Um Estudo Sobre Compartilhamentos entre Contatos via D2D em Serviços de Armazenamento Pessoal em Nuvem

Denilson da Silva Sousa, Glauber Dias Gonçalves

[1] Universidade Federal do Piauí - Campus Picos (UFPI)
denilsondsousa@ufpi.edu.br, ggoncalves@ufpi.edu.br

Resumo. *Serviços de armazenamento pessoal em nuvem são utilizados frequentemente para compartilhamentos de dados, o que os tornam atrativos para usuários, mas aumentam seus custos com transmissões de dados entre vários dispositivos de usuários na Internet. Nesse artigo, investigamos o potencial de compartilhamento de dados entre contatos via o uso da comunicação direta de dispositivo para dispositivo (D2D) para transmissão de dados. Medimos a ocorrência e a duração de encontros entre contatos em compartilhamentos, baseado em dados do Dropbox. Propomos modelos estatísticos para essas medições e mostramos oportunidades amplas para comunicação D2D, dado a alta frequência de encontros com durações longas, superiores a 15 minutos.*

Abstract. *Personal cloud storage services are often used for data sharing, which makes them attractive to users, but increases their costs through data transmissions among various user devices on the Internet. In this article, we investigate the potential for data sharing between contacts via the use of direct device-to-device (D2D) communication for data transmission. We measure the occurrence and duration of meetings between contacts on shares, based on data from Dropbox. We propose statistical models for these measurements and show wide opportunities for D2D communication, given the high frequency of meetings and durations that are longer than 15 minutes.*

1. Introdução

Armazenamento de dados pessoais na nuvem [Gonçalves et al. 2016], é um serviço de Internet popular que oferece uma maneira conveniente e confiável de armazenar dados de usuários domésticos ou corporativos na nuvem para *backup*. Adicionalmente, esse serviço oferece a opção de compartilhar arquivos entre usuários distintos, em tempo quase real, para entretenimento ou trabalho colaborativo. Isso é possível a partir da sincronização automática de dados armazenados na nuvem entre vários dispositivos do usuário como PCs, tablets e smartphones. Essas vantagens oferecidas aos usuários refletem as tendências de crescimento desse serviço, com a previsão de 1,3 ZB de dados armazenados na nuvem em 2022, um crescimento de 4,6 vezes comparado a 2016 [Cisco 2019].

O compartilhamento de dados é um recurso essencial para os usuários de serviços de armazenamento em nuvem, como já demonstraram alguns estudos [Palviainen and Rezaei 2015, Gracia-Tinedo et al. 2016]. De fato, os maiores provedores desse serviço têm um crescente interesse em promover a interação entre os usuários registrados e seus respectivos *contatos* (familiares, amigos, colegas de trabalho), como

uma forma de aumentar a base de usuários registrados no serviço. A exemplo disso, o Dropbox propôs a versão comercial de seu serviço focada no compartilhamento, enquanto o Google Drive procura incentivar o trabalho colaborativo por meio de edição de texto, planilhas e apresentações diretamente na nuvem em tempo real.[1]

Embora os compartilhamentos sejam atrativos para usuários, eles têm custos extras para o provedor do serviço, dado sua arquitetura centralizada na nuvem. Isso porque arquivos em servidores na nuvem, ao serem compartilhados, são sincronizados para vários dispositivos, o que demanda maior capacidade de transferência dos dados da nuvem para os dispositivos, além de maior largura de banda para que essa transferência ocorra com rapidez. Adicionalmente, o tráfego dos dados compartilhados sobre a Internet contribui para o aumento de congestionamentos e custos dos provedores de Internet (ISPs), que já vem crescendo devido à popularização dos dispositivos móveis e conteúdo multimídia. Logo, é necessário abordagens para lidar com compartilhamentos, visando a redução do tráfego de dados sobre servidores na nuvem e ISPs, mas mantendo a qualidade do serviço para os usuários.

