável de saída – engajamento) e assim gerar os modelos $\mathcal{M}$ pretendidos.

O modelo, após sua construção, tem sua eficácia avaliada em um conjunto disjunto do conjunto de treino, chamado de *conjunto de teste*. Ou seja, a admissão é que o conhecimento aprendido no treino é generalizado para outros conjuntos e pode ser usado para descrever e prever engajamento de novos clientes que venham a se juntar a uma transmissão corrente ou futura.

Ainda em relação ao processo de aprendizado, vale destacar que existe mais de uma maneira de dividir um conjunto de dados em conjuntos de treino e teste. No primeiro, conhecido como validação *Holdout*, os dados são divididos em 2 partes, sendo geralmente 70% para treino e 30% para teste. Esse método é recomendado para conjuntos de dados grandes, como no caso desta tese.

O segundo método é conhecido como *validação cruzada com n-subconjuntos*. Nessa abordagem, os dados são divididos em subconjuntos (normalmente 5 a 10), com $n - 1$ conjuntos sendo usados para treino e 1 para teste. Esse processo é repetido variando-se n vezes o conjunto de treino e deixando os restantes como teste. Ao final, um valor médio de precisão é calculado. Essa abordagem é menos sensível a instâncias aberrantes que possam influenciar artificialmente no resultado de precisão. É o método preferível para amostras reduzidas de dados. No entanto ele é mais custoso em termos de gasto de memória principal, sendo mais difícil utilizá-lo em conjuntos muito grandes.

7.2.3 Instâncias dos Modelos

A Seção 7.2.2 mostra que um modelo de previsão aprende a partir de um conjunto de instâncias chamado de conjunto de treinamento e é avaliado em um conjunto de testes. Sendo assim, uma instância de treino/teste é uma unidade fundamental de informação, que contém todas as variáveis independentes e a variável dependente, e que deve possibilitar treinar e testar um modelo de previsão e descrição.

No modelo descritivo as instâncias se referem a totalidade de uma sessão. Com isso, cada sessão dá origem a uma instância no conjunto de dados. Já no modelo preditivo, adotou-se a abordagem de dividir uma sessão em janelas de 1 minuto, onde cada janela armazena o *desempenho acumulado até a janela anterior*. Essa abordagem permite que uma previsão seja efetuada a cada minuto de transmissão. A escolha pelo tempo de 1 minuto para as janelas se deu em virtude de ela oferecer um bom compromisso entre o número de janelas geradas e o tamanho médio das sessões. Usar janelas maiores impossibilitaria a previsão em sessões curtas, que representam uma parte significativa das sessões do conjunto de dados.

Como resultado da estrutura de janelas, é possível tanto prever o tempo restante contínuo de engajamento de um cliente, quanto se ele vai ficar pelos próximos n minutos, com n variando de 1 a 5. Por exemplo, considerando o problema de classificação binária e uma sessão iniciando seu minuto 5, é possível efetuar a previsão se o cliente irá permanecer no minuto 6 com base no desempenho acumulado até o minuto 4.

Estrutura das Instâncias

A Figura 7.1 mostrou um comparativo das variáveis dependentes das instâncias de ambos os modelos. Modelos clássicos de engajamento usam métricas de desempenho de transmissão. Nesse sentido, foram utilizadas a **(1)** *taxa de transmissão média*, **(2)** a *taxa de congelamentos por minuto* e **(3)** a *latência de inicialização*. A literatura mostra que essas métricas possuem grande relação com o engajamento de um usuário [103]. Já o modelo associado ao regime de adaptação, proposto neste capítulo, usa somente dados relativos à dinâmica de adaptação do cliente. São eles a **(1)** *taxa de adaptações positivas e negativas*, **(2)** *distribuição de segmentos de cada taxa de transmissão* e **(3)** *matriz de transição entre taxas de transmissão*. Essas métricas já foram introduzidas no Capítulo 4, Seção 4.4.

Com relação à variável dependente, adotou-se o engajamento relativo, como já foi mencionado anteriormente. Isto é, o engajamento será tempo de sessão normalizado pela duração de transmissão, exceto no caso da classificação do modelo preditivo, que é um modelo binário que determina a presença/abandono do cliente para um futuro próximo na transmissão.

Vale salientar também que, em ambas as propostas, foi incluído como variável independente o *tipo de dispositivo* do cliente, que é uma métrica contextual fortemente associada ao engajamento, como foi apresentado no Capítulo 5.

7.2.4 Algoritmo de Aprendizado

O aprendizado das correlações entre variáveis independentes e dependente é implementado por meio de um algoritmo de aprendizado de máquina. Foram testados vários algoritmos considerados estado-da-arte tanto em classificação quanto em regressão. Em experimentos preliminares foi observado que a melhor relação entre custo de processamento/memória e precisão foi obtida com o modelo *XGBoost* [19].

O *XGBoost* é baseado em árvores de decisão com aumento gradiente (*gradient boosting*). A ideia é usar múltiplas instâncias sequenciais de um modelo fraco, como uma árvore com poucos níveis, para criar um modelo forte no qual cada árvore prevê os erros da árvore anterior. A minimização de erros é feita por gradiente descendente, que é uma técnica amplamente utilizada em diversos algoritmos de aprendizado de máquina.

7.2.5 Processo de Validação de Precisão do Modelo

Modelos de classificação e regressão têm sua precisão avaliada por métricas específicas. No caso da classificação, tem-se que essas métricas avaliam majoritariamente a taxa de acertos nas classes estabelecidas. Por outro lado, a precisão da regressão é medida por meio de métricas que avaliam o erro entre o valor real de engajamento e o valor previsto. As métricas de classificação são as seguintes:

- **Acurácia:** é a fração de previsões certas em relação a todas as previsões. Definida pela fórmula $\frac{TP+TN}{TP+TN+FP+FN}$, onde TP e FP são os verdadeiros e falsos positivos, respectivamente. Já TN e FN corresponde aos verdadeiros e falsos negativos.
- **Medida-F:** é a média harmônica de precisão (P) e revocação (R), dado por $\frac{2*P*R}{P+R}$. A precisão P é a fração de acertos dentre todas as previsões. Já a revocação R é a fração de acertos dentre todos os possíveis. Os valores da medida-F variam entre 0 e 1.

Já na regressão, os valores preditos são contínuos. Desta forma, a precisão é medida pelo erro entre o valor previsto e o valor real, que pode ser estimado por:

- **Erro absoluto médio (MAE):** média das diferenças absolutas entre valor previsto e valor real. É dada pela fórmula $1/n \sum_{i=1}^{n} |y_i - \hat{y}_i|$, onde y_i é o valor previsto e $\hat{y}_i$ o valor real.

– **Raiz do erro quadrático médio (RMSE):** média da raiz do quadrado das diferenças entre valor previsto e valor real. É dada pela fórmula $\sqrt{1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$, onde y_i é o valor previsto e $\hat{y}_i$ o valor real. Essa medida tem a mesma magnitude do MAE, porém é mais sensível à ocorrência de erros maiores.

7.3 Avaliação de Desempenho dos Modelos Descritivos

Esta seção compara o desempenho dos modelos descritivos das duas abordagens, isto é, do modelo clássico, que usa métricas de desempenho de transmissão, e o modelo proposto neste capítulo, que usa atributos relativos ao regime de adaptação do cliente.

A tabela 7.1 mostra os valores de precisão (regressão e classificação) para os modelos descritivos. Como pode ser visto, foram utilizadas diferentes granularidades para a variável dependente da classificação, indo de duas classes ($\{< 50\%, \geq 50\%\}$) até 5 classes ($\{20\%, 40\%, 60\%, 80\%, 100\%\}$). A tabela evidencia que, conforme se adicionam novas classes, reduz-se a precisão de ambos os modelos. Além disso, foi confirmado que a nova proposta não só alcançou, como ultrapassou a acurácia do modelo base, que usa métricas de desempenho de transmissão. Vale ressaltar que a acurácia apresentada pelo modelo tradicional é semelhante a de modelos similares da literatura [8] e também nos estudos anteriores com outros conjuntos de dados [45, 44], o que serve de evidência para confirmar os dados da Tabela 7.1.

Em relação às outras métricas de avaliação, nota-se que a medida F1 é mais alta também para o novo modelo proposto, o que demonstra uma acurácia mais bem distribuída entre todas as classes. Com isso, confirma-se que os atributos relativos à adaptação podem explicar com significativa acurácia o engajamento de um usuário, reduzindo ou mesmo eliminando a necessidade de monitoramento de desempenho de métricas dependentes do funcionamento interno dos clientes, como por exemplo métricas de *buffer* como taxa de congelamentos.

Para a regressão, foi observado um erro (RMSE) de $0,1239$ para o modelo base e $0,0549$ para o modelo baseado em adaptação. Ou seja, houve uma redução de 55% do erro em relação ao modelo tradicional.

Tabela 7.1: Precisão de classificação e regressão (modelos descritivos)

	Modelo clássico		Bas. em Adaptação	
Classificação				
# classes	Acurácia	F1	Acurácia	F1
2	0.6594	0.7851	0.9188	0.9558
3	0.6272	0.7165	0.9169	0.9275
4	0.6114	0.6973	0.9095	0.9173
5	0.6005	0.6836	0.8901	0.8980
Regressão				
	MAE	RMSE	MAE	RMSE
	0.0829	0.1239	0.0319	0.0549

Fonte: Elaborado pelo autor.

7.4 Avaliação de Desempenho dos Modelos Preditivos

A partir da confirmação da melhoria de desempenho obtida pela nova proposta de uso de métricas de adaptação, evidenciado na Seção 7.3, optou-se por focar apenas nessa abordagem para a avaliação dos modelos preditivos.

A Tabela 7.2 apresenta os resultados por horizonte de tempo, isto é, a quantidade em minutos no futuro em que se deseja saber se um usuário vai permanecer ou não. Novamente, nota-se que, à medida que o horizonte aumenta, menor é a acurácia do modelo. Entretanto, um horizonte de tempo de 1 minuto já é significativo, tendo em vista que modelos de previsão de vazão, usados em algoritmos de adaptação trabalham com uma folga muito menor, de segundos no futuro [112, 68].

A tabela mostra que, para 1 minuto, foi possível uma acurácia de 73% e uma medida F1 de $0,7646$, que indica que o modelo consegue uma boa acurácia tanto para prever a continuidade quanto o término das sessões. Com a possibilidade de saber quais sessões irão abandonar em breve a transmissão, é possível empregar medidas que prolonguem o engajamento ou que direcionem recursos às sessões com maior probabilidade de permanência.

Vale ressaltar também a queda de acurácia em relação aos modelos descritivos. Esse fato já era esperado, tendo em vista que o processo de previsão é inerentemente mais complexo que a descrição. Além disso, um cliente pode deixar sua sessão por falta de interesse de seu usuário, mesmo se a sessão dispuser de alta qualidade. A dificuldade de mensurar o impacto do interesse no engajamento contribui para a redução de acurácia.

Tabela 7.2: Precisão de classificação e regressão (modelo preditivo)

Classificação		
Minutos	Acurácia	F1
1	0,7300	0,7646
2	0,7214	0,6298
3	0,7026	0,7540
4	0,7005	0,7476
5	0,6986	0,7437
Regressão		
Engajamento restante	MAE	RMSE
	0,0796	0,1057

Fonte: Elaborado pelo autor.

7.4.1 Especialização por Cenário de Desempenho

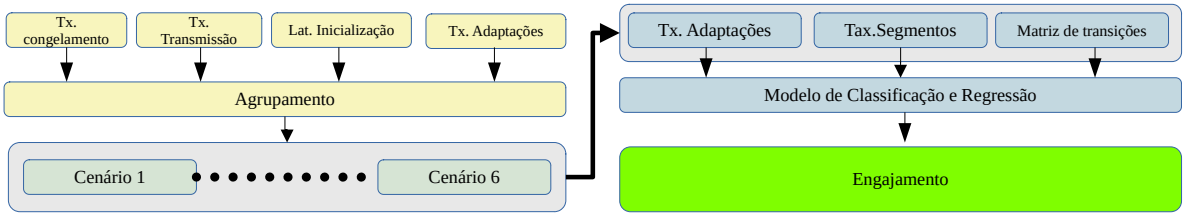
Os capítulos anteriores demonstraram que a modelagem considerando perfis específicos de comportamento ajuda a explicar melhor aspectos relativos à clientes e usuários, em particular o engajamento. Foi assim no Capítulo 5 com os cenários de desempenho, e também no Capítulo 6, que mostrou a presença de múltiplos comportamentos induzidos pelo desempenho experimentado. Assim, com base nessa premissa, esta seção retoma o conceito de cenários de desempenho para avaliar se ele é também benéfico para aumentar a acurácia de modelos de previsão de engajamento.

A abordagem adotada no processo de especialização é construir um modelo híbrido dividido em duas etapas, em que a primeira utiliza as métricas de desempenho de transmissão para agrupamento de janelas de modo similar à criação dos cenários de desempenho, e a segunda, que cria modelos especializados por cenário, utilizando nesses modelos as métricas de adaptação de taxa de transmissão. Ou seja, cada cenário possuirá (1) um centróide, que representa seu desempenho médio e (2) um modelo de classificação binário, responsável por efetuar a previsão levando em conta as características particulares do cenário.

Com o modelo híbrido, passou a fazer parte do treinamento a identificação dos centróides que descrevem os cenários de desempenho. A descoberta dos centróides utiliza aprendizado não-supervisionado, mais especificamente o algoritmo de agrupamento *K-Means*, de forma análoga ao que foi aplicado no Capítulo 5. A diferença reside apenas no fato de que as instâncias são janelas, ao invés de sessões. Após a identificação dos grupos, são treinados modelos especializados de previsão, sendo um para cada grupo/cenário identificado. O processo é ilustrado na Figura 7.3.

Após o treinamento, na etapa de teste, as novas instâncias são agrupadas com base nos centróides identificados no treino e, em seguida, o modelo de previsão associado ao

Figura 7.3: Arquitetura do modelo de previsão baseado em Cenários de desempenho



Fonte: Elaborado pelo autor.

cenário (centróide) de cada instância é usado para executar a tarefa de previsão.

