

Análise do Uso de Comunicação Dispositivo a Dispositivo para Transferência de Dados Compartilhados em Serviços de Armazenamento Pessoal em Nuvem

(Analysis of the Use of Device-to-Device Communication for Transferring Shared Data in Personal Cloud Storage Services)

Denilson da Silva Sousa
CSHNB - UFPI
Picos - Piauí - Brazil
denilsondsousa@ufpi.edu.br

Glauber Dias Gonçalves
CSHNB - UFPI
Picos - Piauí - Brazil
ggoncalves@ufpi.edu.br

RESUMO

Serviços de armazenamento pessoal em nuvem são utilizados frequentemente para compartilhar dados e realizar trabalhos colaborativos, o que os tornam atrativos para usuários. Contudo, compartilhamentos podem aumentar os custos dos provedores do serviço assim como custos de provedores de Internet residenciais com transmissões de dados entre vários dispositivos de usuários sob a Internet. Nesse artigo, investigamos o potencial de compartilhamento de dados entre contatos via o uso da comunicação direta de dispositivo para dispositivo (D2D) em função do comportamento dos usuários. Nesse sentido, medimos a ocorrência e a duração de encontros entre contatos, baseado em dados compartilhados via o Dropbox em redes locais de campus universitários e residências. Propomos modelos estatísticos para representação e reprodução dessas medições e analisamos a probabilidade de sucesso na transferência de dados compartilhados via D2D, considerando vários cenários com a variação nos parâmetros dos modelos. Nossos resultados mostram que as chances de sucesso na transferência de um arquivo compartilhado na nuvem entre dois contatos via D2D varia entre 15% a 72% do cenário mais pessimista para o mais otimista nos diferentes tipos de redes locais, indicando, portanto, oportunidades amplas para o uso comunicação D2D em serviços de armazenamento pessoal em nuvem.

Keywords

Armazenamento Pessoal em Nuvem; Dropbox; D2D; Caracterização; Modelos

RESUMO

Personal cloud storage services are often used to share data and perform collaborative work, which makes them attractive to users. However, sharing can increase costs for service providers as well as costs for residential Internet service

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

providers given the data transmissions between various user devices over the Internet. In this article, we investigate the potential for data sharing between contacts via the use of direct device-to-device (D2D) communication based on the user behavior. In this sense, we measure the occurrence and duration of meetings between contacts, based on the data they shared via Dropbox on local networks of university campuses and homes. We propose statistical models for the representation and reproduction of these measurements and we analyze the probability of success in the transfer of shared data via D2D, considering several scenarios with the variation in the parameters of the models. Our results show that the chances of success in transferring a file shared in the cloud between two contacts via D2D vary between 15% to 72% from the most pessimistic to the most optimistic scenario in the different types of local networks, thus indicating opportunities for the use of D2D communication in personal cloud storage services.

Keywords

Personal Cloud Storage; Dropbox; D2D; Characterization; Model

CCS Concepts

•Networks → Network performance evaluation;

1. INTRODUÇÃO

Armazenamento de dados pessoais na nuvem [11], 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 no relatório da Cisco de projeções para computação em nuvem, com a previsão de 1,3 ZB de dados armazenados na nuvem em 2022, um crescimento de 4,6 vezes comparado a 2016 [6].

Os serviços de armazenamento na nuvem com as funcionalidades acima mencionadas atraíram um grande interesse da indústria. O sucesso de público empurrou muitos prove-

dores para esse mercado, e entre os serviços mais conhecidos, pode-se citar Dropbox, Google Drive, OneDrive, e iCloud. Esses serviços também atraíram a atenção dos pesquisadores, não só devido à sua popularidade entre os usuários, mas também devido ao impacto sobre a Internet e as redes locais. Por exemplo, pesquisadores realizaram várias medições do uso desse tipo de serviço em redes locais. Em provedores de Internet residenciais europeus, as medições mostraram que os serviços Dropbox, Google Drive e OneDrive já estavam presentes em 36% [2], ao passo que em redes universitárias no Brasil, apenas o serviço Dropbox chegou a ocupar 4% do tráfego de rede diário de um campus universitário [8]. Esses estudos indicavam o crescimento desse tipo de serviço nos anos seguintes, o que foi comprovado posteriormente pelo relatório da Cisco [6] acima mencionado.

O compartilhamento de dados é um recurso essencial para os usuários de serviços de armazenamento em nuvem, como já demonstraram alguns estudos [19, 13]. 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 oferece a versão comercial de seu serviço focada no compartilhamento, ao passo que Google Drive e MS Onedrive incentivam 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. 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 [1]. 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 particular 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 e propor modelos que representem essas probabilidades baseados em dados reais. Adicionalmente, foram propostas simulações de vários cenários realistas para analisar as chances de sucesso da comunicação D2D, estendendo a versão preliminar desse trabalho apresentado em congresso [23]. Para conduzir esse estudo, utilizamos uma base de dados do serviço de armazenamento na nuvem Dropbox [11]. Esses dados mostram padrões de compartilhamentos entre usuários desse serviço em ambientes universitários e residenciais.

