Uma Abordagem Não Supervisionada para Inferir Qualidade de Experiência em Redes Sem Fio de Grande Escala

Diogo Menezes Ferrazani Mattos, Dianne Scherly Varela de Medeiros, Natalia Castro Fernandes e Luiz Claudio Schara Magalhães

[1]MídiaCom - Departamento de Engenharia de Telecomunicações
Universidade Federal Fluminense (UFF)

Resumo. *Inferir a qualidade da experiência de usuários de redes sem fio é desafiador, pois o monitoramento da rede não captura a qualidade para cada usuário individualmente. Este artigo propõe uma abordagem não supervisionada, baseada em aprendizado de máquina, para inferir a qualidade de experiência de diferentes perfis de uso de uma rede sem fio de grande escala. A abordagem proposta usa a correlação entre dados de uso de pontos de acesso e estatísticas de fluxos de dados na rede. A ideia central da proposta é coletar dados de utilização de diversos pontos de acesso, correlacioná-los com as estatísticas dos fluxos das conexões que passam pelos pontos de acesso, reportados pelo NetFlow, e, a partir da aplicação do algoritmo de agrupamento* k-means*, inferir diferentes perfis de uso da rede. A abordagem proposta foi avaliada na rede sem fio real de grande escala e os resultados mostram que a separação dos fluxos em cinco agrupamentos permite identificar perfis característicos de estados degradados da rede e situações de sobrecarga em pontos de acesso, considerando apenas as estatísticas de fluxos reportadas.*

Abstract. *Inferring the quality of experience of wireless users is challenging, as network monitoring does not capture quality perception for each user individually. In this paper, we propose an unsupervised approach, based on machine learning, to infer the quality of experience from different usage profiles of a large-scale wireless network. The proposed approach uses the correlation between access point usage data and summaries of data flows in the network. The main idea of the proposal is to collect data on the usage of several access points, to correlate them with summaries of the connections flows that pass through the access points, reported by NetFlow, and to infer, from the application of the k-means clustering algorithm, different network usage profiles. The proposed approach was evaluated in the real wireless network, and the results show that the division of the flows in five clusters allows identifying characteristic profiles of the degraded state of the network and overload conditions in access points, considering only the flow summaries.*

1. Introdução

O uso de dados em dispositivos móveis vem crescendo a cada ano. Em 2016, as redes móveis contaram com um crescimento de 63%, alcançando um crescimento acumulado de dezoito vezes nos últimos 5 anos. Esse crescimento é marcado pela terceirização

Este trabalho foi realizado com recursos do CNPq, CAPES, RNP e FAPERJ.

do tráfego de dados de dispositivos móveis para redes fixas sem fio, que compõe 60% do tráfego móvel global [Cisco, 2017]. Assim, da popularização das redes sem fio padrão IEEE 802.11 as tornam presença constante em diversos ambientes [Mattos et al., 2019, Divgi e Chlebus, 2013]. No entanto, essa mesma popularização fomenta o desafio de controlar e gerenciar esse tipo de rede, já que um ponto de acesso típico está na vizinhança de diversos outros, sofrendo e gerando interferência [Biswas et al., 2015]. A ausência de mecanismos de inteligência e de ação rápida na rede, muitas vezes associada a visões limitadas de ferramentas de controle por fluxo, prejudicam a qualidade de experiência (QoE) de usuários de redes sem fio de grande escala [Jang et al., 2017].

A QoE do usuário de uma rede sem fio é difícil de mensurar, pois o serviço de rede percebido pelo usuário está sujeito a diversas interferências e ao tipo de conteúdo acessado. Assim, a inferência da QoE implica a necessidade de identificação de perfis de uso da rede sem fio. Em especial, a identificação do perfil de uso da rede relaciona-se com a maneira como o ponto de acesso e os usuários conectados a ele interagem com o restante da rede. Ademais, a inferência de perfis do uso de dados em redes sem fio permite definir padrões de uso [Zhang et al., 2016, Ghosh et al., 2011] e propicia que a cobrança pelo uso da rede seja de acordo com a classificação do uso [Sen et al., 2013]. Dessa forma, inferir padrões de uso da rede sem fio que considerem tanto os dados trafegados como informações de associação de usuários a pontos de acesso é essencial para a identificação de pontos de falhas e gargalos de desempenho em redes sem fio de grande escala.

Este artigo propõe uma abordagem de inferência de qualidade de experiência de usuários de rede sem fio de grande escala usando uma abordagem não supervisionada. A ideia central da proposta é gerar um conjunto de dados, correlacionando estatísticas de fluxos de cada usuário da rede sem fio com características dos pontos de acesso a que estão associados e, a partir do conjunto de dados correlacionados, extrair agrupamentos de usuários com perfis semelhante para finalmente identificar a qualidade de experiência de cada perfil extraído da rede. Para tanto, a abordagem proposta conta com uma coleta de dados de estatísticas de fluxos NetFlow, um módulo de coleta de dados da interface sem fio dos pontos de acesso da rede, um módulo de correlação de dados e, por fim, um módulo de processamento e inferência dos perfis. A inferência de perfis de uso é calculada através do algoritmo de agrupamento de dados não supervisionado *k-means*.

Trabalhos anteriores focam em modelar a chegada de novos usuários em redes sem fio [Ghosh et al., 2011], identificar possíveis gargalos de desempenho na interface sem fio da rede [Biswas et al., 2015], caracterizar o uso de dados por aplicações móveis [Qian et al., 2011, Shye et al., 2010] e analisar o desempenho em cenários restritos [Oliveira et al., 2016, Magalhães e Mattos, 2018]. Diferentemente, a proposta deste artigo é coletar tanto dados de tráfego com NetFlow, quanto métricas de associação de usuários em pontos de acesso, correlacionar os dados e, então, inferir perfis de uso. O conjunto de dados[1] gerados se baseia na rede sem fio institucional da Universidade Federal Fluminense que conta com 547 nós, atende a mais de 60 mil usuários e conta com picos de mais de 5.000 usuários conectados simultaneamente. A avaliação da proposta mostra que foram identificados cinco perfis de uso distintos da rede e as características de cada perfil relacionam-se com o tipo de serviço acessado, bytes trafegados, duração dos fluxos e a carga de uso sobre os pontos de acesso.

[1]Os dados coletados são referentes ao período de 17 a 24 de abril de 2018.