A comunicação direta de dispositivo para dispositivo (D2D), está dentre as principais abordagens da literatura para lidar com tráfego de dados gerado por compartilhamentos, como discutimos na Seção 2. Essa comunicação explora especialmente a proximidade entre dispositivos em uma mesma localidade (campus universitário, prédio empresarial ou habitacional) que se comunicam frequentemente por utilizarem uma mesma aplicação, inclusive para o compartilhamento de dados. D2D é um recurso importante para economia de custos com transmissão de dados, assim como redução de atrasos, sendo, portanto incorporado ao padrão 5G a ser adotado por dispositivos móveis mundialmente [Ansari et al. 2017]. Especificamente, D2D é definido, desde o padrão 3G, como a comunicação ad hoc entre dispositivos próximos dentro de uma estação-base (comunicação *inband*) e também a comunicação ad hoc, que usa o espectro não licenciado como as redes WiFi, Wifi-Direct ou Bluetooth (comunicação *outband*).

Nesse trabalho investigamos o potencial de uso da comunicação D2D, em especial a comunicação *outband*, para compartilhamentos em serviços de armazenamento pessoal na nuvem. O foco dessa investigação será analisar a probabilidade de ocorrência e a duração de encontros entre pares de dispositivos que compartilham dados. Para conduzir esse estudo, utilizamos uma base de dados do serviço de armazenamento na nuvem Dropbox [Gonçalves et al. 2016]. Esses dados mostram padrões de compartilhamentos entre usuários desse serviço em ambientes universitários e residenciais.

As contribuições desse artigo estão organizadas em duas análises que oferecem informações para cientistas e desenvolvedores avaliarem o desempenho de aplicações que venham a explorar a comunicação D2D em compartilhamento de dados. São elas:

- a medição da duração e quantidade de encontros entre contatos em serviços de armazenamento em nuvem;
- dois modelos estatísticos para representação e reprodução dessa medição, baseado em caracterização de dados extraídos de compartilhamentos reais.

Nossos resultados apresentados na Seção 3, mostram que encontros entre pares de contatos têm duração suficientemente longa para transmissões de dados via D2D: pelos

[1] https://www.dropbox.com/guide/business/share/collaborate, https://www.google.com/docs/about.

menos 74% dos encontros tem duração maior que 15 minutos. Adicionalmente, encontros entre os contatos ocorrem frequentemente, ou seja, em mais de 47% das requisições de dados em compartilhamentos há um ou mais contatos conectados no serviço de armazenamento, simultaneamente em uma mesma localidade que poderiam, potencialmente, transmitir dados via D2D. Baseados nessas medições com dados do Dropbox, modelamos a duração dos encontros como variáveis aleatórias seguindo a distribuição de Weibull e a quantidade de pares de contatos possíveis para comunicação D2D como uma variável aleatória seguindo a distribuição Geométrica. Em suma, nossas análises indicam que as oportunidades são amplas para aplicação de D2D a compartilhamentos de dados em serviços pessoais de armazenamento.Os modelos propostos podem ser utilizados para avaliar o uso de comunicação D2D em diferentes tipos de aplicações.

2. Trabalhos Relacionados

Este trabalho traz novas análises baseadas em outros trabalhos sobre compartilhamentos em serviços de armazenamento em nuvem. Em [Gonçalves et al. 2016], foi coletado o tráfego de rede associado ao Dropbox em redes universitárias e residenciais e proposto um modelo para geração de cargas sintéticas desse tipo de serviço, contudo não analisamos compartilhamentos e o potencial para comunicação D2D. Em [Gonçalves et al. 2017], foi proposto um modelo para analisar custos e benefícios de políticas de compartilhamento em serviços de armazenamento em nuvem e um mecanismo de incentivo para os usuários ajudarem a reduzir o tráfego de dados via D2D. Nesses trabalhos, contudo, não foram realizadas medições da quantidade e duração de encontros entre contatos que compartilham dados e seus respectivos modelos estatísticos.