As contribuições desse artigo estão organizadas em três 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.
- análise da probabilidade de sucesso na transferência de dados compartilhados via D2D, considerando vários cenários com a variação nos parâmetros dos modelos.

Nossos resultados mostram que encontros entre pares de contatos têm duração suficientemente longa para transmissões de dados via D2D: pelos 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. A seguir, simulamos vários cenários com variações de parâmetros dessas distribuições. Nossos resultados mostram que nos ambientes universitários as chances de sucesso na transferência de um arquivo compartilhado na nuvem via D2D varia entre 15% a 66% do cenário mais pessimista para o mais otimista, enquanto que nos ambientes residenciais essas chances de sucesso variam entre 23% a 72% respectivamente.

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.

Esse artigo tem a seguinte organização de seções. Os trabalhos relacionados são apresentados na Seção 2. Na Seção 3, descrevemos a base de dados do Dropbox utilizada nesse trabalho e os respectivos processamentos para a medição de encontros entre contatos. Os resultados obtidos,

¹<https://www.dropbox.com/guide/business/share/collaborate>,
<https://www.google.com/docs/about>,
<https://www.microsoft.com/pt-br/microsoft-365/free-office-online-for-the-web>

especificamente os modelos, simulações e análises são apresentados na Seção 4. Finalmente, apresentamos as considerações finais sobre esse artigo na Seção 5.

2. TRABALHOS RELACIONADOS

Nesta seção discutimos de forma geral trabalhos relacionados a serviços que possibilitam contatos entre usuários para compartilhamentos de dados. Primeiramente apresentamos trabalhos sobre serviços de armazenamento pessoal em nuvem, a seguir, discutimos os esforços existentes para comunicação direta entre contatos nesses serviços para então discutir, sobre aplicações que tem como foco a comunicação entre pessoas e adicionalmente oferecem opções de compartilhamento de dados.

2.1 Armazenamento Pessoal em Nuvem

Este trabalho traz novas análises baseadas em outros trabalhos sobre compartilhamentos em serviços de armazenamento em nuvem [11, 9, 10].

Gonçalves et al. [11] coletaram o tráfego de rede associado ao Dropbox em redes universitárias e residenciais e propuseram um modelo para geração de cargas sintéticas desse tipo de serviço, contudo não foram analisados compartilhamentos e o potencial para comunicação D2D. Esses dados foram utilizados posteriormente em outros trabalhos dos autores com o foco foi avaliar o impacto de aplicações para compartilhamento de dados no tráfego de redes locais e da Internet.

Em [9], foram analisados o tráfego de rede associado aos compartilhadas do Dropbox em redes universitárias e residenciais. Os autores observaram que a fração desse tráfego relacionada a *downloads* é expressiva (68%) e cerca de 25% desses *downloads* são para atualizar compartilhamentos com dados já existentes nessas redes e, portanto, tais *downloads* poderiam ser evitados. Adicionalmente, Gonçalves et al. [10] propuseram 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 pessoal em nuvem já foi demonstrada em vários estudos. Medições sobre o Dropbox realizadas por Bocchi et al. [2] e no UbuntuOne por Gracia-Tinedo et al. [12] mostraram que volumes de *downloads* superam *uploads* nesses serviços. Adicionalmente, Gracia-Tinedo et al. [13] mostraram evidências de adoção massiva de compartilhamento de dados e a predominância dos usuários que fazem apenas *download* de arquivos. Uma pesquisa conduzida por Palviainen et al. [19] 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.

Em vários trabalhos foram analisados o desempenho e caracterizados o comportamento de usuários em serviços pessoais de armazenamento na nuvem como Dropbox, OneDrive, iCloud e Google Drive [7, 3, 15]. Os autores desses trabalhos sugerem várias otimizações para um serviço de armazenamento quanto à redução de custo com transferên-

cias e sincronizações de dados entre a nuvem e dispositivos de usuários mas com o foco em funcionalidades da aplicação cliente como compressão, deduplicação e particionamento de dados. Novamente, em nenhum desses trabalhos o comportamento dos usuários foram analisados como uma oportunidade de uso de D2D explorando encontros entre contatos com compartilhamentos em comum.

2.2 Comunicação Direta entre Contatos

Análise sobre o potencial de comunicação direta entre contatos em serviços de armazenamento pessoal em nuvem não é novo. Existem trabalhos na literatura onde autores propuseram alternativas ao modelo de comunicação atual de transferência de dados desses serviços que é centralizado em servidores na nuvem, especialmente, explorando protocolos de comunicação par-a-par (P2P) como Bittorrent [20].