O restante do artigo está organizado da seguinte forma. A Seção 2 discute os trabalhos relacionados. A Seção 3 apresenta o problema de caracterização de perfis de uso da rede. A abordagem de inferência de qualidade de experiência é exposta na Seção 4. A proposta é avaliada no cenário de uma rede sem fio de grande escala real e os resultados são discutidos na Seção 5. A Seção 6 conclui o trabalho.

2. Trabalhos Relacionados

As aplicações de dispositivos móveis, como telefones inteligentes (*smartphones*), são as principais geradoras de fluxos em redes sem fio, em especial devido à terceirização dos tráfegos das redes de celular para a rede sem fio infraestruturada [Joe-Wong et al., 2013]. Manweiler *et al.* propõem um sistema de predição do tempo de estadia de clientes em pontos de conexão sem fio. A ideia central do sistema é aprender assinaturas de um conjunto de clientes inicial no ponto de acesso e, então, inferir o tempo de permanência de cada cliente nas proximidades do ponto de acesso sem fio. A predição utiliza diversos sensores nos celulares dos clientes, gerando uma matriz de dados que é passada para um classificador de máquina vetor de suporte (*Support Vector Machine* - SVM) que separa os clientes em classes de comportamento predeterminadas. As predições são geradas em série [Manweiler et al., 2013].

Quian *et al.* desenvolvem a ferramenta ARO (*Application Resource Optimizer*) que visa inferir e otimizar o uso de recursos por aplicações na interface de rádio de dispositivos móveis [Qian et al., 2011]. A proposta se baseia em uma abordagem entre camadas, considerando tanto a aplicação quanto o controle na camada de rádio. Em especial, a proposta desenvolve uma técnica para estimação dos estados de controle da interface de rádio e presume causas para gargalos no uso de tráfego de dados. A proposta extrai perfis de uso de aplicações populares e permite identificar que algumas aplicações têm interações ineficientes com a camada de rádio e, por conseguinte, implica gargalos no tráfego de dados e redução de desempenho nos dispositivos móveis. Por sua vez, Quiang *et al.* identificam perfis de comportamento de uso de aplicações em dispositivos móveis [Xu et al., 2011], mas, diferentemente do trabalho anterior, os autores focam no padrão de uso dos dispositivos e não no consumo de recursos de cada aplicação. Esses trabalhos relacionam-se com a inferência de perfis de uso em redes sem fio na medida em que os dispositivos móveis são os mais numerosos na rede e as aplicações mais prevalentes geram assinaturas de consumo de dados que impactam a rede sem fio de formas distintas [Shye et al., 2010].

Oliveira *et al.* argumentam que entender as atividades dos usuários em redes de acesso sem fio é essencial para garantir a escalabilidade da rede e para atender às futuras demandas de tráfego [Oliveira et al., 2016]. Portanto, os autores investigam e caracterizam as atividades de usuários em redes sem fio e comparam os padrões encontrados entre redes universitárias e redes em áreas urbanas. Os resultados das análises evidenciam que há uma relação linear entre o número de sessões de acesso sem fio e o número de pontos de acesso, mas a relação não é mantida ao se considerar em quantos pontos de acesso um usuário se conecta. Paralelamente, Biswas *et al.* verificam que a monitoração de redes sem fio exige monitorar o espectro usado por cada rede para inferir a interferência gerada por cada uma delas, pois o uso dos canais de redes WiFi é baixo perante o número de redes presentes [Biswas et al., 2015]. Tais resultados são ratificados ao analisar os dados da Universidade Federal Fluminense [Magalhães e Mattos, 2018]. Magalhães e Mattos

evidenciam que muitos usuários tendem a se associar a poucos pontos de acesso durante um dia e que, embora o número de redes sem fio no entorno de um ponto de acesso seja substancialmente alto, a interferência que cada rede gera depende da carga a que está submetida. De forma semelhante, Ghosh *et al.* caracterizam o acesso sem fio em redes de grande escala públicas e propõem um modelo de chegada de usuários [Ghosh et al., 2011]. Os autores combinam um agrupamento estático com regressão de Poisson para modelar o processo de Poisson não-estacionário que define a chegada de novos clientes sem fio. A ideia central é entender a carga típica de redes sem fio de grande escala.

Divgi *et al.* revisam a literatura sobre a caracterização de redes sem fio e analisam os dados de acesso de uma rede sem fio de âmbito nacional gerados em um período de 5 meses [Divgi e Chlebus, 2013]. Os autores caracterizam o uso comercial de redes sem fio, categorizando os usuários pelo tipo de conta de acesso, desenvolvendo métricas que mostram a flutuação da população de usuários e propondo novas métricas para contabilizar tempo de sessão e uso de dados. Já Zhang *et al.* buscam por padrões de sequência de uso em redes sem fio [Zhang et al., 2016]. Os autores inferem a sequência de padrões através de um modelo oculto de Markov. Para aplicar o modelo, os dados são divididos em agrupamentos de usuários usando o método de *k-metoids*. O número ideal de agrupamentos é definido usando a estatística de intervalo (*Gap Statistic*). Nenhum dos trabalhos mencionados relaciona estatísticas de fluxos e métricas de uso da interface sem fio.

Jang *et al.* propõem o sistema RFlow$^+$ [Jang et al., 2017]. Os autores argumentam que a ausência de mecanismos inteligentes e que possam tomar ações rápidas nas redes sem fio associada a ferramentas de monitoramento de tráfego que geram visões limitadas sobre a rede prejudicam a confiabilidade das redes sem fio. Assim, o sistema RFlow$^+$ combina a ideia de redes definidas por *software* e monitoramento de redes sem fio para prover um arcabouço de gerenciamento e controle que permita a tomada de ações rápidas na rede. Para tanto, os autores propõem dois níveis de contadores de dados de fluxo, um local e outro global. Os dados desses contadores de fluxos são usados no controle da rede. Paralelamente, Dely *et al.* argumentam que devido à sobreposição da área de cobertura de pontos de acesso de uma mesma rede, a otimização da rede pode causar a troca constante de associação entre usuário e ponto de acesso [Dely et al., 2012]. Assim, os autores introduzem um problema de otimização linear inteira mista que visa reduzir o números de trocas de reassociação ao considerar o custo da reassociação. Balbi *et al.* abordam os problemas da instabilidade na associação propondo um processo simples de suavização da série temporal de amostras de sinal de recepção [Balbi et al., 2016]. Diferentemente, a proposta deste artigo aborda o gerenciamento de redes sem fio sob a perspectiva da análise de dados e correlaciona os dados de fluxos unitários, colhidos por NetFlow, a dados de associação de usuários a pontos de acesso. A proposta analisa o conjunto de dados contendo tanto dados da interface sem fio quanto os fluxos na rede de transporte para extrair conhecimento útil no controle e monitoramento das redes sem fio.