No agrupamento, o espaço euclidiano tem como suas dimensões as métricas de desempenho de transmissão, e a normalização de todas as instâncias, uma etapa requerida pelo *K-means*, foi feita com relação aos máximos e mínimos do conjunto de treino. Assim, tanto o conjunto de treino quanto o conjunto de teste têm valores na mesma escala. Essa abordagem pode produzir valores normalizados fora do intervalo $[0, 1]$. No entanto, ressalta-se que a ocorrência de tais valores é rara e não interfere significativamente no processo de agrupamento. A avaliação do número de grupos (k) foi feita da mesma forma que nos capítulos antecedentes, considerando a minimização da soma das distâncias de todas as instâncias em relação a seus centróides. Com base nesse critério, foi estipulado um $k = 6$.

A Tabela 7.3 mostra os valores de desempenho de previsão para esse modelo. Conforme a tabela mostra, houve um aumento de precisão em relação ao modelo único. No horizonte de tempo de 1 minuto, nota-se que a acurácia foi de 0.73 para 0.80, indicando uma melhora de $\approx$ 10%. Isso mostra que o uso de um modelo para cada cenário de desempenho é vantajoso para melhorar a precisão da previsão de engajamento.

Tabela 7.3: Especialização por cenários de desempenho

Classificação		
Minutos	Acurácia	F1
1	0.8009	0.8206
2	0.7881	0.8111
3	0.7598	0.7902
4	0.7378	0.7684
5	0.7302	0.7641
Regressão		
Engajamento restante	MAE	RMSE
	0.0758	0.1019

Fonte: Elaborado pelo autor.

No caso da medida F1, é possível observar que seu valor foi de 0.76 para 0.82, indicando que o uso de cenários faz com que a acurácia seja melhor balanceada entre as classes. Por outro lado, nota-se que a redução do erro na regressão foi menor, alcançando uma média de $3,5\%$.

7.5 Sumário das Contribuições

Este capítulo apresentou uma nova proposta para modelos descritivos e preditivos de engajamento. Modelos de engajamento utilizam métricas de desempenho de transmissão e contextuais para descrever aceitação. O modelo proposto avança nesse aspecto, ao descrever engajamento a partir do regime histórico de adaptação do cliente, abordagem que apresentou alta acurácia descritiva e preditiva, e pode ser usado em substituição aos modelos tradicionais de engajamento. Além da acurácia, outra vantagem é que o modelo de previsão, em sua forma pura, sem a utilização de cenários de desempenho, não precisa de dados de desempenho, que por sua vez dependem de fatores internos do cliente como seu nível de *buffer*.

O capítulo também mostra que a adição de modelos especializados em cenários de desempenho pode oferecer uma acurácia de previsão ainda maior, reforçando a tese de que a tolerância a um certo nível de desempenho pode variar com o cenário de desempenho global de um usuário, como já visto nos Capítulos 5 e 6.

No próximo capítulo, será proposto um mecanismo para alocação de recursos que visa alcançar um melhor compromisso entre o interesse do cliente, que é ter alto desempenho individual, e do provedor de conteúdo, que é aumentar o engajamento médio (i.e., número de clientes concomitantes e duração de permanência desses mesmos clientes) em suas transmissões. Isso é feito por meio de um esquema que permite a avaliação de diversos cenários de alocação a fim de escolher o que atinge o melhor compromisso entre os interesses descritos. Esse esquema usa, em particular, o previsor de engajamento apresentado na Seção 7.4.1.

Capítulo 8

Uma Proposta de Mecanismo de Alocação de Recursos

Este capítulo aborda no quarto objetivo de pesquisa (OP4), introduzido no Capítulo 1, que trata do desenvolvimento de uma nova abordagem de alocação de recursos para transmissões adaptativas ao vivo na Internet.

Ao longo do Capítulo 5, diversas análises foram apresentadas visando avaliar como os recursos foram distribuídos entre os clientes. Nesse sentido, foi observado que a abordagem para permitir mais clientes concomitantes foi a limitação da taxa de transmissão das sessões, sendo que esta limitação foi feita de maneira generalizada, possivelmente a partir da premissa de que o impacto de tais restrições no engajamento é similar para todos os usuários.

Em oposição a essa premissa, os resultados apresentados no Capítulo 5 apontam para a necessidade de se considerar combinações específicas de valores de métricas, nomeadas naquela ocasião como cenários de desempenho, para que o impacto de uma possível redução de taxa seja melhor compreendido e minimizado. Com base nessa constatação, o Capítulo 7 explora o conceito de cenários de desempenho para a criação de modelos para previsão e descrição de engajamento, que se mostraram mais precisos e com mais acurácia se comparado à modelos únicos.

Com o auxílio da variante preditiva do modelo proposto no Capítulo 7, esse capítulo apresenta um *mecanismo de alocação de recursos*, que visa efetuar uma intervenção nas decisões de adaptação do cliente, atentando-se ao interesse do seu usuário, que é ter um desempenho de transmissão mínimo, porém com foco especial no interesse do provedor de conteúdo, que é conciliar o engajamento dos seus usuários com a redução do consumo de recursos, abrindo assim espaço para a entrada de novos clientes. Mais especificamente, considerando um instante de tempo t qualquer, o objetivo é *aumentar o número de clientes concomitantes em t em relação à abordagem adaptativa tradicional, onde apenas o cliente tem controle de seu mecanismo de adaptação.*

Esse mecanismo serve como uma prova de conceito que visa validar a hipótese geral da tese, introduzida no Capítulo 1. O funcionamento básico do mecanismo se dá da seguinte forma: a cada minuto, são enviados a um previsor de engajamento os valores das

métricas de desempenho de transmissão e de adaptação do cliente. Esse previsor efetua a seleção de um regime de adaptação que atenda especificamente ao interesse central do provedor, que é aumentar o aproveitamento dos recursos disponíveis em sua infraestrutura (i.e., alcançar mais clientes simultâneos).

O objetivo proposto de atingir um maior número de clientes concomitantes pode ser alcançado de duas formas complementares: **(1)** pela redução dos recursos consumidos pelos clientes correntes garantindo que o seu engajamento não seja afetado, o que abre espaço para a admissão de novos clientes enquanto mantém os clientes atuais ativos, e **(2)** pela prevenção de abandono de clientes prestes a sair devido ao baixo desempenho de transmissão.

Com base em nossa revisão da literatura, apresentada no Capítulo 3, foi constatado que essa é a primeira iniciativa no sentido de abordar preservação de engajamento dentro do estudo do conflito de interesses entre usuários e provedores de conteúdo na utilização de recursos de transmissões adaptativas. Outros trabalhos na literatura focam na melhoria de métricas particulares [83, 64, 65] ou abordam a melhoria de engajamento sem se preocupar diretamente com o conflito de interesses mencionado [8].

Este capítulo está organizado da seguinte forma. Na Seção 8.1, é formalizado o problema do conflito de interesses entre clientes e provedores de vídeo adaptativo via internet. Já na Seção 8.2, é revisitado o processo de alocação de recursos do evento abordado nesta tese e discute seu impacto no engajamento e no desempenho de transmissão, enquanto na Seção 8.3, é apresentada a proposta do novo mecanismo de alocação de recursos sensível a engajamento, que usa o modelo preditivo desenvolvido no Capítulo 7. Na Seção 8.4, por sua vez, é detalhada a simulação do mecanismo. Na Seção 8.5, são apresentados os ganhos de engajamento e recursos obtidos por meio do uso do mecanismo proposto. Finalizando o capítulo, na Seção 8.6 é apresentada uma discussão sobre os desafios relacionados à implementação real do mecanismo e na Seção 8.7 é mostrado o sumário dos resultados produzidos.

8.1 O Conflito de Interesses no Uso de Recursos

Esta seção formaliza o conflito de interesses entre usuários e provedor de conteúdo em relação ao consumo de recursos de uma transmissão adaptativa ao vivo pela Internet.

Considere um provedor de vídeo P com uma largura de banda total L. Considere ainda o conjunto C_t de clientes ativos no instante t, sendo servidos por este provedor. Em cada instante t da transmissão, L está distribuída entre o conjunto de clientes C_t na forma de um conjunto de segmentos $S(C_t)$ em que cada segmento $s \in S(C_t)$ pertence a

uma das taxas de transmissão disponíveis na transmissão.

Em qualquer desses instantes de tempo t, a largura de banda referente ao somatório das taxas de transmissão dos segmentos em $S(C_t)$ isto é, $L_{S(C_t)}$, atende à restrição $L_{S(C_t)} \leq L$. Nesse sentido, clientes e provedor têm interesses distintos quanto a utilização de L:

- É de interesse do provedor de conteúdo aumentar seu conjunto C_t, sendo que o tamanho C_t depende das taxas de transmissão de cada $s \in S(C_t)$. Isto é, segmentos de baixa taxa permitem um C_t maior, já segmentos de alta taxa diminuem o tamanho de C_t.
- Já cada usuário individual, por sua vez, irá preferir a maior taxa compatível com seu cliente, uma vez que seu engajamento têm correlação positiva com a taxa de transmissão, como evidenciado pelo Capítulo 5. Assim, para maximizar o engajamento de todos os usuários, basta sempre atribuir a maior taxa de transmissão como requisitado pelos clientes.

O cenário exposto revela um potencial conflito de interesses entre clientes e provedor de conteúdo. Isto é, se a taxa de transmissão dos clientes aumentar, haverá uma melhora de desempenho de transmissão em cada cliente com impacto positivo para o engajamento, em detrimento de uma redução de C_t. Por outro lado, se diminuirmos a taxa de transmissão dos clientes em C_t, haverá um aumento dos recursos disponíveis que permite novos acessos ao conteúdo, ao preço de que os clientes atuais possam ter uma redução no desempenho de transmissão e, consequentemente, pior engajamento, isto é, permaneçam por menos tempo na transmissão. Ou seja, a possível entrada de novos clientes pode não compensar as saídas produzidas pela queda de desempenho.

Num cenário em que C_t requer um $L_{S(C_t)} \leq L$, fica fácil atingir esse compromisso, bastando disponibilizar a maior taxa compatível com cada cliente. Já nos casos em que são necessárias restrições de recursos em $S(C_t)$, como quando C_t requer um $L_{S(C_t)} > L$, ou quando se deseja economizar recursos, é desejável empregar políticas de restrição que tenham baixo impacto sobre o engajamento de usuários dos clientes C_t. Nesse sentido, uma abordagem para a determinação do melhor $S(C_t)$ envolve o emprego de padrões personalizados de realocação, a partir da avaliação prévia de múltiplos cenários de redução de desempenho e seus custos para o engajamento de cada usuário.

Para medir o custo de uma redução de desempenho no engajamento é utilizado um modelo preditivo. A partir dele, é possível antecipar o resultado de decisões de alocação e adaptação de taxa de transmissão e escolher um conjunto de decisões que seja mais sensível ao engajamento dos usuários, se comparado à observada nos dados originais, aumentando assim C_t.

Nesta tese, as novas decisões de realocação podem aumentar C_t de duas formas complementares: **(1)** conciliação entre redução de consumo de recursos e manutenção de engajamento ou **(2)** estabelecimento de um conjunto de decisões de adaptação que prolongue o engajamento dos usuários prestes a abandonar.

8.2 A Alocação de Recursos no Sistema Alvo

Ao longo das transmissões estudadas, observou-se que jogos de grande audiência registraram sessões com taxas de transmissão significativamente mais baixas, com uma alta correlação negativa entre o número máximo de clientes simultâneos e a taxa de transmissão média oferecida por cliente (Spearman $\rho = -0,91$), em particular para clientes fixos, como foi mostrado na Seção 5.2.1.

Uma visão mais geral a respeito da relação entre carga e taxa de transmissão pôde ser vista na Figura 5.4. Essa figura mostra que transmissões como por exemplo a do jogo França vs Bélgica, que atraiu grande audiência, apresentaram uma taxa de transmissão mais baixa em relação às outras partidas com menor número de clientes.

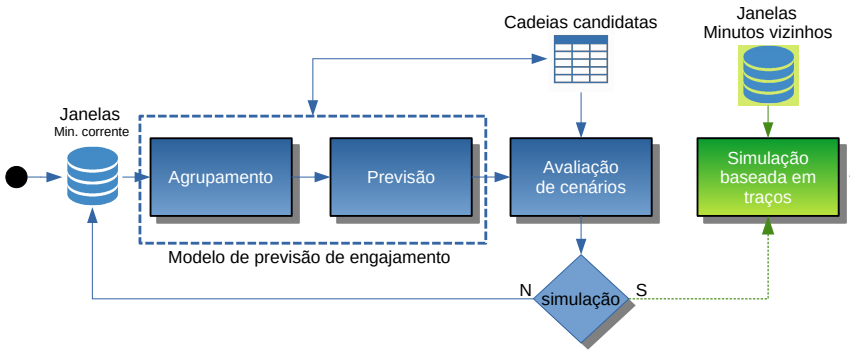
Assim, por meio desses dados, pode-se observar uma possível correlação entre o número de clientes concomitantes e o desempenho, representado neste caso pela a taxa de transmissão, de cada um deles. Em outros termos, há evidências que sustentam a hipótese da existência de uma política de alocação de recursos, em algum ponto da infraestrutura de distribuição, que tende limitar o desempenho de transmissão com o possível objetivo de permitir que todos os clientes possam ter acesso ao conteúdo oferecido, particularmente em cenários com grande número de acessos. Por outro lado, conforme foi apresentado no Capítulo 1, Seção 1.1, houve um significativo aumento na taxa de abandonos de usuários durante o pico de acessos das transmissões, o que indica que a abordagem de limitação global de taxa de transmissão pode não ter sido eficiente para abrir espaço para novos clientes enquanto mantinha o engajamento de clientes atuais.

Curiosamente, o mesmo tipo de restrição de taxa de transmissão não foi observada para clientes móveis (Figura 5.7), que experimentaram uma taxa de transmissão muito mais estável, sendo essa observação baseada na correlação entre o pico de clientes e a taxa de transmissão média (Spearman $\rho = -0,38$). No caso desses dispositivos, a existência de menores taxas pode se dever a uma limitação natural de hardware e conectividade. Além disso, como mostrado na análise de cenários de desempenho (Seção 5.4), clientes móveis são mais flexíveis quanto a reduções de taxa de transmissão, e preservam melhor seu engajamento nessas ocorrências.