O Wuala era um serviço de armazenamento pessoal em nuvem baseado em protocolos P2P (encerrado em 2015). Mager et al. [17] analisou o Wuala, apresentando detalhes da arquitetura e desempenho desse sistema. Quando o Wuala foi lançado, sua arquitetura se baseava na combinação inteligente de armazenamento em nuvem para backup de longo prazo com um cache de arquivos baixados com frequência servidos por participantes da rede via comunicação P2P. Contudo, esse serviço migrou para o modelo centralizado na nuvem, e a baixa disponibilidade de participantes, i.e., encontros entre contatos, foi um dentre os motivadores para essa migração.

Gracia et al. [14] propuseram uma rede social entre contatos onde os dados, e.g., postagens e conteúdo multimídia, são armazenados de forma distribuída nos dispositivos dos usuários e em serviços de armazenamento em nuvem, ou seja um sistema de armazenamento híbrido onde os dispositivos se comunicam via P2P. Assim os autores estudam um modelo para a construção de nuvens sociais altamente disponíveis. Cada membro da nuvem social contribui com armazenamento local e em nuvem para armazenar partes de seus dados e mascarar os períodos recorrentes de indisponibilidade de contatos. Os autores observam que tal sistema de comunicação direta entre dispositivos depende, crucialmente, da coincidência de encontros entre dispositivos e quando essa coincidência tem probabilidade baixa a necessidade de armazenamento em nuvem aumenta.

Chaabouni et al. [5] propõem o uso do protocolo Bittorrent, ou seja, um exame de contatos para baixar conteúdo compartilhado em serviços de armazenamento pessoal em nuvem. Esse protocolo é gerenciado por um algoritmo que troca um *download* do servidor nuvem para o Bittorrent quando há contatos disponíveis. Nesse caso, o *download* é executado apenas se o tempo de download for reduzido. Contudo, as chances de criar um exame de Bittorrent para baixar um arquivo dependem do número de dispositivos disponíveis simultaneamente para fornecê-lo.

As questões observadas nesses trabalhos [17, 14, 5] quanto à disponibilidade de dispositivos que podem fornecer um dado de interesse para algum usuário, nos motivam a investigar as chances de encontro entre contatos e duração dos encontros em compartilhamentos Dropbox para o uso potencial da comunicação D2D em redes locais.

2.3 Aplicações de Comunicação

Compartilhamentos de dados também vem sendo estudado no âmbito de aplicações de comunicação entre contatos

que são pessoas com laços familiares ou profissionais.

O aplicativo WhatsApp foi investigado por Seufert et al. [22] 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.

Vieira et al. [25] analisam também o aplicativo WhatsApp quanto à sua arquitetura de comunicação e o potencial para difusão de informação em massa entre dispositivos via de grupos de contatos. Em suas medições, os autores observaram que 50% dos grupos tinham 20 usuários, com uma média de 6 usuários por grupos, alcançando um grupo de até 211 membros. A difusão de dados em massa ocorre devido à estrutura da rede entre grupos de contatos, amplificada por funções de *forward* e *broadcast*, que permitem transmissões de dados para até 256 contatos simultaneamente.

Mota et al. [18] propuseram 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 trabalhos, em particular [22, 18], 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 [11] 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

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

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 3.1 sumariza as informações dos traços utilizados para as análises por local. 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 essas pastas estão associados 29%, 41%, 21% e 22% dos dispositivos no Campus-1, Campus-2, PoP-1 e PoP-2 respectivamente. Contudo, quando se observa os casos extremos, ou seja, pastas com muitos dispositivos associados, nas residências há pastas compartilhadas com mais de dez dispositivos, alcançando até dezesseis dispositivos no PoP-1. O maior compartilhamento foi observado no Campus-1, contento 24 dispositivos, o que provavelmente esta relacionado a grupos de estudantes ou professores compartilhando conteúdos. 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 pessoal 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.

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).

O mecanismo de compartilhamento varia de serviço para serviço, mas geralmente os dispositivos que têm pastas compartilhadas sincronizam o conteúdo imediatamente se estiverem online, ou assim que ficam on-line, embora os usuários podem pausar/parar manualmente sincronização automática. Portanto, esse mecanismo levam a vários downloads de um único conteúdo de servidores em nuvem pelos vários dispositivos que participam de compartilhamentos. A comunicação D2D *outband* [1] entre os dispositivos que estão numa localidade, por exemplo, possivelmente sob a mesma rede Wifi de um campus universitário ou o mesmo PoP de um ISP, poderia reduzir esses downloads. No entanto, esses dispositivos precisam estar online *simultaneamente* para a comunicação D2D como analisamos a próximo seção.

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

³<https://dropbox.tech/infrastructure/inside-lan-sync>

Table 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

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 conforme a metodologia proposta por Gonçalves et al [10]. 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, ou seja, houve modificações na pasta para ser compartilhada ou um dispositivo entra *online* e requisita as modificações na pasta enquanto estava *offline*. 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. As distribuições analisadas foram Uniforme, Normal, Log-normal, Exponencial, Gamma, Weibull, Pareto, Logística, Log-logística, Cauchy para distribuições contínuas e Poisson, Binomial, Binomial Negativa, Geométrica, Hipergeométrica, e Zipf para distribuições discretas.