3. A Inferência de Perfis de Uso da Rede

Inferir perfis de uso da rede consiste em analisar os dados de uso da rede e descobrir padrões que se repetem em diferentes usuários. Os perfis podem apontar usos característicos de determinadas aplicações [Shye et al., 2010] e o quanto de recursos é consumido por cada aplicação [Qian et al., 2011]. Contudo, em trabalhos anteriores a busca por padrões de uso de aplicações ou padrões de uso de recursos de rede são realizadas de

forma supervisionada, isto é, conjuntos de dados característicos de cada aplicação são usados para treinar um classificador capaz de reconhecer os padrões em um conjunto de dados de teste [Andreoni Lopez et al., 2017]. Ao identificar padrões anteriormente ocultos de uso da rede é possível extrair não somente conhecimento a respeito do funcionamento da rede [Biswas et al., 2015, Ghosh et al., 2011], mas também a respeito dos usuários, como preferências e padrões de mobilidade [Wang et al., 2014, Guo et al., 2014].

O problema da inferência de perfis não supervisionada é complexo, pois *a priori* não são conhecidos os padrões de uso da rede e, assim, são usados algoritmos de agrupamento não supervisionado para identificação de padrões recorrentes [Zhang et al., 2016]. O agrupamento de amostras de dados em conjuntos que definem um padrão é um problema NP-difícil [Zhang et al., 2016]. Os algoritmos de aprendizado de máquina para agrupamento não supervisionado visam buscar vetores que permitam representar agrupamentos de dados com maior espaçamento entre eles e, simultaneamente, reduzindo a distância das amostras até o centro dos agrupamentos. O algoritmo de agrupamento *k-means*, por exemplo, busca k vetores que minimizam a distância de todas as amostras aos vetores centrais de cada agrupamento. O algoritmo atribui o vetor de dados a exatamente um agrupamento. A cada rodada, o algoritmo recalcula o centro de cada agrupamento e termina quando as atribuições aos agrupamentos não mudam entre duas rodadas.

Ao se considerar uma rede sem fio de grande escala, como a rede institucional da Universidade Federal Fluminense, a detecção de padrões de uso torna-se computacionalmente custosa devido à grande massa de dados gerada pelo uso contínuo da rede. Outro fator complicador é que os dados necessários para inferir a qualidade de experiência dos usuários são dispersos em diversas fontes de registros de atividades (*logs*) não correlacionadas. Vale ressaltar que na rede considerada há 547 pontos de acesso, gerando registros de associação e desassociação de usuários, e os fluxos de dados gerados pelos usuários são reportados pelos roteadores de saída da rede usando NetFlow. Contudo, as informações de associação de usuários não se relacionam diretamente com os relatórios NetFlow, já que as informações de associação carregam dados da camada de enlace, enquanto os relatórios NetFlow carregam dados das camadas de rede e superiores. Assim, a variedade de fontes de dados é um desafio para inferir os perfis de uso em redes de grande escala. Dessa forma, com dados sendo gerados em alta velocidade, grande massa de dados e variabilidade nos dados, a inferência de perfis de uso é um problema de *Big Data*.

4. A Abordagem Proposta de Geração de Perfis do Uso da Rede Sem Fio

A rede sem fio institucional da Universidade Federal Fluminense (UFF) é distribuída por mais de 90 prédios em 16 *campi* universitário distintos. A infraestrutura da rede sem fio é baseada no *software* livre SCIFI (Sistema de Controle Inteligente para redes sem fio), originado de um grupo de trabalho da RNP (Rede Nacional de Ensino e Pesquisa) - o GT SCIFI. O SCIFI é composto por um controlador de *software* e um *firmware* instalado em pontos de acesso de baixo custo. O *software* controlador configura os pontos de acesso para usarem canais que minimizem a interferência entre pontos de acesso SCIFI e, também, evitem a interferência com outras redes [Balbi et al., 2012, Magalhães e Mattos, 2018]. A infraestrutura de rede conta com 5 *gateways* de saída de tráfego da rede sem fio para a Internet. Os *gateways* de conexão da rede sem fio com a Internet concentram o tráfego dos maiores *campi*. Este trabalho foca na análise do tráfego de um *campus* da UFF, Praia Vermelha, no período de 17 a 24 de

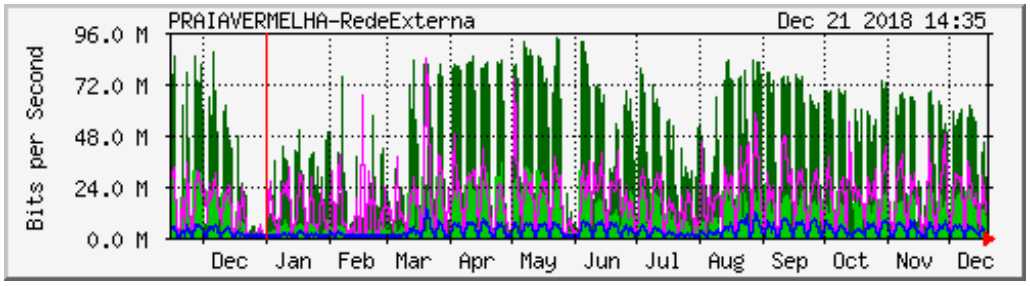


Figura 1. Tráfego encaminhado no *campus* Praia Vermelha da Universidade Federal Fluminense no ano de 2018.

abril de 2018, identificando um total de 6.770 dispositivos únicos que acessaram a rede através de um subconjunto de 363 pontos de acesso. A Figura 1 mostra o tráfego encaminhado no *campus* Praia Vermelha. A infraestrutura nesse *campus* realiza o roteamento de pacotes a taxas de pico próximas a 100 Mb/s. Vale ressaltar que os dados da Figura 1 representam somente o uso de dados da rede sem fio de um único *campus* [2]. No contexto da rede institucional da UFF, a abordagem proposta consiste em analisar os dados de fluxos gerados pelos usuários da rede e correlacionar com dados de associação entre usuários e ponto de acesso. Dessa forma, o *firmware* e o *software* de controle SCIFI são essenciais para a coleta de dados.