A importância de compartilhamentos em serviços de armazenamento na nuvem já foi demonstrada em vários estudos. Medições sobre o Dropbox [Bocchi et al. 2015] e UbuntuOne [Gracia-Tinedo et al. 2015] mostraram que volumes de *downloads* superam *uploads* nesses serviços. Em [Gracia-Tinedo et al. 2016] foi mostrado evidências de adoção massiva de compartilhamento de dados e a predominância dos usuários que fazem apenas *download* de arquivos. Uma pesquisa conduzida em [Palviainen and Rezaei 2015] com usuários de diferentes serviços de armazenamento em nuvem identificou o compartilhamento de dados e a sincronização de dados em vários dispositivos como as principais razões para uso do serviço. Esses trabalhos fornecem evidências que motivam o estudo atual, mas eles não abordam a questão de oportunidades de comunicação D2D.

Compartilhamentos de dados também vem sendo estudado no âmbito de aplicações de comunicação entre pessoas. Em [Seufert et al. 2016], o aplicativo WhatsApp foi investigado quanto aos padrões de comunicação em grupos de pessoas e ao impacto no tráfego das redes. Os autores mostraram que toda comunicação no WhatsApp ocorre via um servidor localizado nos EUA (arquitetura totalmente centralizada), embora a natureza dessa comunicação seja D2D. Baseado em uma coleta de dados de voluntários, os autores mediram uma média de 9 participantes por grupo de comunicação, enquanto o volume das mensagens de conteúdo multimídia tem tamanho médio de 225KB. Em [Mota et al. 2017], foi proposto uma arquitetura chamada *D2D caching* onde dispositivos móveis funcionam como distribuidores de conteúdo via comunicação D2D em cenários onde pessoas com interesse comum em determinado conteúdo se encontram via redes wireless ou Bluetooth. Os autores avaliaram o desempenho dessa arquitetura

em vários cenários hipotéticos via simulações. Nosso trabalho segue a mesma linha de pesquisa desses, mas com o foco em serviços de armazenamento em nuvem.

3. Metodologia

Nessa seção, descrevemos a metodologia para as análises conduzidas nesse trabalho. Primeiramente descrevemos a base de dados utilizada, a seguir explicamos a simulação, medição e caracterização da comunicação D2D.

3.1. Base de Dados

Utilizamos uma base de dados de compartilhamentos de dados no Dropbox, que foi coletada no trabalho de [Gonçalves et al. 2016] e está publicamente disponível.[2] Esses dados foram coletados através do monitoramento passivo do tráfego de rede em dois campi universitários e dois pontos de presença (PoPs) de ISPs residenciais.

A base de dados consiste em traços de acessos de usuários a dados na nuvem via as redes desses campi e PoPs, onde cada dado é um *namespace* do Dropbox, isto é, uma estrutura utilizada nesse serviço para identificar de forma única um arquivo do usuário (documento, áudio, imagem) ou uma pasta. Os acessos são representados por identificadores anônimos (ID), que não oferecem dicas sobre a identidade dos usuários ou o conteúdo armazenado, mas permite analisar os padrões para compartilhamento de um *namespace* no Dropbox, que, por simplicidade, denominamos de agora a diante por *pasta* do usuário no Dropbox. Dentre as informações sobre cada acesso de usuário, utilizamos para as nossas análises: a marca de tempo do acesso, o ID da pasta, o ID do dispositivo do usuário que fez o acesso à pasta e o tipo de acesso, ou seja, *download* ou *upload* de dados. Além dos acessos, os traços também registram eventos de *login* e *logout* que permite extrair o período em que um dispositivo do usuário esteve conectado ao Dropbox, isto é, o tempo de sessão no serviço. A Tabela 1 sumariza as informações dos traços utilizados para as análises por local.

Tabela 1. Sumário das bases de dados de traços do Dropbox.