8.3 Arquitetura do Novo Mecanismo de Alocação

O mecanismo de alocação de recursos proposto neste capítulo tem o objetivo de aprimorar a alocação de recursos em transmissões adaptativas de vídeo ao vivo. A hipótese explorada é que uma alocação que leve em conta os requisitos particulares de desempenho dos usuários é capaz de preservar melhor seus engajamentos se comparado a um modelo de restrições gerais de recursos, aumentando assim o número de clientes simultâneos em qualquer instante t da transmissão.

Figura 8.1: Arquitetura do mecanismo de alocação



Fonte: Elaborado pelo autor.

A Figura 8.1 delineia a arquitetura do mecanismo proposto. Ela mostra que o funcionamento dos módulos se dá da seguinte forma: a cada Δt, com duração de 1 minuto, são coletadas as *sessões em andamento*. Cada sessão é representada por uma *janela*, que armazena um conjunto de dados históricos da sessão que incluem o regime de adaptação, dados contextuais, como sistema operacional, dispositivo e métricas de desempenho de transmissão e métricas de adaptação. Como já mencionado no Capítulo 7, cada minuto de atividade dentro de uma sessão dá origem a uma nova janela, isto é, há uma sincronia entre a geração de janelas e a sua coleta pelo mecanismo de alocação. Nesse sentido, Δt maiores podem ser utilizados. No entanto isso poderia reduzir a utilidade do mecanismo para sessões curtas, que são frequentes dentro das transmissões.

A janela de uma sessão é similar à apresentada na Seção 7.2.3, com a diferença de que, neste caso, os dados se referem somente ao minuto anterior, ao contrário da abordagem de acumular todo o desempenho, desde o início da transmissão, a cada janela. Essa abordagem facilita o cálculo das métricas de desempenho e adaptação em cada janela e não há perda significativa na acurácia de previsão.

Os dados coletados são enviados ao *módulo de agrupamento*, que agrupa as sessões de acordo com seu cenário de desempenho. Em seguida, o *módulo de previsão* é invocado para prever quais clientes irão permanecer e quais irão abandonar suas sessões. Após a

previsão, os dados são submetidos ao módulo de *avaliação de cenários*. Esse módulo é responsável por escolher um novo regime ou arranjo de adaptação para as sessões, a partir de um conjunto de cadeias candidatas. O módulo pode focar nas sessões cuja previsão foi de abandono ou naquelas cuja previsão foi de permanência, de acordo com o *modo de operação* escolhido (Seção 8.3.2).

Por fim, os regimes selecionados pelo avaliador de cenários são enviados aos clientes, que irão orientar seu mecanismo de adaptação conforme o regime proposto. Desta forma, a adaptação é feita de forma a atender os requisitos de desempenho do cliente, mas também considerando os interesses de escalabilidade do provedor.

Vale ressaltar que as sessões que estão em sua primeira janela de avaliação e que, portanto, não possuem dados históricos, podem empregar uma adaptação local temporária, ou receber do provedor uma cadeia de adaptação baseada na média de desempenho de clientes de mesma região geográfica e dispositivo, como por exemplo em [83]. Nesta proposta optou-se pela adaptação inicial local, pois ela permite obter uma melhor estimação da taxa de transmissão inicial do do cliente.

O mecanismo também possui um módulo anexo responsável por fazer uma simulação baseada em dados de transmissões passadas. O módulo está destacado em verde na Figura 8.1 e será detalhado na Seção 8.4. Ele utiliza as informações de janelas dos minutos anteriores ao minuto corrente para estimar, via extrapolação, o impacto de um determinado cenário de alocação, escolhido pelo avaliador de cenários, no engajamento.

8.3.1 Componentes do Mecanismo de Alocação

Conforme mostrado na Figura 8.1, o mecanismo é dividido em módulos sequenciais. A seguir, esses módulos são descritos em detalhes.

- **Agrupador**. O agrupador é responsável por identificar o cenário de desempenho de cada sessão (baseado em sua janela anterior). A associação ao cenário de desempenho é feita partindo da execução de um modelo descritivo dos cenários de desempenho, descrito no Capítulo 7. Especificamente, este modelo consiste de um conjunto de perfis (i.e., centróides) representativos dos cenários de interesse, obtido a partir de um algoritmo de agrupamento (K-Means), conforme descrito no Capitulo 5. Estes perfis são obtidos na etapa de treinamento do modelo. À cada sessão de entrada é associada o centróide que melhor representa o cenário de desempenho refletido na janela atual. Em seguida, as sessões são submetidas ao módulo de previsão. O objetivo da utilização de cenários de desempenho é capacitar o previsor a reagir adequadamente a uma maior

variedade de cenários de carga impostos ao sistema e reduzir a perda de acurácia ao longo do tempo. Ou seja, se o sistema estiver passando por uma sobrecarga e os clientes experimentarem baixo desempenho, então decisões de alocações específicas para esse cenário serão adotadas. Se por outro lado o desempenho for bom, então regimes de adaptação específicos de cenários de alto desempenho serão selecionados.

- **Módulo de previsão**. O módulo de previsão é baseado na abordagem apresentada no Capítulo 7. Ele toma como instâncias janelas de sessão, que contêm os dados de desempenho de transmissão e adaptação do minuto anterior de transmissão a qual a janela foi gerada. Seu aprendizado é feito com um conjunto de transmissões de treino. Uma vez treinado, o modelo pode ser usado em outras transmissões. Os teste feitos em nossos dados demonstra que a previsão em outras partidas atinge níveis de acurácia acima de 76%, o que demonstra boa extrapolação do modelo. Com base nesses dados, o previsor é capaz de prever se um cliente irá permanecer na janela seguinte. Vale recordar ainda que existe um previsor personalizado para cada cenário de desempenho identificado no treinamento.
- **Avaliador de cenários**. Esse módulo é responsável por avaliar diversos regimes de adaptação e escolher algum que seja compatível com o modo de operação do mecanismo de alocação (ver Seção 8.3.2).

  O módulo avaliador de cenários usa como entrada uma tabela de cadeias candidatas. Essa tabela contém todos os regimes de adaptação encontrados no conjunto de treino. Ela é usada como conjunto de entrada para auxiliar na decisão de quais cadeias de adaptação serão enviadas aos clientes. A tabela passa por uma filtragem de acordo com a taxa de transmissão histórica do cliente e, em seguida, é determinado pelo previsor de engajamento às cadeias cuja previsão é de permanência no minuto seguinte. Qualquer uma das cadeias pode ser escolhida. Se a ideia é economia máxima, então um regime de baixa taxa de transmissão pode ser selecionado. Se o objetivo é prolongar o engajamento, então pode ser selecionado um regime com taxa de transmissão próxima à da janela anterior, que aumente a chance de permanência do cliente.

8.3.2 Modos de Operação

Conforme introduzido, há duas maneiras de se abordar o objetivo proposto que é o aumento de clientes simultâneos no sistema. A primeira maneira consiste em mitigar o abandono dos clientes, mantendo seu consumo de recursos, fazendo assim com que haja mais clientes simultâneos. A segunda maneira trata da redução do consumo de recursos

dos clientes correntes sem afetar o engajamento dos seus usuários. Como efeito, isso abre uma margem de recursos adicionais para permanência concomitante de mais clientes.

No mecanismo apresentado neste capítulo, as duas formas de abordagem são contempladas. Isto é, o ele pode ser configurado para trabalhar sobre os clientes que cuja previsão é de abandono iminente (aumento de engajamento) ou sobre o clientes cuja previsão é de permanência (economia de banda). O modo de operação escolhido é empregado na etapa de previsão e avaliação de cenários, descrito na Figura 8.1. A seguir são descritos os dois modos de operação considerados:

- **Modo de aumento de engajamento**. Nesse modo, são selecionados os clientes cuja previsão de abandono é iminente, conforme indicado pelo módulo de previsão. Esses clientes são enviados ao avaliador de cenários para que ele selecione um regime de adaptação compatível com a taxa de transmissão recente do cliente, que permita um possível prolongamento de sua permanência. Ou seja, a ideia é aproveitar o conhecimento global do modelo de previsão treinado, que aprendeu, com certa acurácia, quais regimes de adaptação são mais propensos a aumentar a possibilidade de um cliente permanecer em sua sessão. Esse modo é adequado para cenários em que o sistema não enfrenta sobrecarga na sua infraestrutura de transmissão, uma vez que o prolongamento da permanência de clientes pode aumentar o consumo geral de recursos.
- **Modo de economia de recursos com preservação de engajamento.** Nesse modo, são selecionados os clientes que permanecerão no sistema na próxima janela. Esses clientes são enviados ao avaliador de cenários para que ele selecione um regime de adaptação que reduza suas taxas de transmissão, com o mínimo de interferência no engajamento dos usuários. Dessa forma o objetivo é estabelecer um mecanismo de redução de recursos que minimize a perda de engajamento. Esse modo é adequado para transmissões que estejam passando por uma sobrecarga ou caso o provedor deseje economizar recursos.

8.4 Simulação do Mecanismo de Alocação

Esta seção descreve o processo de simulação do mecanismo de alocação, provido pelo módulo opcional destacado na Figura 8.1. Esse módulo permite estimar como as escolhas do avaliador de cenários impactam no engajamento, sem que seja necessário executar o mecanismo num ambiente de produção. Para tanto, parte-se da premissa que o resultado de determinada ação do avaliador de cenários pode ser estimado por extrapolação, ou seja, a partir do resultado em instâncias de adaptação similares a que

se deseja avaliar [83, 8, 38, 64]. Ou seja, basta avaliar o engajamento de clientes que utilizaram o mesmo arranjo de adaptação pretendido em um passado próximo.

Por exemplo, supondo que o alocador escolhe um arranjo de adaptação A para um cliente $c \in C_t$. Para saber se A interfere positiva ou negativamente no engajamento de c, basta procurar nos minutos anteriores ao da janela atual de c todos os clientes, que foram submetidos a A. Para extrapolação, foi escolhida uma faixa de tempo de 10 minutos anteriores à janela atual. Esse valor foi escolhido porque, em experimentos preliminares, faixas de tempo menores reduziram a probabilidade de encontrar uma quantidade adequada clientes similares para extrapolação, e faixas maiores não produziram melhoras quanto a essa quantidade.

O critério utilizado para declarar a similaridade entre arranjos de adaptação é a comparação de suas distribuições estacionárias, como empregado em [53, 14]. A distribuição estacionária é o conjunto de probabilidades que ocorre após um longo tempo de execução de uma matriz de transições. Isto é, define o comportamento de uma cadeia de adaptação quando ela atinge um estado de estabilidade nas trocas de taxa de adaptação.

A métrica utilizada para comparação das distribuições estacionárias é a distância de Jensen-Shannon, derivada da divergência de Kullback-Leibler (ou entropia relativa). É a medida mais popular para calcular a similaridade entre matrizes de transição [107, 89, 32, 14]. Ela já foi utilizada nesta tese para comparar os modelos de comportamento de clientes (Seção 6.4), onde os detalhes de seu cálculo foram apresentados.

Após a seleção das sessões similares, é determinado o valor da variável dependente (permanece/não permanece no próximo minuto) por meio de votação, onde a prevalece o valor da maioria dos votantes, do mesmo modo que é feito em abordagens similares, como por exemplo no caso de aprendizado baseado em instâncias como o *K-nearest neighbors* e outros. Vale ressaltar que, em caso de empate, o valor da variável dependente é escolhida aleatoriamente.

Ainda em relação a extrapolação, para garantir que ela fosse mais precisa, optou-se por selecionar os clientes para votação que se submeteram a um A com o mesmo tipo de dispositivo e sistema operacional de $c \in C_t$. Devido a essa decisão, poderá haver casos em que não serão encontrados clientes com o arranjo A similar a de $c \in C_t$ para efetuar tal votação e extrapolação. Por isso foi estabelecido um relaxamento do critério de similaridade que permite que clientes cujo A tenha similaridade de, no mínimo, 95% possam ser usadas para extrapolação. Contudo, se mesmo assim não houver clientes similares para efetuar a votação, então o fluxo de adaptação de $c \in C_t$ é mantido como nos dados originais.

Por fim, após o mecanismo de votação ter sido finalizado, há a troca do arranjo de adaptação original pelo arranjo A mais próximo ao sugerido pelo pelo avaliador de cenários, segundo a distância de Jensen-Shannon, escolhido a partir das instâncias de votação. Adotar o arranjo de adaptação de uma das instâncias de votação, ao invés da

cadeia escolhida na tabela de candidatas, ajuda a evitar o fenômeno de *concept drift*, no qual o sistema é levado a um estado desconhecido, não analisado anteriormente, que pode diminuir a precisão da previsão de engajamento.

8.5 Resultados da Simulação do Mecanismo

Esta seção apresenta o resultado da simulação do mecanismo de alocação. Foram consideradas três transmissões com média de 1,5 milhões de sessões. Além disso, para obter uma estratégia de referência, para fins de comparação, também foi desenvolvida uma versão aleatória dos mecanismo de alocação, onde tanto a variável dependente quanto a janela candidata são escolhidos uniformemente. A versão aleatória foi executada três vezes com diferentes sementes.

Foram simulados os dois modos de operação do mecanismo, a saber o modo de aumento de engajamento (Seção 8.5.1) e o modo de redução de banda (Seção 8.5.2). Nas figuras, os ganhos apresentados em ambas as versões do mecanismo são sempre em relação ao sistema de alocação original, cuja dinâmica foi registrada nos *logs* do provedor, que deixa a cargo dos clientes a dinâmica de trocas de taxa de transmissão.