A abordagem proposta se baseia na análise de estatísticas de fluxos da rede. O fluxo de rede pode ser definido como uma sequência unidirecional ou bidirecional de pacotes entre dois nós da rede com algum atributo em comum. Comumente, define-se os atributos em comum como a 5-tupla: endereços de origem e destino, protocolo de transporte e portas de origem e destino [Andreoni Lopez et al., 2017, Mattos et al., 2019]. Os campos capturados na descrição dos fluxos da rede fornecem um conjunto avançado de estatísticas de tráfego, incluindo usuário, protocolo, porta e tipo de serviços que podem ser usados para uma análise ampla com diferentes propósitos, como segurança de rede, monitoramento de rede, análise de tráfego, planejamento de capacidade, classificação de tráfego, contabilidade e faturamento. Ao se considerar a caracterização através da tecnologia NetFlow, o processo geral inclui capturar, amostrar, gerar, exportar, coletar, analisar e visualizar [Li et al., 2013].

NetFlow é uma tecnologia de monitoramento de tráfego desenvolvida na Cisco e define como um roteador exporta informações e estatísticas de fluxos de rede. Os dispositivos de rede examinam os pacotes que chegam nas interfaces e capturam estatísticas de tráfego por fluxo com base na configuração para amostragem ou filtragem, criam um cache do fluxo, agregam e exportam os dados por meio dos protocolos UTP ou SCTP. A entrada de cache do NetFlow é criada pelo primeiro pacote de um fluxo, mantida para características de fluxo semelhantes e exportada para coletores periodicamente com base em temporizadores de fluxo ou no gerenciamento de cache de fluxo. Ao usar amostragem em NetFlow, são introduzidos alguns novos desafios, tais como não contabilizar novos fluxos quando o cache estiver cheio; a redução da precisão na amostragem dos fluxos, especialmente quando a taxa de amostragem é ajustada pela taxa de tráfego; e discrepância na ordem em que os registros de fluxo são exportados em relação à ordem em que o tráfego

[2]Os dados apresentados são reportados pela ferramenta MRTG que consolida informações colhidas por SMNP. Os dados estão disponíveis em http://200.20.0.200/mrtg.

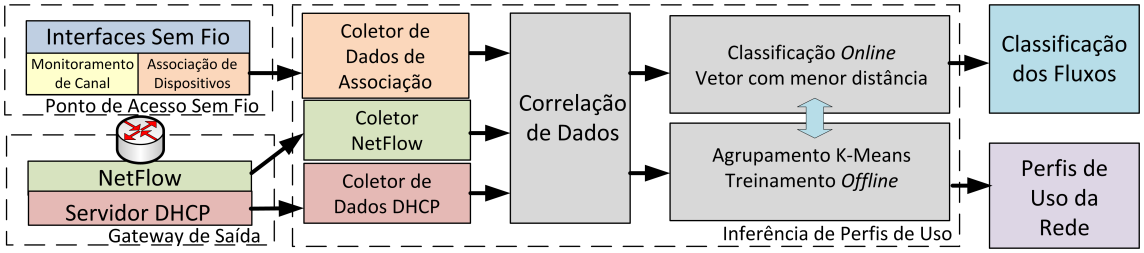


Figura 2. A abordagem proposta consiste na coleta e correlação de estatísticas de fluxos NetFlow e dados de associação de clientes móveis com pontos de acesso. Os dados são correlacionados para a geração de resumos bidirecionais de fluxos e a identificação do dispositivo móvel e do ponto de acesso que geram os dados. Os módulos de processamento *online* e *offline* geram inferências em tempo real do perfil de uso de cada fluxo na rede.

do fluxo chega ao roteador. Existem variedades de coletores NetFlow e ferramentas de análise de fornecedores comerciais. Neste artigo, todos os fluxos são contabilizados, pois é realizada uma amostragem individual de cada fluxo e a taxa de transmissão de linha é inferior à capacidade de encaminhamento do *gateway* de conexão com a Internet, já que o tráfego atinge 100 Mb/s, enquanto o *gateway* tem capacidade para rotear até 1 Gb/s.

A abordagem proposta é evidenciada na Figura 2 e consiste de módulos de coleta de dados, módulos de correlação e módulos de processamento *online* e *offline*. Após o processamento dos dados, o conhecimento extraído é analisado. Os módulos de **coleta de dados** são divididos em duas partes. A primeira parte executa nos elementos de rede, sejam os pontos de acesso, sejam os *gateways* de acesso à Internet, e é responsável por executar a coleta local de dados. Nos pontos de acesso, são coletados os dados de associação de novos dispositivos através de mensagens do *hostapd*[3], o ponto de acesso por *software* executado pelo *firmware* SCIFI. Os *gateways* realizam a coleta de dois conjuntos de dados importantes, os registros de entrega de endereços IP pelo DHCP e as estatísticas dos fluxos roteados na rede. A coleta dos dados locais gerados nos pontos de acesso e dos dados de entrega de IP nos *gateways* é realizada pelo *rsyslog* [4]. Já a coleta das estatísticas NetFlow é realizada pela ferramenta de coleta e análise `SiLK` [5]. Após a coleta, todos os dados são concentrados em um servidor de gerenciamento.

O módulo de **correlação de dados** limpa, funde e correlaciona os dados gerados pelas diversas fontes. O foco desse módulo é gerar, para cada registro de fluxo reportado pelo NetFlow, características extras, como a identificação do dispositivo do cliente que realiza o fluxo, o ponto de acesso em que o cliente está conectado no momento em que realiza o fluxo e o número de clientes ativos, ou seja, que estão com fluxos ativos, no mesmo ponto de acesso. Para considerar o número de clientes ativos, foi avaliado o número de fluxos ativos de clientes distintos em uma janela de 10 s. Ao realizar a fusão de dados, o módulo agrega mais informações à análise e permite identificar correlações entre características distintas.

O módulo de **processamento dos dados** é particionado em processamento *online* e em processamento *offline*. O processamento *online* realiza classificação dos registros en-

[3]Disponível em http://w1.fi/hostapd/.
[4]Disponível em https://www.rsyslog.com/.
[5]Disponível em https://tools.netsa.cert.org/silk/.

trantes conforme chegam ao módulo. Essa classificação é realizada de acordo com os vetores de identificação dos agrupamentos de dados calculados pelo processamento *offline*. O processamento *offline* realiza a análise de dados históricos armazenados. Para tanto, nesse trabalho, a análise realizada é não supervisionada e consiste no pré-processamento dos dados e na aplicação do algoritmo *k-means* para realizar o agrupamento. O pré-processamento realizado exclui características textuais, constantes ou vazias reportadas tanto pelo NetFlow quanto pelo módulo de correlação de dados.