Local	Dispositivos	Pastas	Acessos (x1000)	Sessões (x1000)	Período
Campus-1	10516	12418	2560	3324	04/14-06/14
Campus-2	2216	4095	500	77	04/14-06/14
PoP-1	9347	12759	1645	1077	10/13-04/14
PoP-2	3336	4954	1293	3792	07/13-05/14

O compartilhamento de pastas entre dispositivos varia de acordo ao ambiente de trabalho do usuário. De forma geral, nas redes universitárias (Campus-1 e Campus-2) o compartilhamento é maior, onde pelo menos 33% das pastas tem mais de um dispositivo associado, enquanto nas redes residenciais (PoP-1 e PoP-2) essa porcentagem é ligeiramente inferior e alcança 24% das pastas. A média de dispositivos por pasta é de 2,72 e 2,60 nas redes universitárias para o Campus-1 e Campus-2 respectivamente, e nas redes residenciais essa média é de 2,31 e 2,44 no PoP-1 e PoP-2 respectivamente. Essas médias indicam que no ambiente universitário os serviços de armazenamento em nuvem são mais explorados para realização de trabalhos colaborativos ao passo que nas residências esse serviço é mais utilizado com o intuito de realizar *backup* de dados.

[2]Base de dados disponível em https://sites.google.com/a/ufpi.edu.br/traces.

3.2. Simulação da Comunicação D2D

Serviços de armazenamento em nuvem, tipicamente, não utilizam comunicação D2D. A arquitetura de comunicação centralizada é a padrão, onde o dispositivo que gera o dado a ser compartilhado, o envia primeiramente para servidores na nuvem, que depois o distribui aos demais dispositivos nos casos de compartilhamentos. Alguns serviços chegam a utilizar comunicação D2D, mas ainda de forma limitada. Por exemplo, o cliente Dropbox possui o módulo *LAN Sync*[3] para compartilhamento direto de dados apenas entre computadores pessoais (PCs) dentro de uma mesma rede local (LAN).

Para analisar o potencial da comunicação D2D, conduzimos simulações dirigida pelos traços de usuários do Dropbox para cada um dos locais da base de dados. Nesse sentido, assumimos a hipótese razoável que dois ou mais dispositivos com sessões do Dropbox abertas simultaneamente (*online*) sob o mesmo campus universitário ou PoP, ou seja, a mesma infraestrutura de rede, podem compartilhar dados via comunicação D2D.

As simulações dirigidas por traços foram realizadas com a seguinte metodologia. Primeiro, identificamos os conjuntos de dispositivos associados a cada pasta compartilhada, ou seja, grupos de contatos. A seguir, rastreamos os períodos em que cada contato está com sessões abertas e fechadas, ou seja, os períodos *online* e *offline*, para assim, calcular o tempo de duração dos encontros entre pares de contatos por pasta compartilhada. Adicionalmente, calculamos o número de contatos *online* no instante em que algum contato requisita um dado da pasta compartilhada. Todos os dados da simulação foram registrados para medições e análises.