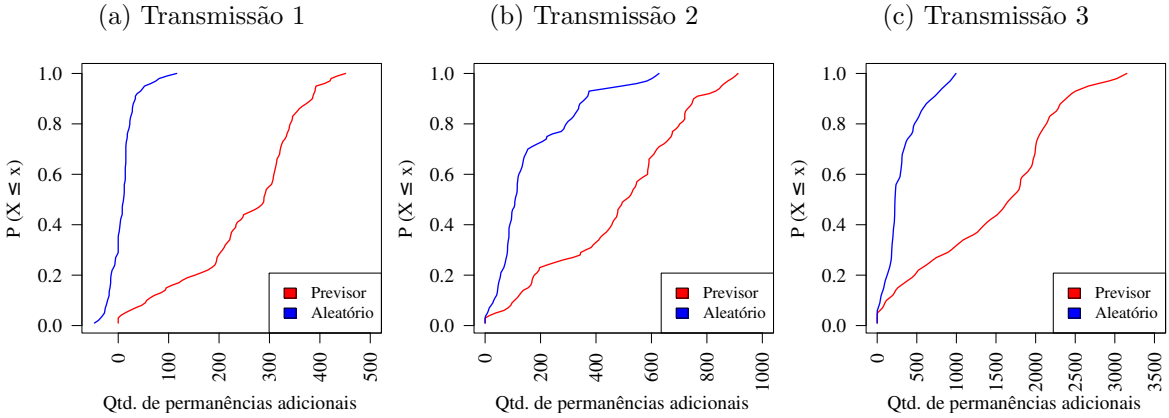
8.5.1 Modo de Aumento de Engajamento

O modo de aumento de engajamento busca mitigar o abandono de usuários. Os clientes são submetidos ao avaliador de cenários para receber novas matrizes de transição, com taxas de transmissão médias similares às originais, mas que usam o conhecimento aprendido sobre desempenho versus engajamento para promover um aumento da probabilidade de permanência.

A Figura 8.2 mostra as funções de distribuição acumulada do ganho em número de sessões por minuto (i.e., sessões que adiaram seu abandono), em relação aos dados originais, obtido com o mecanismo de alocação para as três transmissões consideradas, já descontados os abandonos de sessões provocadas por escolhas equivocadas do avaliador de cenários. Os gráficos apresentam os dados da execução para as duas versões do mecanismo, a saber, a que usa o modelo de previsão de engajamento e a que usa uma escolha aleatória. Em todas as transmissões, é possível notar que o uso do modelo preditivo é mais eficaz que a alocação aleatória. Em particular, ressalta-se que a abordagem

de alocação aleatória chega a registrar uma perda no número de sessões (i.e., valores negativos no eixo x do gráfico) para a transmissão 1 durante 21% do tempo de transmissão (Figura 8.2a). Por outro lado, considerando o uso do previsor de engajamento, foi possível registrar ganhos médios de 262, 459 e 1445 sessões por minuto para as transmissões 1, 2 e 3, respectivamente, se comparado com os dados originais.

Figura 8.2: Aumento de engajamento em número de sessões comparado aos dados originais



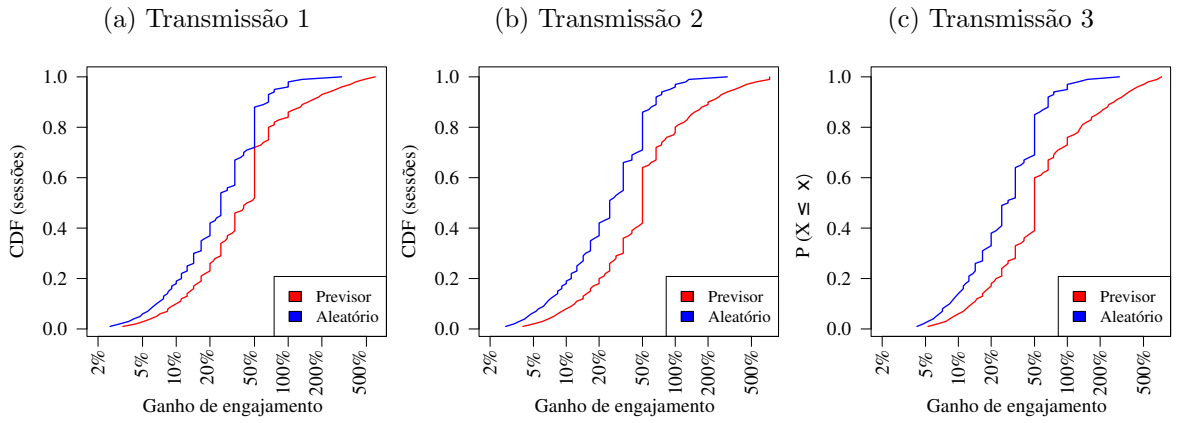
Fonte: Elaborado pelo autor.

Por sua vez, a Figura 8.3 mostra a o aumento percentual do engajamento do usuário (tomando os dados originais como referência) obtida pela alocação associada ao previsor de engajamento versus a abordagem aleatória. Nota-se novamente que com o previsor, há um ganho maior de engajamento nas sessões. Por exemplo, a transmissão 3 registrou um ganho médio de 100% por meio do uso da alocação associada a previsão contra 35,5% da abordagem aleatória. Ainda na transmissão 3, foi registrado que 27% das sessões duplicaram seu engajamento, sendo que apenas 5% das sessões alcançaram a mesma melhoria na abordagem aleatória. Considerando as demais transmissões, foram observados ganhos médios de 94% e 92% para as transmissões 1 e 2, respectivamente, usando a alocação associada ao previsor de engajamento.

8.5.2 Modo de Economia de Recursos

No modo de economia de recursos, a meta é reduzir a largura de banda consumida pelos clientes sem que haja um impacto significativo no engajamento dos usuários, visando alcançar uma maior margem de recursos livres para o aumento do número de clientes atendidos. Durante a execução dos testes, foi registrado que o alocador aleatório leva a uma perda muito expressiva de clientes. Por essa razão, optou-se por focar as discussões

Figura 8.3: Ganho percentual de engajamento por sessão



Fonte: Elaborado pelo autor.

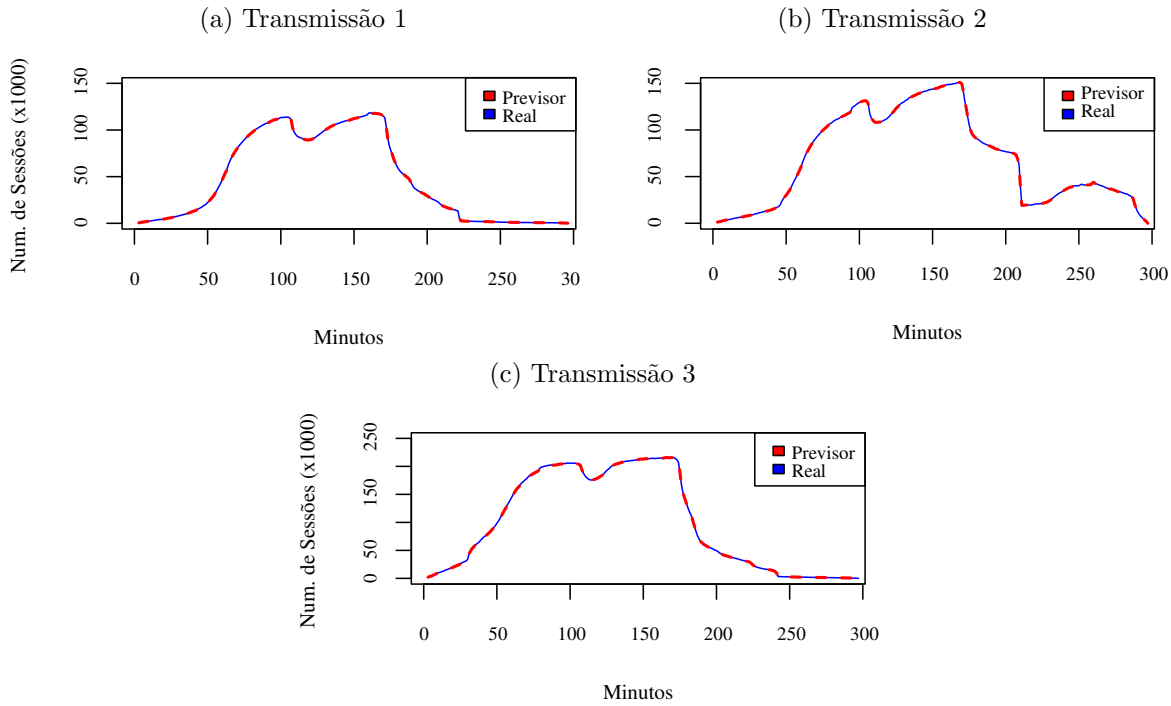
desta sessão aos resultados da versão do mecanismo que usa o previsor de engajamento.

Em primeiro lugar, é necessário verificar se a redução de taxa de transmissão interfere no engajamento dos usuários. Para isso é mostrada na Figura 8.4 a variação ao longo do tempo da carga de trabalho expressa em número de sessões nas três transmissões selecionadas. Como é possível observar, as linhas tracejadas vermelhas, que representam número de sessões admitidas pelo alocador baseado no modelo preditivo, acompanham as linhas azuis, que representam a variação real do número de sessões. As figuras apresentam uma confirmação visual de que há baixo impacto sobre o engajamento dos usuários decorrente da diminuição de banda dos clientes. Em termos quantitativos, a perda média de sessões de clientes que já estavam na transmissão foi de 0,09%, 0,39% e 0,31% por minuto para as transmissões 1,2 e 3 respectivamente.

A Figura 8.5 mostra as funções de distribuição acumulada da economia de banda por minuto obtida com o uso do mecanismo de alocação. Como pode ser visto, há picos de redução de mais de 800 Gigabytes por minuto, com um impacto de menos de 1% no engajamento, como foi mencionado anteriormente. Usando a transmissão 1 como exemplo, tem-se que a redução média de banda é de 28 gigabytes por minuto, o que permitiria a entrada de mais de 71 mil novos clientes na maior taxa de transmissão (3127 kbps), sem que os usuários correntes tenham seu engajamento reduzido significativamente.

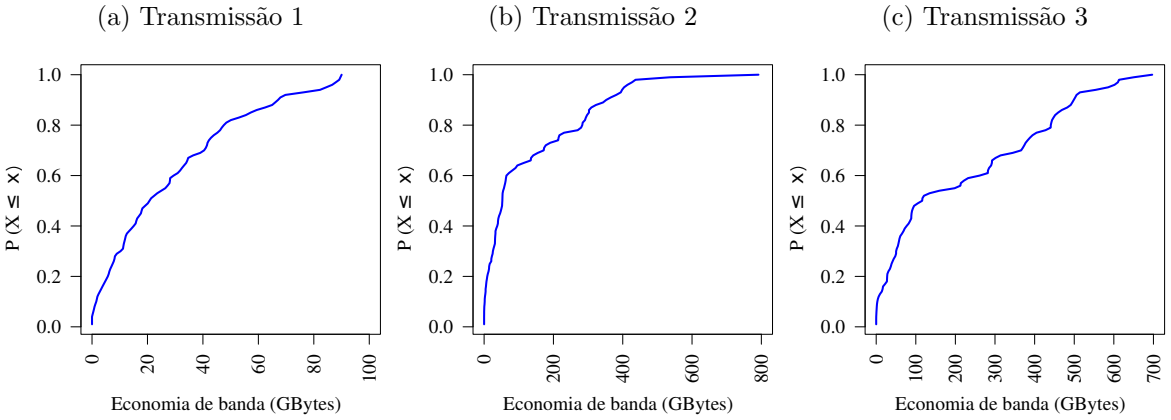
Além do aspecto da economia em si, é relevante analisar o padrão de limitação efetuada pelo alocador para entender como foi possível obter um melhor compromisso entre economia de recursos e preservação do engajamento dos usuários. Para auxiliar nessa análise, a Figura 8.6 mostra a fração de segmentos de cada taxa de transmissão registrada nos dados reais e no mecanismo proposto. Na Figura 8.6, é possível ver que a diferença mais marcante é o fato de a maior parte dos segmentos, que antes eram da maior taxa de transmissão (3127 kbps), passaram a ser da segunda maior taxa (2085 kbps). Nesse sentido, o engajamento foi mantido porque essa queda de taxa pode não apresentar uma

Figura 8.4: Quantidade de sessões ativas durante transmissões (alocador vs. dados reais)



Fonte: Elaborado pelo autor.

Figura 8.5: Largura de banda economizada por minuto nas transmissões



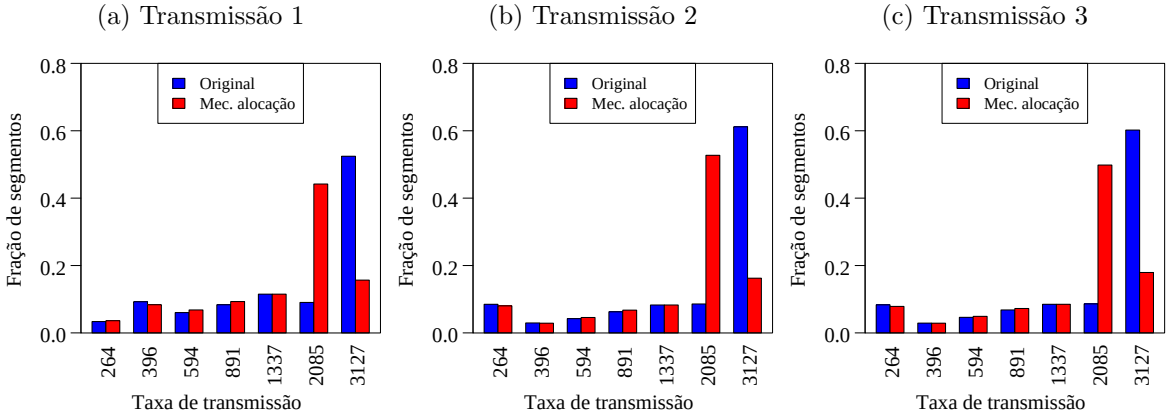
Fonte: Elaborado pelo autor.

diferença substancial na qualidade de imagem percebida pelo usuário, fato que foi capturado pelo avaliador de cenários. Analisando o arquivo *.m3u8* que lista as representações disponíveis para a transmissão (Figura 4.1), é visto que a degradação visual pode de fato ter sido menos evidente, uma vez que as duas taxas de transmissão mencionadas têm a mesma resolução, de 720 linhas horizontais.

Em suma, por meio dos resultados apresentados, é possível obter evidências de que o conhecimento da relação entre desempenho de transmissão, adaptação e o engajamento

pode ser efetivamente explorado em um processo de alocação mais aprimorado, que permite um uso mais racional dos recursos de transmissão, que ao mesmo tempo preservam os requisitos de desempenho do usuários e o interesse do provedor, que é um aumento do número de clientes concomitantes.