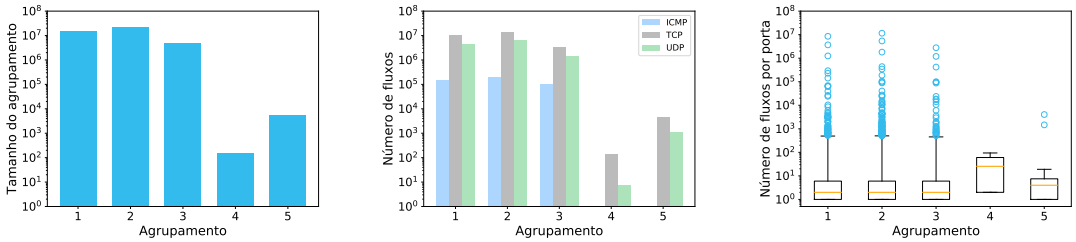
Após o processamento dos dados, é possível identificar diferentes padrões de uso mais prevalentes na rede. Assim, a proposta reporta a classificação dos fluxos *online*, *i.e.*, assim que são correlacionados e classificados. Ademais, quando executada em tempo diferenciado sobre uma base histórica de uso da rede, a abordagem proposta permite identificar padrões prevalentes na rede que são as inferências de perfis de uso da rede. A inferência desses perfis implica em maior distância[6] entre os vetores que representam cada agrupamento. Sendo assim, os perfis inferidos são os que permitem maior diferenciação do uso da rede. Vale ressaltar que a inferência de qualidade de experiência está associada à análise dos perfis representados pelos agrupamentos identificados pelo algoritmo *k-means*. O algoritmo *k-means* é usado pois é simples e não paramétrico, o que permite a extração mais imediata de perfis sem o conhecimento prévio dos dados. A qualidade da experiência consiste em analisar os dados de cada agrupamento e inferir o que os clientes classificados naquele perfil têm em comum. Em especial, verifica-se a quantidade de clientes associados a um mesmo ponto de acesso como métrica de competição por recursos e a banda de transmissão média alcançada pelos clientes em cada agrupamento.

5. Resultados Experimentais

A avaliação da abordagem proposta foi realizada sobre os dados coletados da rede sem fio do *campus* Praia Vermelha, da Universidade Federal Fluminense. Para tanto, os pontos de acesso foram configurados para reportar os dados para o servidor responsável por executar o *software* controlador SCIFI da rede. Os dados coletados foram consolidados e processados em um servidor equipado com processador Intel Core i5 4460, com 4 núcleos de processamento, e 32 GB de memória RAM. O processamento dos dados *offline* foi implementado em Python com uso das bibliotecas Pandas[7] e Sklearn[8] e com apoio da ferramenta KNIME 3.5.3[9]. Em uma análise preliminar, foi executada a estatística de intervalo (*Gap Statistic*) para determinar o número de agrupamentos ideal [Zhang et al., 2016] para descrever o conjunto de dados. A estatística de intervalo busca o número de agrupamentos que maximiza a distância entre os agrupamentos, enquanto minimiza a distância entre as amostras em um mesmo agrupamento. A estatística é projetada para ser agnóstica ao algoritmo de agrupamento utilizado. A estatística de intervalo foi executada em uma amostra do conjunto de dados e resultou em 5 agrupamentos. Portanto, nas demais avaliações foram considerados 5 agrupamentos.

As análises a seguir são apresentadas prioritariamente em relação aos fluxos de descarga de conteúdos da Internet para a rede sem fio. A análise foca nos fluxos de descarga, pois como os usuários da rede sem fio estão atrás de *firewall* e usando endereços não

[6] A proposta considera a distância euclidiana como métrica para indicar a separação entre agrupamentos.
[7] Disponível em https://pandas.pydata.org/.
[8] Disponível em https://scikit-learn.org/stable/.
[9] Disponível em https://www.knime.com/.



(a) Tamanho em número de fluxos dos agrupamentos.

(b) Distribuição de protocolos por agrupamento.

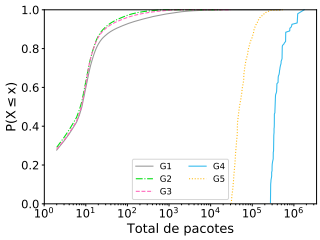
(c) Distribuição do uso de portas por agrupamento.

Figura 3. Agrupamentos calculados pelo algoritmo *k-means*. Gráficos apresentados em escala logarítmica. a) Os agrupamentos 1, 2 e 3 são os mais prevalentes, concentrando maior número de fluxos. b) Os perfis de tráfego dos agrupamentos 1, 2 e 3 são semelhantes, concentrando fluxos TCP, UDP e ICMP. c) Distribuição do números de fluxos em cada porta nos agrupamentos calculados. Os *outliers* indicam que poucas portas têm número de fluxos muito superior à média, indicando a tendência no comportamento do agrupamento.

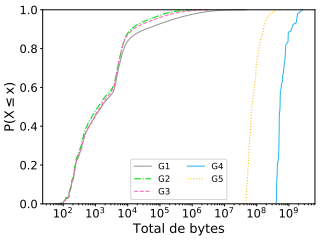
roteáveis locais fornecidos por um NAT (*Network Address Translation*), não é possível que um nó externo à rede sem fio abra uma conexão com um usuário da rede sem fio. Ademais, caso existam fluxos entre usuários da rede sem fio, esses fluxos não são reportados pelo NetFlow, já que a coleta de dados é realizada pelo *gateway* de saída da rede sem fio para a Internet.