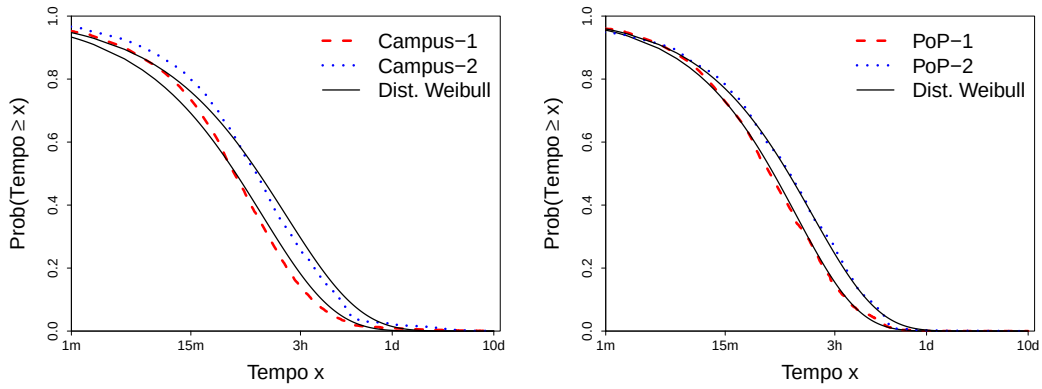
O modelo estatístico mais adequado para caracterizar os dados medidos nas simulações foram escolhidos entre vários modelos usados na literatura da seguinte forma. Para cada modelo, os parâmetros da distribuição que mais se aproximaram dos dados são optimizados usando o método de estimativa por máxima verossimilhança (MLE). Após definição dos parâmetros, a distribuição contínua com menor distância de Kolmogorov-Smirnov (KS) ou a distribuição discreta com o menor erro quadrático (LSE) em relação aos dados foi escolhida.[4] As implementações de MLE, KS e LSE do pacote estatístico R [Venables and Ripley 2002] foram utilizados para essa caracterização.

4. Resultados

Nesta seção apresentamos nossos resultados. Primeiramente, analisamos a duração dos encontros entre pares de contatos que compartilham a mesma pasta, ou seja, o tempo em que ambos estão *online* simultaneamente para, possivelmente, transferirem dados via D2D. Essa análise foca no percentual de pastas que são compartilhadas por mais de um dispositivo em cada local. Modelamos o tempo de duração dos encontros como uma variável aleatória x, então calculamos a probabilidade de se observar um encontro com o tempo x. A Figura 1 mostra distribuições de probabilidades do tempo de forma cumulativa, separados entre campi e PoPs para melhor visibilidade. As curvas tracejadas foram calculadas com as medições das simulações (distribuições empíricas), ao passo que as

[3]https://dropbox.tech/infrastructure/inside-lan-sync

[4]As distribuições analisadas foram: Uniforme, Normal, Log-normal, Exponencial, Gamma, Weibull, Pareto, Logística, Log-logística, Cauchy (contínuas) e Poisson, Binomial, Binomial Negativa, Geométrica, Hipergeométrica, e Zipf (discretas).



(a) (b)

Figura 1. Tempo de duração de encontro de pares de dispositivos que compartilham uma mesma pasta, caracterizado pela distribuição de Weibull[5] com parâmetros: $\kappa = 0.615$ e $\lambda = 4539.469$ (Campus-1); $\kappa = 0.598$ e $\lambda = 7776.43$ (Campus-2); $\kappa = 0.712$ e $\lambda = 4501.971$ (PoP-1); $\kappa = 0.675$ e $\lambda = 6517.757$ (PoP-2). Eixo x tem escala logarítmica para melhor visualização dos dados.

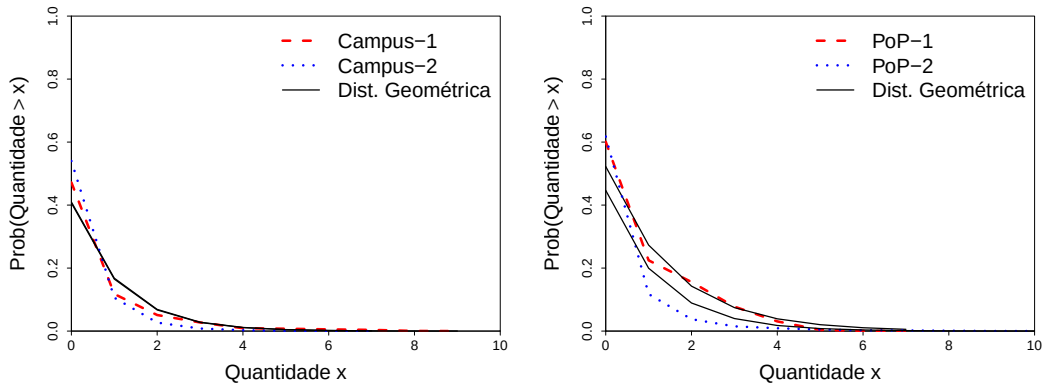
curvas contínuas representam distribuições de probabilidade da literatura que melhor se ajustam às medições (distribuições teóricas).