Para cada modelo, os parâmetros da distribuição que mais se aproximaram dos dados são otimizados 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. As implementações de MLE, KS e LSE do pacote estatístico R [24] foram utilizados para essa caracterização.

4. RESULTADOS

Nesta seção apresentamos nossos resultados. Primeiro propomos modelos estatísticos para as medições sobre compartilhamentos entre contatos no Dropbox. A seguir, aplicamos esses modelos em análises de vários cenários hipotéticos de uso da comunicação D2D em serviços de armazenamento em nuvem.

4.1 Modelos Estatísticos

Analisamos a quantidade de contatos *online* no momento em que algum contato requisita um dado, i.e., *download*, em uma pasta compartilhada. O uso de D2D é possível quando há pelo menos um contato *online*. Modelamos esse número número de contatos como uma variável aleatória x , e calculamos a probabilidade de se observar x contatos no momento de uma requisição. Nesse sentido, focamos apenas

nas requisições de dados a pastas associadas a mais de um dispositivo, isto é, as pastas compartilhadas.

A Figura 1 mostra distribuições de probabilidades da quantidade de dispositivos 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 curvas contínuas representam distribuições de probabilidade da literatura que melhor se ajustam às medições (distribuições teóricas).

As distribuições para universidades e residências obtiveram o melhor ajuste para a mesma distribuição teórica, que é a distribuição Geométrica.⁴ Para isso, assumimos que o parâmetro p dessa distribuição é a probabilidade de ausência de dispositivos para a comunicação D2D no momento da requisição, ou seja, a ocorrência de uma falha. Em nossas medições observa-se chances consideráveis de haver 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. O parâmetro p estimado para a distribuição Geométrica aproxima-se dessas medições como mostra a Figura 1. Portanto, nossas medições indicam que as oportunidades são relevantes para aplicação de D2D nesses dois ambientes quanto à ocorrência de encontros entre pares de contatos que compartilham dados.

Agora, 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. Novamente, essa análise foca nas medições 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 , e assumimos que esse tempo é independente para cada par de dispositivo *online* simultaneamente. A seguir, calculamos a probabilidade de se observar um encontro com o tempo x . A Figura 2 mostra distribuições de probabilidades do tempo de forma cumulativa, separados entre campi e PoPs para melhor visibilidade, assim como na figura anterior.

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 Essa distribuição é usualmente utilizada para modelar o tempo até que uma falha ocorra em um processo, o que pode ser aplicável a serviços de Internet. Por exemplo, a distribuição Weibull foi utilizada por Liu et al. [16] para modelar o tempo de visualização de páginas web por um usuário. Em nosso modelo, assim como nas propostas da literatura, o parâmetro λ representa a escala de tempo observada nas medições ao passo que o parâmetro κ é adimensional e repre-

⁴ Distribuição de Probabilidade de Massa da distribuição Geométrica: $f(x) = (1-p)^x p$

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

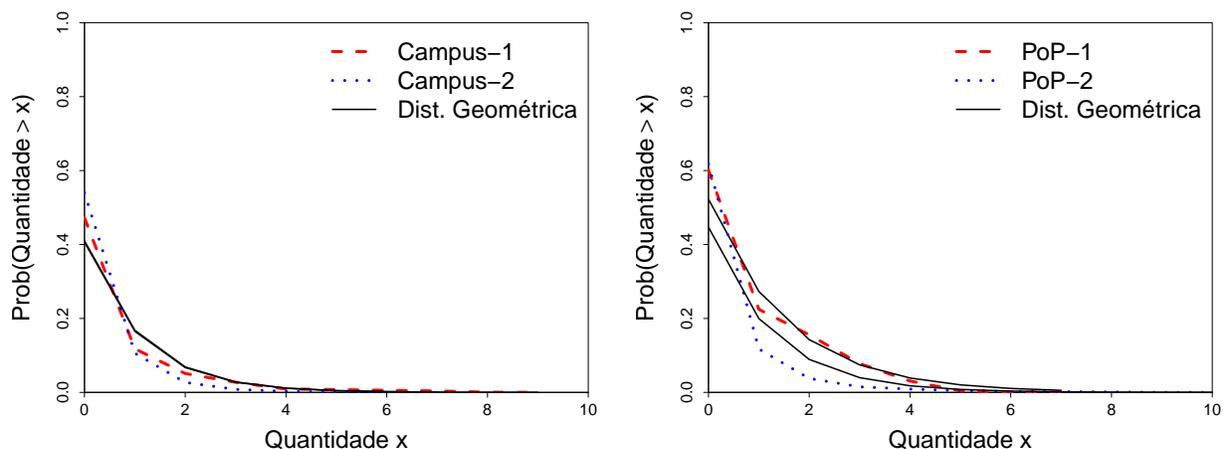


Figure 1: Quantidade de dispositivos no momento de uma requisição em uma pasta compartilhada, caracterizado pela distribuição Geométrica⁴ com parâmetros: $p = 0.590$ (Campus-1); $p = 0.593$ (Campus-2); $p = 0.477$ (PoP-1); $p = 0.553$ (PoP-2).