Após a execução do algoritmo *k-means*, os dados do conjunto de dados foram classificados de forma não supervisionada nos 5 agrupamentos. Cada agrupamento define um perfil de uso da rede. Três agrupamentos contêm da ordem de 10^7 fluxos em cada, conforme mostrado na Figura 3(a), enquanto os dois outros agrupamentos contêm menos fluxos, da ordem de 10^4 e 10^2 fluxos. Quando comparadas as proporções do uso de diferentes protocolos de transporte em cada agrupamento, conforme Figura 3(b), verifica-se que o TCP é o protocolo mais comum em todos os agrupamentos. Contudo, há uma presença constante de sinalização ICMP nos agrupamentos 1, 2 e 3. Nos três agrupamentos, o número de fluxos ICMP é semelhante. Diferentemente, nos agrupamentos 4 e 5 os fluxos do protocolo ICMP são ausentes e destaca-se que no agrupamento 4 há uma maior predominância de fluxos TCP, o que indica um perfil diferente do agrupamento 4 em relação aos demais. Na Figura 3(c), analisa-se o uso das portas de destino dos fluxos na rede, pois as portas de destino evidenciam quais serviços são acessados pelos usuários da rede sem fio. Verificou-se que as portas mais prevalentes em todos os fluxos são as UDP 53, TCP 80 e TCP 443 [Andreoni Lopez et al., 2017]. Tal comportamento é o esperado, pois os acesso mais comuns são a conteúdo *web* ou a serviços que usam chamadas *web*, tais como aplicativos de redes sociais ou mensagens instantâneas. O comportamento de maior uso dessas portas nos agrupamentos é marcado pela presença de *outliers* na Figura 3(c). De forma semelhante à análise anterior, os agrupamentos 4 e 5 se destacam por apresentarem menos *outliers* e uma distribuição maior do uso de portas.

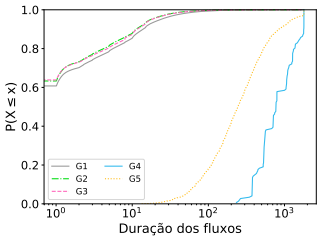
Ao avaliar a distribuição de probabilidades de número de pacotes por agrupamento, Figura 4(a), *bytes* trafegados por cada fluxo no agrupamento, Figura 4(b), e a duração dos fluxos, Figura 4(c), verifica-se uma clara diferenciação entre os fluxos nos agrupamentos 1, 2 e 3 para os demais fluxos nos agrupamentos 4 e 5. Os fluxos nos agru-



(a) Distribuição do número de pacotes por agrupamento.

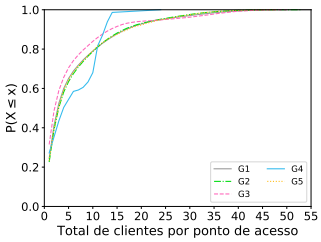


(b) Distribuição do tamanho em bytes por agrupamento.

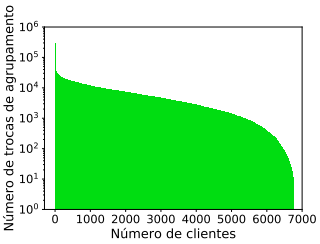


(c) Distribuição da duração dos fluxos por agrupamento.

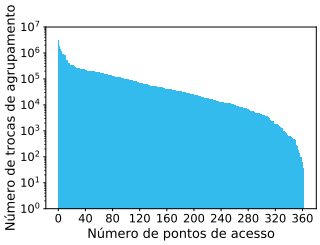
Figura 4. Probabilidade cumulativa da distribuição de caracterização dos agrupamentos. a) Agrupamentos 1, 2 e 3 tendem a ter menos pacotes que os agrupamentos 4 e 5. b) De forma semelhante, os agrupamentos 4 e 5 são responsáveis pelo tráfego de maior quantidade de *bytes*. c) O agrupamento 4 é o que tende a ter os fluxos mais duradouros.



(a) Distribuição do número de clientes por ponto de acesso em cada agrupamento.



(b) Número de trocas de classificação para cada cliente.

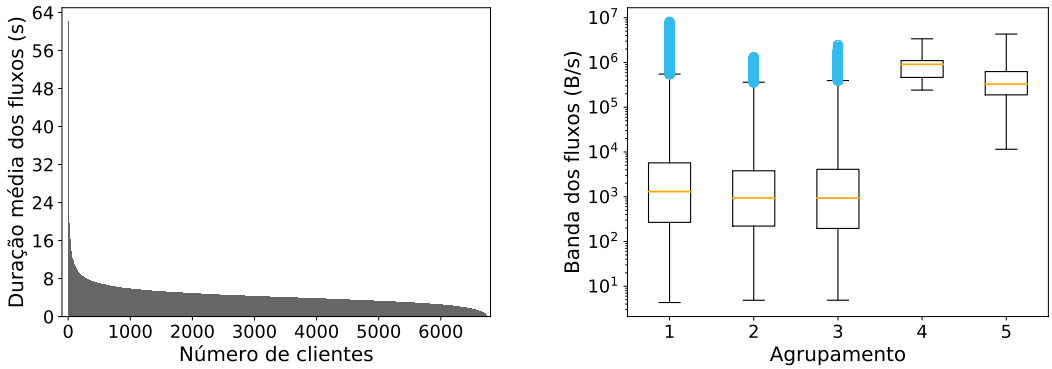


(c) Número de trocas de classificação para cada ponto de acesso.

Figura 5. Caracterização do comportamento dos clientes nos agrupamentos. a) Probabilidade cumulativa do número de clientes conectados a um mesmo ponto de acesso por agrupamento. b) Número de vezes que um cliente tem seu perfil recalculado como outro agrupamento na rede. c) Número de vezes que um ponto de acesso tem o agrupamento de que participa trocado.

pamentos 1, 2 e 3 tendem a ser mais curtos e, portanto, trafegam menos pacotes e menor quantidade de *bytes*. Enquanto os fluxos dos agrupamentos 4 e 5 são fluxos mais duradouros com maior quantidade de *bytes* trafegados. Observando a diferença de duração entre os fluxos dos agrupamentos 4 e 5, verifica-se que a probabilidade de apresentar fluxos longos converge, ao passo que ao avaliar o número de pacotes e *bytes*, verifica-se que os fluxos do agrupamento 4 tendem a apresentar maior volume de dados. Essas características indicam que fluxos do agrupamento 4 tendem a apresentar maior taxa de transmissão que os do agrupamento 5.

Ao considerar o número de clientes que acessam um ponto de acesso simultaneamente, há uma diferenciação no comportamento dos agrupamentos calculados. A Figura 5(a) mostra que o agrupamento 3 tende a ter menos clientes conectados por ponto de acesso que os demais agrupamentos. O agrupamento 4, que conta com fluxos com maior duração do que os demais agrupamentos, tende a ter mais clientes conectados por ponto de acesso do que qualquer outro agrupamento em até 80% dos fluxos. Contudo, o agrupamento 4 atinge o limite de clientes conectados por ponto de acesso, com aproximadamente 25 clientes por ponto de acesso, acima desse valor, não há mais fluxos classificados no



(a) Duração média dos fluxos por cliente.