Figura 8.6: Distribuição da taxa de transmissão (dados reais e mecanismo de alocação)



Fonte: Elaborado pelo autor.

8.6 Questões Relativas à Implementação

Um dos requisitos não-funcionais do mecanismo de alocação de recursos é que ele deve ser capaz de avaliar as sessões e selecionar as novas cadeias dentro da janela corrente, neste caso 1 minuto, para que essas cadeias possam ser retornadas aos clientes e eles as utilizem para proceder a adaptação de taxa de transmissão. Com base nesse requisito houve o esforço de efetuar o processamento das janelas em paralelo, utilizando todos os núcleos do processador. Além disso, o mecanismo também suporta a utilização de processamento distribuído.

Com base nessa arquitetura, foi efetuada a avaliação utilizado uma estação *desktop* comum, com processador de 12 núcleos e 16 GB de memória, onde foi observado que o processamento das sessões ativas na sequência de módulos descrita ocorre em uma média de 30 segundos, isto é, dentro da janela de 1 minuto corrente. No entanto, apesar da avaliação preliminar se mostrar promissora, vale ressaltar que o mecanismo se trata ainda de uma prova de conceito. Isto é, há ainda questões de implantação que devem ser avaliadas.

Uma primeira questão relacionada à implantação real do serviço é integração do mecanismo com os *logs* de requisição de segmentos do servidor. Essa integração deve

ser feita de maneira a permitir computar rapidamente os dados das sessões dos clientes e criação das instâncias de previsão. Uma abordagem para alcançar esse objetivo seria armazenar as requisições em um banco de dados relacional. Avaliações feitas no âmbito do presente mecanismo de alocação mostraram que o uso de banco de dados com *índices* foi capaz de acelerar a leitura das sessões ativas. Por exemplo, considerando a execução de um comando *SQL* de contagem de números de clientes ativos em cada minuto de transmissão, foi observado que o tempo gasto foi de 6,5 segundos usando índices contra 36 segundos sem indexação, isto é, uma redução de 82% no tempo de processamento. Sendo assim, com base nesse resultado, é possível que essa melhoria possa ser estendida também aos *logs* brutos de requisição de segmentos.

Uma segunda questão está relacionada a escala do mecanismo. Apesar do bom desempenho em uma transmissão individual, é sabido que as plataformas de mídia atuais podem lidar com milhares de transmissões simultâneas. Nesse sentido, uma abordagem que pode ser adotada é a replicação do mecanismo ao longo da infraestrutura de transmissão, hospedando componentes do mecanismo de alocação nos servidores da *CDN* do provedor e até mesmo nos clientes. Para isso é necessária a utilização de uma arquitetura adequada, como por exemplo *microsserviços*, em que os módulos se comunicam por meio de um protocolo padronizado como *HTTP*.

Por fim, vale ressaltar que a periodicidade de alocação também pode ser alterada, como mostrado no Capítulo 7. Em outras palavras, tamanhos de janelas de tempo maiores podem ser considerados para que haja mais tempo tanto para recuperação dos dados dos *logs* de segmento, quanto para seu processamento em transmissões com um número mais elevado de clientes.

8.7 Sumário

Este capítulo apresentou uma proposta de mecanismo de alocação de recursos para transmissões adaptativas ao vivo via Internet. O objetivo do mecanismo proposto é obter um melhor compromisso entre o interesse do usuário, que é ter bom desempenho de transmissão, e o interesse do provedor de conteúdo, que é ter bom desempenho geral, o que aumenta o engajamento dos clientes e a escala de transmissão.

O mecanismo utilizou como base os conhecimentos construídos nos capítulos precedentes, e explora a premissa geral de que o mapeamento da relação entre desempenho de sessão e engajamento pode auxiliar no desenvolvimento de mecanismos de alocação de recursos que lidem melhor com o conflito de interesses entre usuários e provedores de conteúdo.

Em primeiro lugar, foi apresentada uma formalização do conflito de interesses entre usuários e provedores de conteúdo. Em seguida o capítulo apresentou detalhes do mecanismo adaptativo padrão adotado na transmissão considerada nesta tese e buscou mostrar evidências de como a infraestrutura lida com o desafio de manter o sistema ativo em cenários de sobrecarga. Foi mostrado que em geral há uma limitação generalizada de taxa de transmissão que, como já se sabe pelos capítulos anteriores, pode afetar os usuários de maneiras diferentes dependendo do contexto. Com isso é necessário desenvolver esquemas de alocação mais personalizados, que levem em conta a expectativa de desempenho de cada usuário.

O mecanismo pode funcionar de duas maneiras distintas. Na primeira o objetivo consiste em recuperar o engajamento de clientes prestes a sair. Já a segunda visa reduzir o consumo de clientes dispostos a permanecer no sistema. Nos dois casos, foi possível constatar que o uso do conhecimento da relação das métricas de desempenho de transmissão e adaptação e o engajamento foi capaz de aumentar a permanência dos usuários, duplicando-a em muitos casos. Já dentro do cenário de restrições de recurso, foi possível reduzir o consumo de banda a níveis que permitem o ingresso e permanência de milhares de novos clientes, com baixo impacto sobre o engajamento de usuários que já estão na transmissão.

É importante relembrar que o mecanismo proposto se trata de uma prova de conceito para a validação da hipótese geral da tese. Isso significa que diversas questões práticas podem ser suscitadas durante sua implementação em um sistema real. Um exemplo é a coleta das informações sobre desempenho do cliente, que devem ser processadas e disponibilizadas ao sistema de alocação em tempo hábil para que seja possível a criação dos cenários de desempenho. Outra questão poderia se referir ao tempo que o modelo de previsão leva para perder acurácia, momento em que deve ser retreinado para refletir as relações atuais entre desempenho/adaptação e engajamento.

Capítulo 9

Conclusões e Trabalhos Futuros

Neste capítulo serão apresentadas, de forma sumarizada, as principais conclusões e contribuições desta tese bem como potenciais direções futuras de pesquisa que podem usar como ponto de partida as contribuições do presente trabalho. Em primeiro lugar, a Seção 9.1 apresenta os principais resultados obtidos, categorizados conforme os objetivos de pesquisa estipulados no Capítulo 1 deste texto. Em seguida, na Seção 9.2, será apresentada uma discussão sobre as questões de pesquisa que permanecem em aberto e que, portanto, podem ser exploradas no futuro.

9.1 Resultados Obtidos

Como já abordado no Capítulo 1, esta tese estabeleceu quatro objetivos de pesquisa:

- OP1 - Caracterização de desempenho de transmissão e sua relação com engajamento
- OP2 - Modelagem do comportamento de clientes em transmissões adaptativas ao vivo
- OP3 - Modelagem de engajamento para vídeos adaptativos ao vivo
- OP4 - Melhoria da alocação de recursos em transmissões adaptativas ao vivo

A seguir, serão apresentados os resultados obtidos associados a cada um destes objetivos.

9.1.1 OP1 - Caracterização de Desempenho de Transmissão e sua Relação com Engajamento

Esse objetivo de pesquisa focou na caracterização do desempenho de um conjunto de transmissões ao vivo na Internet e no entendimento de como aspectos ligados ao desempenho de transmissão interferem no engajamento de um usuário. Para tanto, foi utilizado um conjunto de dados associados a 48 transmissões ao vivo, coletadas entre junho e julho de 2018, que totalizam juntas mais de 62 milhões de sessões, o que permitiu a investigação da dinâmica da relação entre clientes e provedores em um cenário real, de difusão de conteúdo em larga escala.

Em particular, foram abordadas métricas de desempenho relativas ao funcionamento da aplicação cliente que são ligadas a fatores percebidos visualmente pelo sistema sensorial do usuário, como a ocorrência de congelamentos e mudanças na taxa de transmissão. Esses fatores possuem a capacidade de influenciar indiretamente o engajamento e a percepção de qualidade do usuário. Também foram abordados aspectos contextuais como tipo de dispositivo, o provedor, e a localização geográfica, entre outros. Nesse sentido, as principais contribuições associadas ao entendimento da relação entre desempenho de transmissão e engajamento são os seguintes:

- Análise da influência de fatores contextuais no desempenho de transmissão de clientes e no engajamento dos usuários associados. Essa análise utiliza o conceito de ganho de informação para determinação dos fatores mais relevantes, viabilizando assim o uso de estratégias de otimização de desempenho focadas em entidades específicas como um dispositivo, região geográfica ou provedor em particular. Nesse sentido, a análise feita evidenciou que existe um impacto não desprezível de fatores como o tipo de dispositivo, provedor, período da transmissão, tanto para desempenho de transmissão quanto no engajamento.
- Caracterização do desempenho de transmissão por tipo de dispositivo. Foram utilizadas as métricas de desempenho de transmissão mais populares na academia, como latência de inicialização e taxas de congelamentos. Também foi incorporado o papel da dinâmica de trocas de taxa de transmissão, ainda pouco explorado [103], por meio das taxas de adaptações positivas e negativas. Os resultados mostram que uma significativa parcela de clientes ainda sofre com problemas de desempenho. Por exemplo, 40% das sessões de clientes fixos possui latência acima de 10 segundos e 10% das sessões experimentam 1 ou mais congelamentos por minuto. Além disso, foi apresentada uma análise da relação entre taxa de transmissão e escala, onde foi possível observar como a infraestrutura de transmissão lida com altas cargas de trabalho. Nessa análise, constatou-se que existe uma redução de taxa de transmissão proporcional ao número de clientes fixos

concomitantes, o que evidencia um potencial desafio futuro no sentido de conciliar aumento de escala e alto desempenho de transmissão geral.

- Avaliação do impacto do desempenho de transmissão sobre o engajamento. Demonstrou-se que existe uma redução do engajamento à medida em que crescem as taxas de adaptação e congelamentos, bem como a latência de inicialização. O aumento da taxa de transmissão, por outro lado, tem impacto positivo sobre o engajamento. Além disso, foram medidos as correlações entre cada métrica e engajamento, o que permitiu definir o nível de importância de cada métrica para o mesmo. Foi observado que diferentes dispositivos têm diferentes prioridades de desempenho, com clientes móveis sendo menos sensíveis a longas latências iniciais, por exemplo.
- Proposta de uma metodologia de caracterização para avaliar de forma integrada, por meio de cenários, o impacto de métricas de desempenho no engajamento. Um cenário de desempenho é definido como uma combinação específica de valores das métricas de desempenho de transmissão. A vantagem de estudar a relação entre uma métrica particular e o engajamento dentro de um cenário, é a incorporação da contribuição de todas as métricas de desempenho do conjunto. Foi utilizado um processo de agrupamento para a construção dos cenários de desempenho e, para cada cenário, foi construído um modelo de regressão para aprender as correlações entre métricas de desempenho e engajamento. Essa abordagem permitiu constatar que a relação entre engajamento e desempenho pode variar entre os usuários. Por exemplo, foi observado que um usuário em cliente fixo tolera longas esperas iniciais se o seu cenário geral foi de baixo desempenho. O mesmo não acontece se ele estiver em um cenário de alto desempenho.

Estudar o impacto do desempenho no engajamento pode direcionar o projeto de sistemas que realizem a alocação de os recursos de transmissão de acordo com os interesses de desempenho dos usuários, evitando assim o desperdício desses recursos, o que permitiria atender mais clientes concomitantemente.

As contribuições relativas a esse objetivo de pesquisa foram apresentadas em um artigo de conferência [47] e em um periódico [48].

9.1.2 OP2 - Modelagem do Comportamento de Clientes em Transmissões Adaptativas Ao Vivo

No objetivo de pesquisa 2, o foco foi na caracterização e modelagem do comportamento de clientes de transmissões adaptativas ao vivo. Foi observado que os modelos disponíveis na literatura, em sua maioria [43, 77, 85], não abordam diretamente o impacto

da adaptação sobre o consumo de recursos, o que pode levar a um dimensionamento equivocado da infraestrutura necessária para uma transmissão. Outro fator, também pouco abordado, é o impacto do desempenho de transmissão sobre o comportamento dos clientes. A razão para a exploração desse potencial impacto é a observação feita quanto a influência do desempenho sobre o engajamento do usuário, conforme reportado na seção anterior. A fim de estudar esse impacto, foi desenvolvido um modelo hierárquico que captura explicitamente a contribuição da adaptação sobre o consumo de recursos do cliente, bem como o impacto do desempenho sobre diversos fatores relacionados à permanência de clientes, como seus tempos de sessão, tempos de ausência e número de retornos. Nesse sentido, as principais contribuições no contexto do objetivo de pesquisa 2 foram:

- O desenvolvimento de um modelo hierárquico multi-camadas do comportamento de clientes em transmissões adaptativas ao vivo, parametrizado a partir de um conjunto com milhões de sessões. Este modelo avança sobre modelos anteriores ao propor (1) uma camada dedicada a descrever a atuação do algoritmo de adaptação, permitindo uma estimação de consumo de banda mais realista e (2) um método para a identificação e parametrização personalizada de diferentes tipos de comportamento, induzidos pelos diferentes níveis de desempenho registrados. Por meio desse método, foi mostrado que a taxa de transmissão recebida pelo cliente interfere no engajamento de seu usuário associado (i.e., duração de permanência, duração da ausência e número de retornos) e em sua dinâmica de adaptação de taxa de transmissão. Além disso, foi observado que a parametrização dos modelos especializados em cada comportamento produziu melhores ajustes de curvas se comparado a um modelo global que descreve indistintamente o comportamento de todos os clientes e o engajamento de seus usuários associados.
- Desenvolvimento e validação de um gerador de cargas sintéticas (*AdpGen*) que reproduz os tempos de permanência e ausência de um cliente, bem como sua quantidade de sessões. Além disso, também são reproduzidos aspectos como a sua taxa de adaptação e número de segmentos em cada taxa de transmissão. O gerador leva em consideração a parametrização personalizada em cada comportamento para gerar clientes em todos os níveis de desempenho registrados. Na validação, foi demonstrado que o AdpGen produziu cargas muito mais próximas à carga real, com redução de 65,5% no erro para representação dos tempos entre sessões e 60,5% para a taxa de transmissão média, se comparado com um modelo genérico que não captura os múltiplos comportamentos registrados.