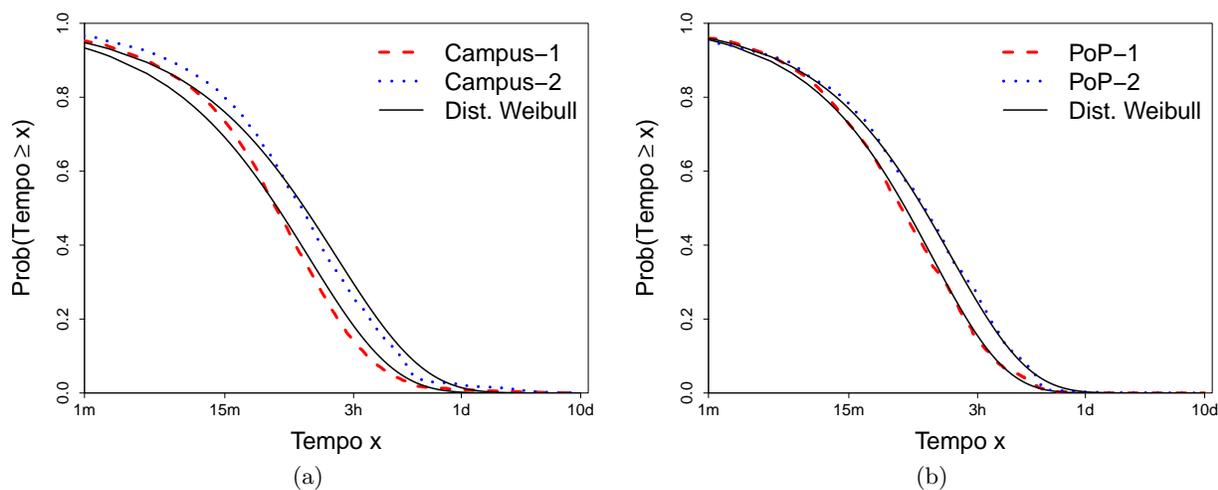


Figure 2: Tempo de duração de encontro de pares de dispositivos que compartilham uma mesma pasta, caracterizado pela distribuição de Weibull⁵ com parâmetros: $\kappa = 0.615$ e $\lambda = 4539.469$ (Campus-1); $\kappa = 0.598$ e $\lambda = 7776.43$ (Campus-2); $\kappa = 0.712$ e $\lambda = 4501.971$ (PoP-1); $\kappa = 0.675$ e $\lambda = 6517.757$ (PoP-2). Eixo x tem escala logarítmica para melhor visualização dos dados.

senta o formato da curva da distribuição. O valor de $\kappa < 1$ indica que a maioria das falhas ocorre em períodos curtos de tempo, mas à medida em que se aumenta o período observado para um processo é menos provável que ocorra falhas instantâneas, fenômeno denominado *envelhecimento negativo*. Por outro lado, o valor de $\kappa = 1$ indica que a ocorrência de falhas é independente do tempo (distribuição Exponencial), ao passo que $\kappa > 1$ indica que as chances de falhas aumentam com o tempo de observação [21].

No modelo que propomos, mostrado na Figura 2, há diferenças nas parametrizações das distribuições de Weibull para cada local, 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). Quanto ao padrão das curvas, nota-se que o parâmetro $\kappa < 1$ para as distribuições Weibull ajustadas evidenciam que as durações de encontros seguem um envelhecimento negativo, isto é, a maioria das durações são curtas, contudo à medida que elas continuam tendem a se perpetuarem por horas ou até dias.

É importante observar que para ambos os modelos mostrados nas Figuras 1 e 2 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 com potencial para o uso de D2D em aplicações, como analisamos na próxima seção.

4.2 Cenários de Comunicação D2D

Para demonstrar o uso dos modelos, analisamos uma questão crítica para a utilização de comunicação D2D em compartilhamentos de dados na nuvem: a chance de sucesso em uma requisição de arquivo. Em outras palavras, analisamos a probabilidade de haver encontros entre contatos de um mesmo compartilhamento no momento em que um arquivo é requisitado no compartilhamento, e que a duração do encontro seja suficiente para completar a transferência do arquivo entre dois contatos via D2D.