(b) Taxa de transmissão alcançada em cada agrupamento.

Figura 6. Avaliação dos perfis de uso. a) Distribuição da duração dos fluxos por clientes identificados na rede. A duração segue uma distribuição de cauda longa. b) Avaliação da taxa de transmissão alcançada em cada agrupamento indica que os agrupamentos 4 e 5 apresentam as taxas de transmissão mais elevadas.

agrupamento 4. Acima de 25 usuários o canal sem fio passa a ser um entrave para manter um fluxo ativo por grande período de tempo e com alta taxa de encaminhamento de dados. Esse resultado reforça a ideia de que a previsão máxima ideal de carga para pontos de acessos é entre 20 e 30 usuários. Os agrupamentos 1, 2 e 5 obtiveram resultados semelhantes quanto ao uso concomitante dos pontos de acesso.

As Figuras 5(b) e 5(c) evidenciam a troca de classificação entre agrupamentos por cliente e por ponto de acesso respectivamente. Vale ressaltar que o comportamento do número de trocas de classificação entre agrupamentos segue a mesma distribuição para clientes e para pontos de acesso. A métrica de trocas de classificação considera que a classificação por agrupamento é atribuída ao cliente ou ao ponto de acesso e, então, verifica ao longo do conjunto de dados quantas vezes há mudança na classificação atribuída ao cliente ou ao ponto de acesso. A ideia central dessa métrica é avaliar se há dependência entre a classificação e cliente ou ponto de acesso. No entanto, como a distribuição das trocas de classificação é semelhante entre clientes e pontos de acesso, verifica-se que a classificação é independente desses fatores, e, assim, a classificação é fortemente dependente dos fluxos que são realizados na rede.

A Figura 6(a) revela a duração média dos fluxos por cliente. Esse resultado evidencia que a duração dos fluxos segue uma distribuição de cauda longa, em que poucos clientes executam fluxos longos, enquanto a maioria dos clientes executam fluxos de curta duração. Tal resultado relaciona-se com a caracterização do tráfego da rede que aponta que os fluxos mais prevalentes são os de consulta DNS e fluxos de requisição HTTP e HTTPS. Por fim, foi avaliada a taxa de transmissão alcançada pelos fluxos em cada agrupamento. Os resultados mostrados na Figura 6(b) revelam que os agrupamentos 4 e 5 são os que apresentam as taxas de transmissão mais altas, com destaque para o agrupamento 4 que agrega fluxos com taxas de transmissão da ordem de 8 Mb/s. Os agrupamentos 1, 2 e 3 apresentam taxas de transmissão variáveis, com mediana próxima a 10 kb/s.

Vale notar que após a caracterização dos agrupamentos calculados pelo algoritmo *k-means* é possível definir que cada agrupamento representa um perfil de uso da rede sem

fio e, a partir das análises, define-se níveis de qualidade de experiência para os clientes. Assim, verifica-se que clientes que têm fluxos classificados nos agrupamentos 4 e 5 são aqueles que experimentam uma boa qualidade de serviço na rede. Em especial, os clientes que têm fluxos classificados nos agrupamentos 4 e 5 têm acesso a uma taxa de transmissão da ordem de alguns Mb/s. Ao verificar que um cliente tem fluxos classificados no agrupamento 4 é possível inferir, com mais de 90% de certeza, que o cliente está conectado em um ponto de acesso com menos de 25 clientes, dado o resultado exposto na Figura 5(a). Um cliente, ao ter seus fluxos classificados nos agrupamentos 1, 2 e 3, pode estar passando por dificuldades de conexão devido a excesso de uso na rede sem fio ou, simplesmente, pode estar usando serviços que exigem poucos recursos da rede sem fio, tais como redes sociais e aplicativos de mensagens. Esses serviços se caracterizam por fluxos de curta duração e pouco tráfego de dados. Portanto, o impacto na rede é reduzido. Esse comportamento é mais marcante em fluxos dos agrupamentos 2 e 3 que tendem a ser menores que os do agrupamento 1.

Outro ponto interessante ressaltado pela Figura 5(a) é que os fluxos do agrupamento 3 tendem a ser de clientes que estão associados a pontos de acesso com menos clientes conectados e, portanto, experimentam condições de rede melhores que os clientes dos agrupamentos 1 e 2. Sendo assim, é possível inferir que clientes que são classificados no agrupamento 4 são aqueles que experimentam as melhores condições da rede sem fio, seguidos pelos clientes do agrupamento 5 e, em seguida, pelos do agrupamento 3. Os clientes dos agrupamentos 1, 2 e 3 têm perfis de uso semelhantes, mas os clientes do agrupamento 3 experimentam redes menos congestionadas na interface sem fio, pois em 90% das vezes estão conectados em pontos de acesso abaixo da situação de sobrecarga. Os clientes dos agrupamentos 1 e 2 experimentam a rede em condições mais adversas que os demais, mas os clientes do agrupamento 2 atingem quantidades trafegadas de dados maiores que os do agrupamento 1, indicando uma melhor experiência na rede. Já os clientes do agrupamento 1 experimentam o uso da rede sem fio prioritariamente para uso de fluxos curtos e que trafegam pouca quantidade de dados.

6. Conclusão

O cálculo de perfis de uso da rede para inferir a qualidade de experiência de clientes somente com base em estatísticas de fluxos e dados de associação de dispositivos a pontos de acesso é um problema complexo. As estatísticas de fluxos não revelam questões importantes da rede sem fio como o uso do espectro e a quantidade de clientes que compartilham o mesmo canal. Esses desafios são ainda mais críticos em redes sem fio de grande escala em que há concentração de usuários em locais e períodos específicos. Este artigo propôs uma abordagem não supervisionada de geração de perfis de uso da rede sem fio para a inferência da qualidade de experiência do usuário final. Para tanto, a abordagem proposta correlaciona dados de diversas fontes, como estatísticas de fluxos NetFlow e dados de associação de clientes a pontos de acesso reportados pelos próprios pontos de acesso. A correlação dos dados ainda usa relatórios da distribuição de endereços IP pelo servidor DHCP para identificar os clientes e seus fluxos. A abordagem foi empregada para analisar o conjunto de dados de uso da rede sem fio institucional da Universidade Federal Fluminense e identificou 5 perfis distintos de uso da rede. Dentre os perfis identificados, é possível verificar que o perfil com melhor qualidade de experiência é associado a fluxos de usuários que estão conectados a pontos de acesso com menos de 25 usuários e, em até

90% dos casos, com menos de 15 usuários. O perfil de melhor qualidade de experiência alcança taxa de transmissão da ordem de 10 Mb/s, enquanto o perfil com menor qualidade de experiência apresenta taxa de transmissão bastante variável. O perfil com menor qualidade de experiência registra picos de mais de 50 clientes conectados em um mesmo ponto de acesso. Contudo, esse perfil se caracteriza por fluxos rápidos e com pouca transmissão de dados típico de aplicativos de mensagens instantâneas ou redes sociais. Como trabalhos futuros, pretende-se estender a proposta para o processamento *online* de fluxos para o cálculo de novos perfis e a adaptação de perfis existentes em tempo real.