Podemos observar que as distribuições do tempo para universidades (campi) e residências (PoPs) têm o mesmo formato, ou seja, elas obtiveram o melhor ajuste para a mesma distribuição teórica, que é a distribuição de Weibull.[5] Modelagens com essas distribuições são vastas, e com relação à serviços de Internet, ela já foi utilizada para modelar o tempo de visualização de páginas web por um usuário [Liu et al. 2010]. Contudo, há diferenças nas parametrizações da distribuições de Weibull para cada local, mostradas na Figura 1, o que torna as curvas ligeiramente diferentes. Por exemplo, a probabilidade de um encontro entre um par de contatos durar 15 minutos ou mais está entre 74 - 80% em universidades (Campi-1 - Campi-2), ao passo que a probabilidade desse mesmo tempo ocorrer em residências está entre 74 - 79% (PoP-2 - PoP-1). É notável também que os ajustes entre distribuições empíricas e teóricas não são perfeitos, especialmente nos campi devido as diferentes rotinas de usuários (professores e estudantes) nesse ambiente. As variações entre as curvas das distribuições empíricas entre os campi e os PoPs, assim como seus ajustes às distribuições teóricas são esperadas para medições reais, e de modo geral, consideramos os padrões de comportamento dos usuários nesses dois ambientes similares para o uso de D2D em aplicações.

Para demonstrar o uso do modelo, analisamos um cenário crítico para a utilização de D2D em aplicações de comunicação e compartilhamento, que é a probabilidade de completar a transferência de um arquivo entre dois dispositivos [Mota et al. 2017]. Nesse caso, a duração de encontro entre um par de dispositivos requisitante-fornecedor do arquivo deve ser suficientemente longa para permitir a transferência. Formalmente, seja x a duração de um encontro, t o tamanho do arquivo e L a largura de banda da comunicação. A probabilidade de transmitir o arquivo é dada por:

$$Prob(transmissao) = Prob(x) \geq \frac{t}{L}.$$

[5] Função de densidade de probabilidade da distribuição de Weibull: $f(x) = \frac{\kappa}{\lambda}\frac{x}{\lambda}^{\kappa-1}e^{-(x/\lambda)^{\kappa}}$



(a) (b)

Figura 2. Quantidade de dispositivos no momento de uma requisição em uma pasta compartilhada, caracterizado pela distribuição Geométrica[6] com parâmetros: $p = 0.590$ (Campus-1); $p = 0.593$ (Campus-2); $p = 0.477$ (PoP-1); $p = 0.553$ (PoP-2).

Se considerarmos um cenário típico onde a maioria dos arquivos têm tamanho (t) menor que 10 Mbytes, que é o caso de 90% das transferências de dados no Dropbox [Gonçalves et al. 2016], e a largura de banda (L) é de 2 Mbps, que é o padrão das redes Wifi 802.11, temos 90% de chances de transferir dados com sucesso via D2D nas universidades e residências. Isso corresponde à probabilidade de um encontro entre um par de contatos durar mais que 42 segundos (t/L). Se considerarmos outro tipo de arquivo como vídeos do Youtube, onde 95% das transferências de vídeos são menores que 20 Mbytes [Mota et al. 2017], as chances de transferências com sucesso ainda são altas (88%), o que corresponde à probabilidade de um encontro durar mais que 84 segundos.

Agora, analisamos a quantidade de contatos *online* no momento em que algum contato requisita um dado em uma pasta compartilhada, que é uma questão igualmente importante à duração dos encontros. O uso de D2D é possível quando há pelo menos um contato *online*. Modelamos a quantidade de contatos como uma variável aleatória x, e calculamos probabilidade de se observar x contatos no momento de uma requisição. Nesse sentido, focamos apenas nas pastas associadas a mais de um dispositivo e suas respectivas requisições de dados, isto é, os acessos para *download* de dados da nuvem.