As contribuições relativas a esse objetivo de pesquisa foram apresentadas em um artigo de periódico [48]. Além disso, o gerador de carga, parametrizado com dados coletados em 2014[1] e 2018[2] estão disponíveis publicamente para a academia. O gerador

também foi cedido para ser utilizado em projetos internos da *Samsung Research Brazil*[3], no ano de 2020.

9.1.3 OP3 - Modelagem de Engajamento para Vídeos Adaptativos Ao Vivo

No objetivo de pesquisa 3, os esforços se concentraram na criação de um modelo de engajamento. O objetivo da criação deste tipo de modelo é correlacionar aspectos relativos à transmissão, em particular métricas de desempenho de transmissão e métricas de adaptação, com o tempo de permanência do usuário no sistema. Modelos de engajamento podem ser usados para simular valores específicos nas métricas de desempenho e obter o engajamento resultante. Adicionalmente, também é possível a criação de modelos preditivos, que preveem o engajamento futuro a partir do desempenho histórico de um cliente.

A utilização de engajamento como uma medida de aceitação em modelos descritivos e preditivos ainda é pouco explorada, tendo em vista que o engajamento é uma medida susceptível à influência de fatores subjetivos, como o interesse do usuário. Nesse sentido, os poucos modelos do tipo registrados na literatura [8, 115, 114] são restritos quanto às aplicações em que podem ser usados e/ou possuem uma acurácia apenas moderada. Sendo assim, com o objetivo de preencher essa lacuna, as seguintes contribuições foram feitas nesse objetivo de pesquisa:

- Proposição de uma nova abordagem de modelagem, na qual uma sessão não é descrita apenas pelo desempenho das métricas de desempenho de transmissão (e.g., taxa de congelamentos e latência de inicialização), como nos modelos clássicos da literatura [8, 115, 114], mas também por meio de sua atividade de adaptação. Essa abordagem utiliza o mesmo conceito desenvolvido para a camada de segmento do modelo de comportamento de clientes, desenvolvido no OP2. Nesse conceito, a sessão de um cliente é descrita por sua matriz de transição de taxas de transmissão e outras métricas relacionadas à adaptação. A partir da comparação dessa nova abordagem com um modelo baseado em métricas clássicas de desempenho de transmissão, foi constatado que a nova abordagem produz uma melhora da acurácia em modelos descritivos, que chega a 39.3% na classificação, e uma redução de 55% no erro de regressão.

[3] https://research.samsung.com/srbr
[3] https://github.com/thiagoguarnieri/adpgen-adaptive-workoad-generator-2014
[3] https://github.com/thiagoguarnieri/adpgen-adaptive-workoad-generator-2018

- Proposição de um modelo preditivo de engajamento, onde é possível prever se um cliente irá permanecer em sua sessão pelos próximos n minutos, com n variando de 1 até 5, baseado em seu desempenho histórico. Ao contrário do modelo descritivo, no modelo preditivo é utilizada uma etapa prévia com métricas de desempenho de transmissão para agrupar os clientes em cenários de desempenho, de forma análoga ao que foi proposto no OP1. Em seguida, um modelo preditivo é treinado em cada cenário com as mesmas métricas relativas a adaptação utilizadas para o modelo descritivo. A partir dessa abordagem, foi possível alcançar uma acurácia de previsão de 80% para o horizonte de tempo de 1 minuto e 76,41% para 5 minutos.

Com o auxílio de modelos preditivos, é possível antecipar possíveis abandonos e estabelecer medidas para mitigá-los, por exemplo realocando os recursos em uma sessão para priorizar as métricas de desempenho/adaptação que mais estimulem a permanência do cliente. As contribuições desse OP foram publicadas em artigos de conferências [44, 45, 46].

9.1.4 OP4 - Melhoria da Alocação de Recursos em Transmissões Adaptativas Ao Vivo

Nesse objetivo de pesquisa, o objetivo foi a validação das ideias exploradas nas etapas anteriores pela proposição de um mecanismo de alocação de recursos para transmissões adaptativas. O mecanismo proposto visa a melhoria do compromisso entre o interesse dos usuários, que é ter o maior desempenho compatível com seus recursos de banda e *hardware*, e o interesse dos provedores de conteúdo, que é aumentar o número de clientes concomitantes e o engajamento médio nas suas transmissões, pelo aumento do desempenho geral em sua plataforma de disseminação. Tais interesses são conflitantes, porque a redução de recursos permite a entrada de novos clientes, mas reduz o engajamento dos usuários que já estão na transmissão. Por outro lado, aumentar os recursos em cada cliente faz com que o engajamento dos usuários aumente, porém com uma redução na quantidade de usuários concomitantes.

Com base nesse cenário, para auxiliar na tarefa de permitir que usuários usufruam de alto desempenho de transmissão ao mesmo tempo em que os provedores consigam ter seu interesse de aumento de escala de transmissão atendido, a seguinte contribuição foi produzida:

- Proposta de um mecanismo de alocação de recursos para transmissões adaptativas. Esse mecanismo age monitorando as sessões ativas na transmissão. Em cada minuto,

essas sessões têm suas métricas de desempenho e adaptação extraídas. Em seguida, esses valores são submetidos a um identificador de cenário de desempenho, como proposto no OP1. Posteriormente é executado, para cada sessão, o previsor de engajamento correspondente ao cenário associado. Após a previsão, é feito no mecanismo um rearranjo do regime de adaptação de cada cliente, de forma que os interesses do provedor sejam melhor contemplados. O mecanismo pode funcionar de duas formas. Na primeira, são selecionadas as sessões cuja previsão foi de permanência para o próximo minuto para que suas adaptações sejam reconfiguradas a fim de reduzir o consumo de recursos mantendo o engajamento. Na segunda forma, as sessões cujas previsões foram de abandono são coletadas e seus regimes de adaptação são reconfigurados para aumentar as chances de permanência no próximo minuto. Em ambos os casos, há uma melhora no no atendimento dos interesses das partes porque há uma preocupação em atender aos interesses do provedor sem prejudicar a qualidade, e consequente engajamento, de cada usuário individualmente. A partir de simulações baseadas em dados reais, foi possível observar situações em que o mecanismo é capaz de um ganho médio de 100% no engajamento dos clientes e uma economia média de 28% GB por minuto, que permite o ingresso de 71 mil novos clientes. Vale enfatizar que a previsão de engajamento, desenvolvida no OP3, é fator chave para determinar como alterar a cadeia de adaptação do cliente sem prejudicar significativamente o engajamento do usuário.

Vale salientar que abordar manutenção de engajamento dentro do conflito de interesses entre usuários e provedores é um aspecto pouco explorado, sendo os principais trabalhos voltados para outros fins como a melhoria de métricas de desempenho individuais ou medidas de aceitação objetivas como MOS [83, 38, 64, 29].

9.2 Trabalhos Futuros

Nesta seção, serão discutidas direções para futuras pesquisas e aprimoramentos. Os tópicos que serão abordados se relacionam tanto aos objetivos abordados nesta tese quanto a problemas relacionados, que eventualmente podem se beneficiar das contribuições apresentadas por esta tese.

- **Estudo do impacto do interesse no engajamento.** Um dos problemas do uso do engajamento como uma medida de aceitação é a dificuldade de estimar o efeito do interesse do usuário em seu tempo de permanência na transmissão. Esse problema é refletido nos modelos de engajamento, que em muitas situações não conseguem atingir uma boa acurácia de previsão e descrição de engajamento. Nesse sentido, já começam

a surgir trabalhos que objetivam incluir o efeito do interesse para previsão de engajamento [115]. No entanto, ainda não há um consenso sobre uma métrica relacionada a interesse na literatura.

Um outro caminho que pode ser trilhado para medir interesse em um contexto específico é entender a correlação entre as interações do usuário por meio de curtidas, comentários e contribuições financeiras sobre o engajamento. Plataformas que permitem esse tipo de interação entre usuário e produtor de conteúdo em vídeos ao vivo tem alcançado grande popularidade, principalmente após o advento da pandemia de COVID 19, e diversos trabalhos em outras áreas de conhecimento têm procurado entender o papel desse aspecto do interesse [52, 75, 66].

- **Ampliação do estudo da relação entre desempenho e engajamento.** Ao longo da tese, foram vistas novas abordagens para caracterizar a relação entre desempenho e engajamento. Foi proposto um modelo de agrupamento associado a modelos de regressão para estudar o impacto da variação de uma métrica de desempenho dentro de um cenário integrado, com todas as métricas. Como resultado, foi possível observar co-dependências entre essas métricas, que devem ser consideradas ao analisar uma variação de particular de desempenho.

  Nesse sentido, uma extensão dessa análise seria utilizar o modelo descritivo apresentado no Capítulo 7, que é mais rico que o modelo do Capítulo 5, associado a ferramentas de inteligência artificial explicável, como por exemplo o estudo das árvores de decisão produzidas pelo algoritmo de classificação.

  Outro caminho seria explorar a causalidade entre desempenho e engajamento. Para isso é possível usar regras de associação, que se valem do conceito lógico de implicação entre uma ou mais proposições antecedentes e uma consequente. Outro exemplo é o Projeto Quasi-Experimental [72]. Ele funciona da seguinte forma: assumindo que o desempenho de uma sessão é descrito por n métricas, são comparados 2 clientes, A e B, com desempenho similar em $n-1$ métricas, sendo que A tem baixo engajamento, e B alto. A partir desse pareamento, é possível medir o impacto isolado da métrica restante, que não foi considerada no conjunto das $n-1$ métricas mencionado, sobre o engajamento. Esse método pode ser adaptado para considerar apenas sessões de um mesmo cliente ou sessões com contexto similar, como por exemplo, mesmo dispositivo e provedor de Internet.

- **Melhorias na modelagem de comportamento de clientes.** O modelo de comportamento de clientes em sistemas adaptativos ao vivo, apresentado no Capítulo 6, produziu avanços importantes ao considerar o impacto da adaptação no consumo de recursos de um cliente, além de propor uma abordagem para identificar e parametrizar múltiplos comportamentos induzidos pela taxa de transmissão oferecida aos usuários. Nesse sentido, existem ainda lacunas que podem ser exploradas a fim de evoluir o

modelo apresentado. Um exemplo é a simulação do nível de *buffer* do cliente, que permitiria capturar métricas importantes para o engajamento tais como a taxa de congelamentos e latência de inicialização.

- **Melhorias na escalabilidade e descentralização do mecanismo de alocação.** Plataformas de disseminação de vídeo geralmente transmitem milhares de vídeos ao vivo simultaneamente. Nesse sentido, é necessário que o mecanismo de alocação seja escalável. Na proposta apresentada no Capítulo 8 é mencionado que o mecanismo pode funcionar de maneira distribuída.

  No entanto, um caminho para obter uma escalabilidade ainda maior seria a adoção de uma arquitetura de microsserviços. Nessa arquitetura, os diversos módulos de alocação podem executar em diferentes computadores, físicos, virtualizados ou em contêineres, e se comunicam por troca de mensagens. Com isso os módulos podem ser replicados, permitindo que o processo de alocação de recursos seja gerenciado não só no nível de provedor de serviços, mas também no provedor de conectividade com a Internet, ou mesmo na rede local dos usuários, ou em seus dispositivos, que atualmente já contam com grande poder de processamento.

  Um outro aspecto que precisa ser estudado é a integração do mecanismo de alocação com o sistema de armazenamentos de *logs* do provedor. É necessário que as requisições de segmentos sejam rapidamente ordenadas em termos de sua ocorrência no tempo para permitir construir as métricas de sessão e a instância de previsão. Uma direção nesse sentido é o uso de bancos de dados não relacionais e o uso de índices, que em estudos preliminares dessa tese demonstraram um potencial de redução de mais de 80% no tempo de recuperação das informações das sessões.

- **Melhorias na previsão de engajamento e alocação de recursos.** Em muitos serviços de *streaming* existe a categoria dos usuários pagantes e aquela em que o acesso é gratuito. Nesse sentido, o simulador pode ser aprimorado para garantir uma prioridade maior aos usuários pagantes, isto é, reduzir a qualidade de clientes gratuitos a um nível que preserve seu engajamento, e permitir desempenho mais alto para os usuários que efetivamente produzem lucro para o serviço. Esse mesmo raciocínio poderia ser aplicado em uma abordagem par-a-par, onde os clientes são bonificados com maior desempenho se contribuírem para a disseminação do conteúdo.

  Ainda em relação ao mecanismo de alocação, foi possível observar que ele atinge um melhor compromisso entre o interesse do cliente, que é um desempenho mínimo, e do provedor, que são mais clientes simultâneos. Apesar disso, o modelo não garante uma solução ótima para o problema. Nesse sentido, o modelo descritivo poderia ser melhor explorado para produzir um conjunto de equações de custo-benefício que maximizem o compromisso descrito, nos moldes do que foi apresentado em [41, 42].

Já no caso do algoritmo de previsão de engajamento, há a possibilidade de ir em novas direções em busca de uma maior acurácia de previsão. Uma alternativa é o uso de séries temporais para explicar engajamento (e.g. variação temporal da taxa de transmissão, ou número de congelamentos) e a adição de mais fatores contextuais como o tipo de conteúdo e o interesse.

- **Reprodutibilidade em outros conteúdos e contextos.** As caracterizações, análises, modelos e mecanismos propostos ao longo desta tese estão fundamentados num conjunto de transmissões de conteúdo específico, que no caso são partidas de futebol da Copa do Mundo.

  Esse conteúdo é interessante sob diversos aspectos. Em particular, jogos da copa gozam de grande apelo junto aos brasileiros, sua principal audiência. Com isso, a falta de interesse como razão para redução de engajamento tem menos relevância, e assim é possível estabelecer com mais exatidão a contribuição do desempenho para o engajamento.

  Uma segunda vantagem é que a taxa de acessos nesse tipo de conteúdo é fortemente concentrada no tempo, fato que permite observar cenários de sobrecarga e também como o provedor reage a surtos repentinos de chegada de clientes.