Referências

[Andreoni Lopez et al., 2017] Andreoni Lopez, M., Silva, R. S., Alvarenga, I., Rebello, G., Sanz, I. J., Lobato, A., Mattos, D., Duarte, O. C. M. B. e Pujolle, G. (2017). Collecting and characterizing a real broadband access network traffic dataset. Em *2017 1st Cyber Security in Networking Conference (CSNet'17)*, Rio de Janeiro, Brazil.

[Balbi et al., 2012] Balbi, H., Fernandes, N., Souza, F., Carrano, R., Albuquerque, C., Muchaluat-Saade, D. e Magalhaes, L. (2012). Centralized channel allocation algorithm for ieee 802.11 networks. Em *2012 Global Information Infrastructure and Networking Symposium (GIIS)*, p. 1–7.

[Balbi et al., 2016] Balbi, H., Passos, D., Carrano, R., Magalhaes, L., e Albuquerque, C. (2016). Análise e solução para o problema da instabilidade de associação em redes IEEE 802.11 densas. Em *XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC'2016*.

[Biswas et al., 2015] Biswas, S., Bicket, J., Wong, E., Musaloiu-E, R., Bhartia, A. e Aguayo, D. (2015). Large-scale measurements of wireless network behavior. Em *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, SIGCOMM '15, p. 153–165, London, United Kingdom. ACM.

[Cisco, 2017] Cisco, V. N. I. (2017). Global mobile data traffic forecast update, 2016–2021 white paper. *Document ID*, 1454457600805266.

[Dely et al., 2012] Dely, P., Kassler, A., Bayer, N., Einsiedler, H. e Peylo, C. (2012). Optimization of wlan associations considering handover costs. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):255.

[Divgi e Chlebus, 2013] Divgi, G. e Chlebus, E. (2013). Characterization of user activity and traffic in a commercial nationwide wi-fi hotspot network: global and individual metrics. *Wireless Networks*, 19(7):1783–1805.

[Ghosh et al., 2011] Ghosh, A., Jana, R., Ramaswami, V., Rowland, J. e Shankaranarayanan, N. K. (2011). Modeling and characterization of large-scale wi-fi traffic in public hot-spots. Em *2011 Proceedings IEEE INFOCOM*, p. 2921–2929.

[Guo et al., 2014] Guo, X., Chan, E. C. L., Liu, C., Wu, K., Liu, S. e Ni, L. M. (2014). Shopprofiler: Profiling shops with crowdsourcing data. Em *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, p. 1240–1248.

[Jang et al., 2017] Jang, R., Cho, D., Noh, Y. e Nyang, D. (2017). Rflow+: An sdn-based wlan monitoring and management framework. Em *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, p. 1–9.

[Joe-Wong et al., 2013] Joe-Wong, C., Sen, S. e Ha, S. (2013). Offering supplementary wireless technologies: Adoption behavior and offloading benefits. Em *2013 Proceedings IEEE INFOCOM*, p. 1061–1069.

[Li et al., 2013] Li, B., Springer, J., Bebis, G. e Gunes, M. H. (2013). A survey of network flow applications. *Journal of Network and Computer Applications*, 36(2):567 – 581.

[Magalhães e Mattos, 2018] Magalhães, L. C. S. e Mattos, D. M. F. (2018). Caracterização do uso de uma rede sem fio de grande porte distribuída por uma ampla Área. *XVII Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPerformance - CSBC 2018)*, 17(1/2018).

[Manweiler et al., 2013] Manweiler, J., Santhapuri, N., Choudhury, R. R. e Nelakuditi, S. (2013). Predicting length of stay at wifi hotspots. Em *Proceedings IEEE INFOCOM 2013*, p. 3102–3110.

[Mattos et al., 2019] Mattos, D. M. F., Velloso, P. B. e Duarte, O. C. M. B. (2019). An agile and effective network function virtualization infrastructure for the Internet of Things. *Journal of Internet Services and Applications*, 10(1):6.

[Oliveira et al., 2016] Oliveira, L., Obraczka, K. e Rodríguez, A. (2016). Characterizing user activity in wifi networks: University campus and urban area case studies. Em *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '16, p. 190–194, Malta. ACM.

[Qian et al., 2011] Qian, F., Wang, Z., Gerber, A., Mao, Z., Sen, S. e Spatscheck, O. (2011). Profiling resource usage for mobile applications: A cross-layer approach. Em *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, p. 321–334, Bethesda, Maryland, USA. ACM.

[Sen et al., 2013] Sen, S., Joe-Wong, C., Ha, S. e Chiang, M. (2013). A survey of smart data pricing: Past proposals, current plans, and future trends. *ACM Comput. Surv.*, 46(2):15:1–15:37.

[Shye et al., 2010] Shye, A., Scholbrock, B., Memik, G. e Dinda, P. A. (2010). Characterizing and modeling user activity on smartphones: Summary. Em *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '10, p. 375–376, New York, New York, USA. ACM.

[Wang et al., 2014] Wang, Y., Yang, J., Chen, Y., Liu, H., Gruteser, M. e Martin, R. P. (2014). Tracking human queues using single-point signal monitoring. Em *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '14, p. 42–54, New York, NY, USA. ACM.

[Xu et al., 2011] Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J. e Venkataraman, S. (2011). Identifying diverse usage behaviors of smartphone apps. Em *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, p. 329–344, Berlin, Germany. ACM.

[Zhang et al., 2016] Zhang, X., Wang, C., Li, Z., Zhu, J., Shi, W. e Wang, Q. (2016). Exploring the sequential usage patterns of mobile internet services based on markov models. *Electronic Commerce Research and Applications*, 17:1 – 11.