Nesse sentido, calculamos as chances de sucesso de comunicação D2D em diferentes cenários onde consideramos variações hipotéticas, mas realistas, nos parâmetros dos dois modelos estatísticos propostos para o comportamento dos contatos em compartilhamentos na nuvem. Primeiro, propomos três configurações para o número de contatos online simultaneamente no momento de uma requisição, variando o parâmetro p da distribuição geométrica, i.e., a probabilidade de falha de comunicação. Obtivemos dessa forma as configurações para o número de contatos *online* Média, Pequena e Grande, onde a distribuição desse número observada no Dropbox representa a configuração Média, ao passo que, essa distribuição três vezes menor (em média) representa a configuração Pequena e três vezes maior (em média) representa a configuração Grande.

A seguir, propomos três configurações de tempo de duração entre pares de contatos variando o parâmetro λ da distribuição Weibull (a escala), mas fixando o parâmetro κ para manter o padrão de envelhecimento negativo caracte-

rizado para os dados Dropbox. Obtivemos dessa forma as configurações para o duração de encontros Média, Curta e Longa, onde a distribuição desse número observada no Dropbox representa a configuração Média, ao passo que, essa distribuição com média dez vezes menor (i.e., $\lambda/10$) representa a configuração Curta, e a distribuição com média dez vezes maior (i.e., $\lambda * 10$) representa a configuração Longa.

As variações mencionadas para número de contatos e duração de encontros com os respectivos valores para os parâmetros são exemplificadas através de distribuições cumulativas na Figura 3, apenas para o Campus-1 e PoP-1, dado que outros locais apresentam curvas com formato similares. Nota-se que as curvas das configurações médias representam o comportamento dos contatos conforme as caracterizações das medições no Dropbox, discutidos na Seção 4.1.

Dado as três configurações para cada modelo estatístico, propomos nove cenários para a nossa análise, combinando uma configuração de número de contatos com cada configuração de duração de encontros. Assumimos então que o número de contatos *online* simultaneamente no momento da requisição em um compartilhamento é um evento independente da duração dos encontros entre os contatos. Dessa forma, cada cenário consiste em uma simulação de várias requisições de dados via comunicação D2D, onde analisamos a porcentagem de sucesso considerando uma amostragem de 1000 requisições.

Em mais detalhes, uma requisição obtém sucesso quando pode ser servida por ao menos um dentre n dispositivos *online* aleatoriamente esboçados da distribuição geométrica. No caso de $n > 0$, é esboçado também o valor x_i aleatório da distribuição de Weibull para representar a duração do encontro entre o dispositivo requisitante e o i -ésimo dispositivo fornecedor dentre n possíveis fornecedores do arquivo. A duração do encontro deve ser suficientemente longa para a transferência do arquivo em *ao menos um par de dispositivos requisitante-fornecedor*. Formalmente, seja x_i a duração de um encontro, t o tamanho do arquivo e L a largura de banda da comunicação, a quantidade de dispositivos que podem atender a requisição é expressa por:

$$\sum_{i=1}^n \mathbb{1}[x_i \geq \frac{t}{L}],$$

onde $\mathbb{1}[x_i \geq \frac{t}{L}]$ é uma função indicadora que soma valor 1 quando a duração do encontro é maior que o tempo de transferência do arquivo (t/L). Portanto, uma requisição obtém sucesso se essa equação é maior ou igual a 1, caso contrário, ocorre uma falha de transmissão via comunicação D2D. Consideramos para os nove cenários o tamanho do arquivo requisitado (t) fixo em 10MB e largura de banda (L) de 2 Mbps, que são, respectivamente, o tamanho usual de arquivos multimídia em redes de compartilhamentos e a velocidade padrão das redes Wifi 802.11 [18].

A Figura 4 mostra as chances de sucesso nos nove cenários simulados, onde cada cenário é representado por uma barra: a cor da barra indica a configuração para número de contatos enquanto a posição da barra no eixo das abcissas indica a configuração para duração dos encontros. Por sua vez, o eixo das ordenadas indica a chance de sucesso, ou seja, a porcentagem de sucesso para cada cenário. A Figura 4(a) mostra as chances de sucesso para os campi e a Figura 4(b) mostra as chances de sucesso para as residências.