  Apesar dessas vantagens, é importante também, como trabalho futuro, analisar a reprodutibilidade das análises aqui mostradas em outros contextos, como transmissões ao vivo com durações mais longas, como *podcasts* ou *reality shows*. Por exemplo, em [23] os autores usam dados de outras transmissões da Globo e concluem que a natureza do conteúdo afeta a relação entre desempenho e engajamento. Também nesse estudo foi mostrado que o tipo de conexão de Internet do cliente e sistema operacional são fatores importantes para o engajamento e não apenas o tipo de dispositivo.

Referências

[1] Adobe. Http dynamic streaming specification version 3.0 final. https://ossrs.io/lts/en-us/assets/files/adobe-hds-specification-8885755a21097e36f659cfb4e6044ad5.pdf, Acesso em setembro de 2020.

[2] A. Ahmed, Z. Shafiq, H. Bedi, and A. Khakpour. Suffering from buffering? detecting QoE impairments in live video streams. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*, 2017.

[3] A. Ahmed, Z. Shafiq, and A. Khakpour. QoE analysis of a large-scale live video streaming event. In *Proc. of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, 2016.

[4] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis. What happens when http adaptive streaming players compete for bandwidth? In *Proceedings of the 22Nd International Workshop on Network and Operating System Support for Digital Audio and Video*, 2012.

[5] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon. Analysis of educational media server workloads. In *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2001.

[6] Anatel. Dados de acessos de comunicação multimídia. https://dados.gov.br/dataset/dados-de-acessos-de-comunicacao-multimidia, Acesso em outubro de 2020.

[7] Anatel. Quantidade mensal de acessos em serviços do serviço móvel pessoal - smp. https://dados.gov.br/dataset/acessos-autorizadas-smp, Acesso em outubro de 2020.

[8] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. *SIGCOMM Comput. Commun. Rev.*, 43(4):339–350, 2013.

[9] N. Barman and M. G. Martini. Qoe modeling for http adaptive video streaming - a survey and open challenges. *IEEE Access*, 7:30831–30859, 2019.

[10] A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann. A survey on bitrate adaptation schemes for streaming media over http. *IEEE Communications Surveys Tutorials*, 21(1):562–585, 2019.

[11] A. Borges, P. Gomes, J. Nacif, R. Mantini, J. M. de Almeida, and S. Campos. Characterizing SopCast Client Behavior. *Comput. Commun.*, 35(8):1004–1016, 2012.

[12] K. Brunnström, K. Moor, A. Dooms, S. Egger-Lampl, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, C. Larabi, B. Lawlor, P. Le Callet, S. Möller, F. Pereira, M. Pereira, A. Perkis, A. Pinheiro, U. Reiter, P. Reichl, R. Schatz, and A. Zgank. *Qualinet White Paper on Definitions of Quality of Experience*. European Network on Quality of Experience in Multimedia Systems and Services, 03 2013.

[13] M. C. Calzarossa, L. Massari, and D. Tessera. Workload characterization: A survey revisited. *ACM Comput. Surv.*, 48(3), Feb. 2016.

[14] J. A. C. Carvalho. Nonparametric Off-Policy Policy Gradient. Master's thesis, Albert-Ludwigs-Universitat Freiburg, Geremany, 2019.

[15] I. M. Chakravarti, R. G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics*, volume 1. John Wiley and Sons, 1967.

[16] J. Chen and J. Wu. Dynamic adaptive streaming based on deep reinforcement learning. *Journal of Physics: Conference Series*, 1237(2):022124, jun 2019.

[17] L. Chen, Y. Zhou, and D. M. Chiu. Video browsing - a study of user behavior in online vod services. In *2013 22nd International Conference on Computer Communication and Networks (ICCCN)*, 2013.

[18] L. Chen, Y. Zhou, and D. M. Chiu. Video browsing - a study of user behavior in online vod services. In *2013 22nd International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7, 2013.

[19] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785?794, 2016.

[20] Cisco. Cisco annual internet report (2018–2023) white paper. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, Acesso em setembro de 2020.

[21] Clappr. Clappr: An extensible media player for the web. https://github.com/clappr/clappr, Acesso em março de 2023.

[22] Cnet.com. Disney plus launch glitches out with service failures, login problems. https://www.cnet.com/news/disney-plus-launch-glitches-service-failures-login-problems/, Acesso em novembro de 2019.

[23] D. V. Correa da Silva, P. B. Velloso, and A. A. d. A. Rocha. Using data mining techniques to extract key factors in mobile live streaming. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, 2019.

[24] C. P. Costa, Í. S. Cunha, A. B. Vieira, C. V. Ramos, M. V. de Melo Rocha, J. M. Almeida, and B. A. Ribeiro-Neto. Analyzing client interactivity in streaming media. In *WWW '04*, 2004.

[25] W. de Almeida Junior, B. Almeida, I. Cunha, J. M. Almeida, and A. B. Vieira. Caracterizacão da transmissão de um grande evento esportivo. In *33o. Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, 2015.

[26] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens. Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching. *IEEE Transactions on Broadcasting*, 59(1):47–61, 2013.

[27] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. In *Proceedings of the ACM SIGCOMM 2011 Conference*, 2011.

[28] J. G. Donat. Php user agent parser. https://github.com/donatj/PhpUserAgent, Acesso em março de 2023.

[29] R. Dubin, R. Shalala, A. Dvir, O. Pele, and O. Hadar. A fair server adaptation algorithm for http adaptive streaming using video complexity. *Multimedia Tools and Applications*, pages 11203–11222, 2019.

[30] Ericsson. Ericsson mobility report. ericsson.com/en/reports-and-papers/mobility-report, Acesso em novembro de 2020.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226?231. AAAI Press, 1996.

[32] M. Falkhausen, H. Reininger, and D. Wolf. Calculation of distance measures between hidden markov models. In *In Proc. Eurospeech*, pages 1487–1490, 1995.

[33] M. Fiedler and K. K. P. Reich. From quality of service to quality of experience. In *Dagstuhl Seminar*, 2009.

[34] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto. Trendlearner: Early prediction of popularity trends of user generated content. *CoRR*, abs/1402.2351, 2014.

[35] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto. On the dynamics of social media popularity: A youtube case study. *ACM Transactions on Internet Technology (TOIT)*, 14(4), 2014.

[36] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 3(21):768–769, 1965.

[37] J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 10 2001.

[38] A. Ganjam, F. Siddiqui, J. Zhan, X. Liu, I. Stoica, J. Jiang, V. Sekar, and H. Zhang. C3: Internet-Scale control plane for video quality optimization. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, pages 131–144. USENIX Association, May 2015.

[39] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant. Study of the effects of stalling events on the quality of experience of mobile streaming videos. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 989–993, 2014.

[40] G. D. Gonalves, I. Drago, A. B. Vieira, A. P. Couto da Silva, J. M. Almeida, and M. Mellia. Workload models and performance evaluation of cloud storage services. *Comput. Netw.*, 109(P2):183?199, Nov. 2016.

[41] G. D. Gonçalves, I. Drago, A. V. Borges, A. P. Couto, and J. M. de Almeida. Analysing costs and benefits of content sharing in cloud storage. In *Proceedings of the 2016 Workshop on Fostering Latin-American Research in Data Communication Networks*, 2016.

[42] G. Gonçalves, A. B. Vieira, I. Drago, A. P. Couto Da Silva, and J. M. Almeida. Cost-benefit tradeoffs of content sharing in personal cloud storage. In *2017 IEEE 25th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 2017.

[43] A. Gouta, C. Hong, D. Hong, A. Kermarrec, and Y. Lelouedec. Large scale analysis of http adaptive streaming in mobile networks. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, 2013.

[44] T. Guarnieri, J. Almeida, and A. Vieira. An adaptation aware model to predict engagement on http adaptive live streaming. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, 2019.

[45] T. Guarnieri, A. Vieira, I. Cunha, and J. Almeida. Previsão de engajamento de usuários durante transmiss ão adaptativa de vídeo ao vivo. In *Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, 2018.

[46] T. Guarnieri, A. B. Vieira, and J. Almeida. Um modelo sensível a adaptação para previsão de qualidade de experiência em vídeos na internet. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, 2019.

[47] T. Guarnieri, Ítalo Cunha, J. Almeida, I. Drago, and A. B. Vieira. Characterizing QoE in large-scale live streaming. In *Proc. of the IEEE GLOBECOM*, 2017.

[48] T. A. Guarnieri, I. Drago, Í. S. Cunha, B. M. de Almeida, J. M. Almeida, and A. B. Vieira. Modeling large-scale live video streaming client behavior. *Multimedia Systems*, pages 1–24, 2021.

[49] Z. Guo, Y. Wang, and X. Zhu. Assessing the visual effect of non-periodic temporal variation of quantization stepsize in compressed video. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3121–3125, 2015.

[50] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross. A measurement study of a large-scale p2p iptv system. *IEEE Transactions on Multimedia*, 9(8):1672–1687, 2007.

[51] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross. A measurement study of a large-scale p2p iptv system. *IEEE Transactions on Multimedia*, 9(8):1672–1687, 2007.

[52] Z. Hilvert-Bruce, J. T. Neill, M. Sjöblom, and J. Hamari. Social motivations of live-streaming viewer engagement on twitch. *Computers in Human Behavior*, 84:58–67, 2018.

[53] L. Hindersin, B. Wu, A. Traulsen, and J. García. Computation and simulation of evolutionary game dynamics in finite populations. *Scientific Reports*, 9, 05 2019.

[54] T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.

[55] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen. Initial delay vs. interruptions: Between the devil and the deep blue sea. In *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012.

[56] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. Quantification of youtube qoe via crowdsourcing. In *2011 IEEE International Symposium on Multimedia*, 2011.

[57] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner. Assessing effect sizes of influence factors towards a qoe model for http adaptive streaming. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2014.

[58] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, page 187–198, 2014.

[59] Q. Huynh-Thu and M. Ghanbari. Temporal aspect of perceived quality in mobile video broadcasting. *IEEE Transactions on Broadcasting*, 54(3):641–651, 2008.

[60] R. Huysegems, B. De Vleeschauwer, K. De Schepper, C. Hawinkel, T. Wu, K. Laevens, and W. Van Leekwijck. Session reconstruction for http adaptive streaming: Laying the foundation for network-based qoe monitoring. In *2012 IEEE 20th International Workshop on Quality of Service*, 2012.

[61] V. Ishakian, R. Sweha, A. Bestavros, and J. Appavoo. Cloudpack* exploiting workload flexibility through rational pricing. In *Proceedings of the 13th International Middleware Conference*, pages 374–393, 2012.

[62] ITU-T. P.800.1 : Mean opinion score (mos) terminology. Standard, ITU Telecommunication Standardization Sector, 2016.

[63] ITU-T. P.1203 : Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport. Standard, ITU Telecommunication Standardization Sector, 2017.

[64] J. Jiang, V. Sekar, H. Milner, D. Shepherd, I. Stoica, and H. Zhang. CFA: A practical prediction system for video qoe optimization. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, 2016.

[65] J. Jiang, S. Sun, V. Sekar, and H. Zhang. Pytheas: Enabling data-driven quality of experience optimization using group-based exploration-exploitation. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, 2017.

[66] H. Jodén and J. Strandell. Building viewer engagement through interaction rituals on twitch.tv. *Information, Communication & Society*, 25(13):1969–1986, 2022.

[67] P. Juluri, V. Tamarapalli, and D. Medhi. Measurement of quality of experience of video-on-demand services: A survey. *IEEE Communications Surveys and Tutorials*, 18(1):401–418, 2016.

[68] K. Kim, B. Y. Cho, and W. W. Ro. Server side, play buffer based quality control for adaptive media streaming. *Multimedia Tools and Applications*, 75(10):5397–5415, 2016.

[69] S. Kim and C. Kim. Xmas: An efficient mobile adaptive streaming scheme based on traffic shaping. *IEEE Transactions on Multimedia*, 21(2):442–456, 2019.

[70] J. Klink and S. Brachmaski. An impact of the encoding bitrate on the quality of streamed video presented on screens of different resolutions. In *The 30th International Conference on Software, Telecommunications and Computer Networks (SoftCon)*, 10 2022.

[71] C. Kreuzberger, B. Rainer, H. Hellwagner, L. Toni, and P. Frossard. A comparative study of dash representation sets using real user characteristics. In *Proceedings of the 26th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2016.

[72] S. S. Krishnan and R. K. Sitaraman. Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. In *Proceedings of the 2012 Internet Measurement Conference*, 2012.

[73] P. Lebreton and K. Yamagishi. Study on user quitting rate for adaptive bitrate video streaming. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2019.

[74] P. Lebreton and K. Yamagishi. Study on user quitting in the puffer live tv video streaming service. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 19–24, 2021.

[75] A. Leith and E. Gheen. Twitch in the time of quarantine: The role of engagement in needs fulfillment. *Psychology of Popular Media*, 11, 11 2021.

[76] B. Li and H. Yin. Peer-to-peer live video streaming on the internet: issues, existing approaches, and challenges [peer-to-peer multimedia streaming]. *IEEE Communications Magazine*, 45(6):94–99, 2007.

[77] C. Li, J. Liu, and S. Ouyang. Large-scale user behavior characterization of online video service in cellular network. *IEEE Access*, 4:3675–3687, 2016.

[78] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652, Acesso em março de 2023.

[79] Z. Li, A. C. Begen, J. Gahm, Y. Shan, B. Osler, and D. Oran. Streaming video over http with consistent quality. In *Proceedings of the 5th ACM Multimedia Systems Conference*, MMSys '14, page 248–258, 2014.

[80] Z. Li, G. Xie, M. A. Kaafar, and K. Salamatian. User behavior characterization of a large-scale mobile live streaming system. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[81] LightTerra. Video encoding settings for h.264 excellence. http://www.lighterra.com/papers/videoencodingh264/, Acesso em outubro de 2019.

[82] C. Liu, I. Bouazizi, and M. Gabbouj. Rate adaptation for adaptive http streaming. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, 2011.

[83] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang. A case for a coordinated internet video control plane. In *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2012.

[84] Y. Liu, T. Lin, Z. Liu, and L. Dai. A cache-aware approach for dynamic adaptive video streaming over http. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, 2019.

[85] B. Machado, A. Vieira, I. Cunha, and A. Ziviani. Evolução do comportamento do usuário em eventos de larga escala na internet. In *Anais do XVIII Workshop em Desempenho de Sistemas Computacionais e de Comunicação*, 2019.