A Figura 2 mostra distribuições de probabilidades da quantidade de dispositivos de forma cumulativa, assim como na figura anterior. As distribuições para universidades e residências obtiveram o melhor ajuste para a mesma distribuição teórica, que é a distribuição Geométrica.[6] Assim, assumimos que o parâmetro p dessa distribuição é a probabilidade de um usuário requisitar um dado em uma pasta compartilhada no momento em que há x contatos *online*. Para nossas medições observa-se frequentemente algum contato ($x > 0$) no momento da requisição, isto é, em pelo menos 47% das requisições em universidades e 60% das requisições em residências seria possível a transferência via D2D. Em suma, nossas análises indicam que as oportunidades são vastas para aplicação de D2D nesses dois ambientes, tanto em termos de duração quanto ocorrência de encontros entre pares de contatos que compartilham dados.

[6] Distribuição de Probabilidade de Massa da distribuição Geométrica: $f(x) = (1-p)^x p$

5. Conclusões e Trabalhos Futuros

Nesse trabalho investigamos o potencial do uso da comunicação D2D em serviços de armazenamento em nuvem, visando a transmissão de dados diretamente entre contatos de um compartilhamento, e por consequência, reduzir custos e atrasos na transmissão. Nossas análises indicam que as oportunidades são vastas para o uso de D2D em ambientes universitários e residencias, tanto em termos da duração quanto ocorrência de encontros entre pares de contatos que compartilham dados. Nesse sentido, propomos modelos estatísticos ajustados com medições da probabilidade de encontros de contatos e suas durações em um sistema de armazenamento em nuvem real. Nossos modelos podem ser utilizados para avaliação de desempenho de aplicações quanto à comunicação D2D. Em trabalhos futuros pretendemos estender esses modelos e análises, considerando a integração desses em um único modelo e as demais variáveis que impactam no ajuste do modelo a dados reais de compartilhamentos.

Referências

Ansari, R. I. et al. (2017). 5g d2d networks: Techniques, challenges, and future prospects. *IEEE Systems Journal*, 12(4):3970–3984.

Bocchi, E., Drago, I., and Mellia, M. (2015). Personal Cloud Storage: Usage, Performance and Impact of Terminals. In *Proc. of the IEEE CloudNet*.

Cisco (2019). Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper. Disponível em `https://www.cisco.com`.

Gonçalves, G. et al. (2017). Cost-benefit tradeoffs of content sharing in personal cloud storage. In *Proc. of the IEEE MASCOTS*.

Gonçalves, G. D. et al. (2016). Workload models and performance evaluation of cloud storage services. *Computer Networks*, 109:183–199.

Gracia-Tinedo, R. et al. (2015). Dissecting ubuntuone: Autopsy of a global-scale personal cloud back-end. In *Proc. of the IMC*.

Gracia-Tinedo, R., García-López, P., Gómez, A., and Illana, A. (2016). Understanding data sharing in private personal clouds. In *Proc. of the IEEE I3C*.

Liu, C., White, R. W., and Dumais, S. (2010). Understanding web browsing behaviors through weibull analysis of dwell time. In *Proc. of ACM SIGIR*.

Mota, V. F. et al. (2017). Incentivando o compartilhamento de conteúdo via comunicação dispositivo-a-dispositivo. In *Proc. of COURB*.

Palviainen, J. and Rezaei, P. P. (2015). The next level of user experience of cloud storage services: Supporting collaboration with social features. In *Proc. of ASWEC*.

Seufert, M., Hoßfeld, T., Schwind, A., Burger, V., and Tran-Gia, P. (2016). Group-based communication in whatsapp. In *Proc. of IEEE IFIP*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, USA.