Primeiramente, observamos que para todos os cenários as

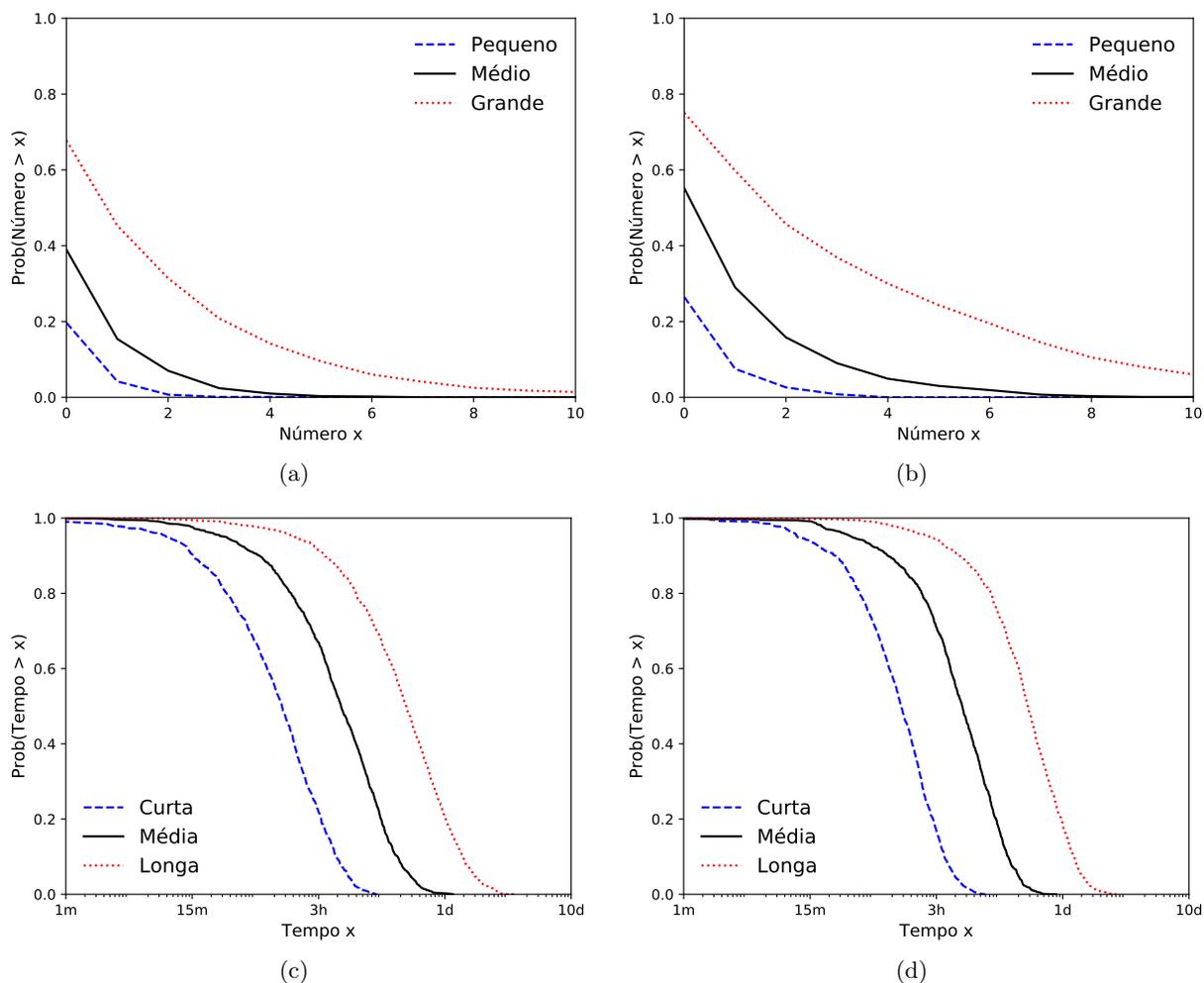


Figure 3: Variação dos modelos estatísticos para análise de chances de sucesso na comunicação D2D, exemplificados para os locais Campus-1 (primeira coluna) e PoP-1 (segunda coluna): (a-b) configurações de número *Pequeno*, *Médio* e *Grande* para a distribuição de contatos *online* no momento da requisição; (c-d) configurações de tempos *Curto*, *Médio* e *Longo* para a distribuição da duração de encontros entre dispositivos.

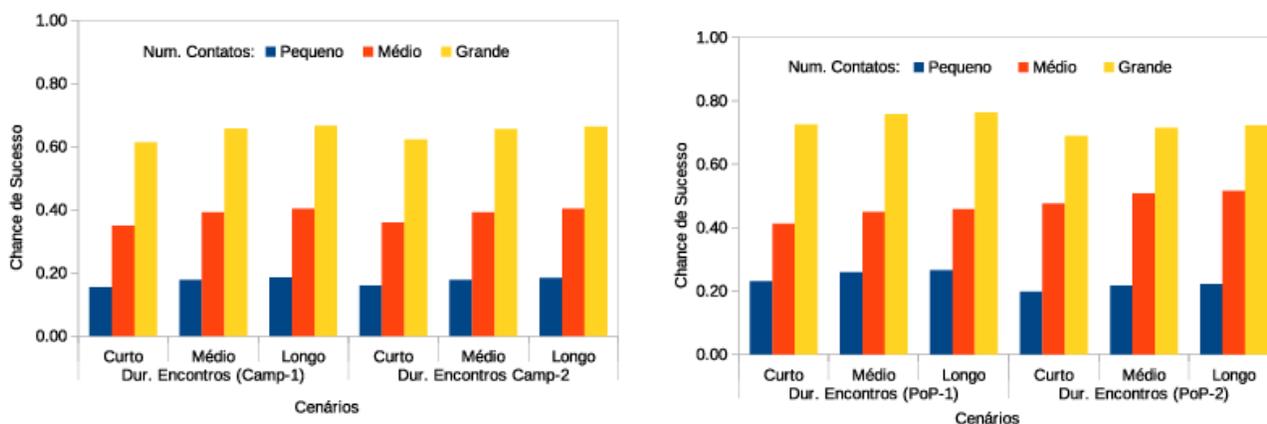


Figure 4: Chances de sucesso para transmissão de um arquivo entre dois contatos que compartilham dados em serviços de armazenamento em nuvem considerando nove cenários distintos em cada local para números de contatos *online* e duração de encontros entre os contatos.