[86] H. Mao, R. Netravali, and M. Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, 2017.

[87] Maxmind. Geoip databases and services. https://www.maxmind.com/en/geoip2-services-and-databases, Acesso em março de 2023.

[88] W. Milliken, T. Mendez, and C. Partridge. Rfc 1546. https://rfc-editor.org/rfc/rfc1546.txt, Acesso em março de 2020.

[89] M. Mohammad and W. Tranter. Comparing distance measures for hidden markov models. In *Proceedings of the IEEE SoutheastCon 2006*, pages 256–260, 2006.

[90] R. K. Mok, E. W. Chan, X. Luo, and R. K. Chang. Inferring the qoe of http video streaming from user-viewing activities. In *Proceedings of the First ACM SIGCOMM Workshop on Measurements up the Stack*, 2011.

[91] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang. Measuring the quality of experience of http video streaming. In *12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops*, 2011.

[92] C. Moldovan, F. Wamser, and T. Hoßfeld. User behavior and engagement of a mobile video streaming user from crowdsourced measurements. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019.

[93] F. Murtagh and P. Legendre. Ward?s hierarchical agglomerative clustering method: Which algorithms implement ward?s criterion? *Journal of Classification*, 31:274–295, 2014.

[94] H. Nam, K. Kim, and H. Schulzrinne. Qoe matters more than qos: Why people stop watching cat videos. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, 2016.

[95] H. Nam and H. Schulzrinne. Youslow : What influences user abandonment behavior for internet video? In *Columbia University, Tech*, 2016.

[96] OpenIPTV. Volume 2a - http adaptive streaming. http://www.oipf.tv/web-spec/volume2a.htm, Acesso em outubro de 2020.

[97] J. Padhye and J. Kurose. An empirical study of client interactions with a continuous-media courseware server. In *NOSSDAV 1998*, 1998.

[98] R. Pantos and W. May. Http live streaming. rfc 8216. https://www.rfc-editor.org/rfc/rfc8216.html, Acesso em janeiro de 2023.

[99] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 365–374, 2013.

[100] Y. Qi and M. Dai. The effect of frame freezing and frame skipping on video quality. In *2006 International Conference on Intelligent Information Hiding and Multimedia*, 2006.

[101] N. Ramzan, H. Park, and E. Izquierdo. Video streaming over p2p networks: Challenges and opportunities. *Signal Processing: Image Communication*, 27(5):401–411, 2012. ADVANCES IN 2D/3D VIDEO STREAMING OVER P2P NETWORKS.

[102] D. Z. Rodriguez, J. Abrahao, D. C. Begazo, R. L. Rosa, and G. Bressan. Quality metric to assess video streaming service over tcp considering temporal location of pauses. *IEEE Transactions on Consumer Electronics*, 58(3):985–992, 2012.

[103] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia. A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys Tutorials*, 17(1):469–492, 2015.

[104] H. Shen and Z. Li. New bandwidth sharing and pricing policies to achieve a win-win situation for cloud provider and tenants. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014.

[105] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino. Quality of experience estimation for adaptive http/tcp video streaming using h.264/avc. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pages 127–131, 2012.

[106] L. Skorin-Kapov and M. Varela. A multi-dimensional view of qoe: the arcu model. In *2012 Proceedings of the 35th International Convention MIPRO*, 2012.

[107] I. Sledge and J. Príncipe. Reduction of markov chains using a value-of-information-based approach. *Entropy*, 21(4):349, 2019.

[108] A. Sobhani, A. Yassine, and S. Shirmohammadi. A video bitrate adaptation and prediction mechanism for http adaptive streaming. *ACM Trans. Multimedia Comput. Commun. Appl.*, 13(2), mar 2017.

[109] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman. Bola: Near-optimal bitrate adaptation for online videos. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, 2016.

[110] M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

[111] T. Stockhammer. Dynamic adaptive streaming over http –: Standards and design principles. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, 2011.

[112] Y. Sun, X. Yin, J. Jiang, V. Sekar, F. Lin, N. Wang, T. Liu, and B. Sinopoli. Cs2p: Improving video bitrate selection and adaptation with data-driven throughput prediction. In *Proceedings of the 2016 ACM SIGCOMM Conference*, 2016.

[113] N. Tahir, T. Minhas, and M. Fiedler. Impact of disturbance locations on video quality of experience. In *2nd Workshop QoEMCS*, 2011.

[114] S. Takahashi, K. Yamagishi, and J. Okamoto. Classification of viewing abandonment reasons for adaptive bitrate streaming. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2020.

[115] X. Tan, Y. Guo, M. Orgun, L. Xue, and Y. Chen. An engagement model based on user interest and qos in video streaming systems. *Wireless Communications and Mobile Computing*, 2018:1–11, 1 2018.

[116] R. I. Tavares da Costa Filho, W. Lautenschlager, N. Kagami, V. Roesler, and L. P. Gaspary. Network fortune cookie: Using network measurements to predict video streaming performance and qoe. In *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016.

[117] R. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[118] TvLine. Hbo go feels game of thrones fans' wrath due to technical difficulties. https://tvline.com/2019/04/28/hbo-go-outage-game-of-thrones-battle-of-winterfell-episode/, Acesso em outubro de 2019.

[119] E. Veloso, V. Almeida, W. M. Jr., A. Bestavros, and S. Jin. A Hierarchical Characterization of a Live Streaming Media Workload. *IEEE/ACM Trans. Netw.*, 14(1):133–146, 2002.

[120] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer Publishing Company, Incorporated, 2010.

[121] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

[122] N. Wehner, M. Seufert, S. Egger-Lampl, B. Gardlo, P. Casas, and R. Schatz. Scoring high: Analysis and prediction of viewer behavior and engagement in the context of 2018 FIFA WC live streaming. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 807–815, 2020.

[123] S. Wu, M.-A. Rizoiu, and L. Xie. Beyond views: Measuring and predicting engagement in online videos. In *2018 The International AAAI Conference on Web and Social Media*, 2018.

[124] J. Yao, S. S. Kanhere, I. Hossain, and M. Hassan. Empirical evaluation of http adaptive streaming under vehicular mobility. In *NETWORKING 2011*, 2011.

[125] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over http. *SIGCOMM Comput. Commun. Rev.*, 45(4):325–338, 2015.

[126] A. Zambelli. Iis smooth streaming technical overview. microsoft corpoation. https://learn.microsoft.com/en-us/iis/media/on-demand-smooth-streaming/smooth-streaming-technical-overview, Acesso em setembro de 2020.

[127] D. Zegarra Rodríguez, R. Lopes Rosa, E. Costa Alfaia, J. Issy Abrahão, and G. Bressan. Video quality metric for streaming service using dash standard. *IEEE Transactions on Broadcasting*, 62(3):628–639, 2016.

[128] T. Zinner, T. Hossfeld, T. N. Minhas, and M. Fiedler. Controlled vs. uncontrolled degradations of qoe : The provisioning-delivery hysteresis in case of video. In *New Dimensions in the Assessment and Support of Quality of Experience (QoE) for Multimedia Applications*, 2010.

Apêndice A

Parâmetros dos Modelos de Comportamento

Tabela A.1 apresenta os parâmetros de ajuste de curvas para o modelo de comportamento do Capítulo 6. O processo de agrupamento originou 8 grupos de comportamento. Como pode ser visto, os grupos têm diferentes regimes, que são refletidos na variabilidade dos ajustes encontrados. A seguir estão detalhadas as distribuições encontradas.

- A PMF da distribuição binomial negativa é $p(k, n, p) = \binom{k+n-1}{n-1} p^n (1-p)^k$ para $k \geq 0$.
- A PDF Para Weibull exponencial é $f(x, a, c) = ac(1-exp(-x^c))^{a-1} exp(-x^c) x^{c-1}$ com $x > 0, a > 0, c > 0$ e a e c como parâmetros de forma.
- A PDF da distribuição Log-normal potência é $f(x, c, s) = \frac{x}{cs}\phi(log(x)/s)\Phi(-log(x)/s))^{c-1}$ para $x \geq 0$ e $a > 0$. ϕ é a PDF da distribuição Normal, Φ é a CDF da distribuição Normal, e $x > 0, s, c > 0$. c e s são os parâmetros de forma.
- A PDF da distribuição Log-normal é $f(x, s) = \frac{1}{sx\sqrt{2\pi}} e^{(-1/2(\frac{log(x)}{s})^2)}$ com $x, s > 0$ e s como parâmetros de forma.
- A PDF da distribuição Gamma $f(x, a) = \frac{x^{a-1} exp(-x)}{\Gamma(a)}$ para $x \geq 0$ e $a > 0$. A distribuição $\Gamma(a)$ é a função Gamma e a é o parâmetro de forma.
- A distribuição Erlang é um caso especial da distribuição Gamma com o parâmetro de forma a como inteiro.
- A PDF da distribuição Gamma Generalizada é $f(x, a, c) = \frac{|c| x^{ca-1} exp(-x^c)}{\Gamma(a)}$ com $x \geq 0$, $a > 0$ e $c \neq 0$. a e c are como parâmetros de forma e $\Gamma(a)$ é a função Gamma.
- A PDF da distribuição Weibull (*Frechet right*) é $f(x, c) = cx^{c-1} exp(-x^c)$ para $x > 0$ e $c > 0$ e c como parâmetros de forma.
- A PDF da distribuição Potência Exponencial é $f(x, b) = bx^{b-1} exp(1 + x^b - exp(x^b))$ com $x \geq 0$ e $b > 0$ como parâmetros de forma.
- A PDF da distribuição Beta é $f(x, a, b) = \frac{\Gamma(a+b) x^{a-1} (1-x)^{b-1}}{\Gamma(a)\Gamma(b)}$ para $0 \leq x \leq 1, a > 0, b > 0$ onde Γ é a função Gamma.

Tabela A.1: Parâmetros para modelos especializados (topo) e modelo único (abaixo).

Especialização por Comportamento										
Grupo	Best fit.	Med.	Desv. Pad.	Forma. 1	Shape Par. 2	Loc.	Scale	MSE	Goodness (KS and AD)	P-value
Ontime (s)										
Fixo-Única-BQ	Power Lognormal	530.4023	1 407.786	0.0389	0.4113	-5.0557	13.0226	138 416	0.07	0.06
Fixo-Mult-BQ	Lognormal	373.2391	705.3332	1.3613	–	-2.6475	158.4657	1 984	0.05	0.44
Móvel-Mix-BQ	Lognormal	585.8209	1 177.214	1.7491	–	1.0545	164.2389	143 649	0.05	0.29
Fixo-Mix-MQ	Exp. Power	3 725.837	3 228.845	0.7553	–	0	6 573.4892	169 466	0.06	0.10
Fixo-Mix-AQ	Exp. Power	3 367.94	3 210.551	0.6759	–	0	6 115.9294	146 195	0.06	0.10
Móvel-Mix-AQ	Exp. Weibull	1 167.461	1 616.276	1.6167	0.5739	0	505.7510	8 944	0.03	0.92
Fixo-Mix-VQ	Beta	2 608.156	2 868.757	0.5377	3.5551	0	2 1268.1875	176 410	0.05	0.44
Móvel-Mix-VQ	Exp. Weibull	888.7709	1 359.043	1.8355	0.4921	0	267.3873	13 746	0.03	0.85
Offtime (s)										
Fixo-Mult-BQ	Weibull	1 387.88	1 933.087	0.6424	–	179.9999	893.6772	10 062	0.04	0.62
Móvel-Mix-BQ	Gen. Gamma	1 212.077	1 798.686	2.0688	0.4070	179.9999	91.5716	11,812	0.04	0.67
Fixo-Mix-MQ	Gen. Gamma	1 082.398	1 288.366	2.3915	–	902.5628	864.0239	175 090	0.06	0.29
Fixo-Mix-AQ	Gamma	1 174.829	1 499.34	0.4443	–	179.9999	1 973.9611	51 235	0.05	0.29
Móvel-Mix-AQ	Gen. Gamma	1 563.053	2 048.097	1.2894	0.5630	179.9999	590.1127	33 926	0.03	0.72
Fixo-Mix-VQ	Gen. Gamma	1 144.255	1 397.881	2.1517	0.4234	179.9999	86.8629	43 858	0.02	0.98
Móvel-Mix-VQ	Exp. Weibull	1 619.592	2 090.083	1.8307	0.4926	179.9999	443.8503	89 607	0.04	0.53
Número de sessões. Nota: valor crítico para AD is 0.325										
Fixo-MLT-BQ	Neg. binomial	3.9351	2.8177	0.61	0.24	2	–	$2.62 * 10^{-6}$	-1.10	0.25
Fixo-Mix-BLQ	Neg. binomial	1.8244	1.6736	0.46	0.38	1	–	$7.88 * 10^{-6}$	-0.22	0.25
Fixo-Mix-MQ	Neg. binomial	1.4171	1.0191	0.33	0.38	1	–	0.00035	-0.57	0.25
Fixo-Mix-AQ	Neg. binomial	1.3698	0.9374	0.56	0.60	1	–	$7.30 * 10^{-5}$	-0.89	0.25
Móvel-Mix-AQ	Neg. binomial	1.6880	1.1574	0.83	0.55	1	–	$4.28 * 10^{-7}$	-1.33	0.25
Fixo-Mix-VQ	Neg. binomial	1.3835	1.0193	0.33	0.44	1	–	0.00013	-1.26	0.25
Móvel-Mix-VQ	Neg. binomial	1.4957	1.0071	0.50	0.50	1	–	$1.34 * 10^{-6}$	0.094	0.25
Modelo Geral										
Ontime (s)										
	Weibull	1 806.646	2 563.408	0.6332	–	0	1 301.0819	275 164	0.03	0.96
Offtime (s)										
	Expon-weib.	629.3417	1 423.243	2.0198	0.4312	179.9999	251.5142	119 985	0.03	0.92
Número de sessões										
	Neg. binom.	1.5940	1.2965	0.33	0.30	1	–	0.00022	-0.47	0.25
Notes: O valor crítico para AD é 0.325 e o p-value deve ser maior que 0.05										

Fonte: Elaborado pelo autor.