chances de sucesso na requisição de um arquivo são maiores no ambiente residencial. Tomando como referência os cenários médios nos quatro locais, que representam as medições no Dropbox, as chances de sucesso de uma requisição via D2D no ambiente universitário é de 39-40%, ao passo que no ambiente universitário essa porcentagem varia de 45-51%. Isso se explica pela probabilidade de falha (parâmetro p da distribuição geométrica) ser menor nas residências (Figura 1), mesmo embora o tempo de duração dos encontros entre pares de contatos nessas redes seja ligeiramente menor que os ambientes universitários (Figura 2).

Dado essa primeira análise, observamos a seguir o impacto da variação de número de dispositivos em comparação à variação da duração dos encontros na chance de sucesso. É notável que a primeira variação tem maior impacto que a segunda. Considere, por exemplo, o Campus-1, onde os cenários com número de dispositivos Pequeno, Médio e Grande e duração de contatos Curta obtêm chances de sucesso 15%, 34% e 61%, ao variarmos a duração de contatos para Longa, as chances de sucesso aumentam levemente para 18%, 40% e 66% respectivamente. Esse padrão é qualitativamente similar para os outros locais como pode ser observado na Figura 4.

Portanto, as variações de duração dos encontros tem menor impacto e mantém a chance de sucesso das requisições em encontros curtos quase similar aos encontros longos. Isso ocorre porque há múltiplas tentativas para uma requisição se temos $n > 1$ contatos para uma requisição de arquivo. Esse resultado mostra a importância em desenvolver protocolos de comunicação D2D que explorem a possibilidade de redundância em número de dispositivos *online* simultaneamente para reduzir a aleatoriedade de duração de encontros. Adicionalmente, protocolos clássicos de comunicação par a par como Bittorrent podem ser utilizados para aumentar a velocidade de transferências quando há vários dispositivos fornecedores de uma requisição, como proposto por [4].

Contudo, observamos que a duração dos encontros têm um impacto relevante para o sucesso de requisições de arquivos via D2D nos casos em que contatos de um compartilhamento raramente se encontram, como exemplificado em cenários com número de dispositivos *pequeno* (barras de cor azul). Por exemplo, os cenários com número de contatos pe-

queno no PoP-1, a duração de contatos curto e longo obtêm chances de sucesso 23% e 26%, ou seja, houve um aumento de 15% de chances de sucesso. Nos cenários onde contatos se encontram frequentemente, o tempo de duração desses encontros não é um fator tão relevante para o sucesso da transmissão D2D, dado que observamos um aumento de 5% quando esse tempo transita de curto para longo. É importante observar nesse caso, que múltiplas tentativas de contatos para uma requisição deve ser explorada para reduzir o impacto da duração dos encontros.

Enfim, destacamos as chances de sucesso de transmissão de um arquivo entre o cenário mais pessimista e otimista de nossas simulações. Nos ambientes universitários observamos que essas chances variam entre 15% a 66%, enquanto que nos ambientes residências as chances de sucesso variam entre 23% a 72%. Portanto, há um potencial relevante para o uso de comunicação D2D em compartilhamentos de serviços de armazenamento em nuvem, mesmo no cenário pessimista, se exploradas as redes Wifi locais, o número de contatos *simultâneos* e a duração dos encontros entre esses contatos. O desenvolvimento de protocolos D2D nesses cenários podem reduzir os custos com largura de banda dos servidores de nuvem, assim como reduzir o tráfego de provedores de Internet.

5. CONCLUSÕES

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 em uma mesma rede local, por exemplos, um campus universitário ou um PoP de ISPs residenciais. Por consequência, o sucesso no em transmissão de dados via D2D pode reduzir custos dos provedores de serviço assim como ISPs, assim como diminuir atrasos na transmissão.

Nossas análises indicam que as oportunidades são amplas para o uso de D2D em ambientes universitários e residências, 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. Adicionalmente demonstramos o uso desses modelos em análises da probabilidade de sucesso na transferência de dados compartilhados via D2D, considerando vários cenários realistas, onde as chances de sucesso podem variar entre 15% a 72% em função de características das redes locais.

Em trabalhos futuros pretendemos estender esses modelos e análises, considerando outras variáveis extraídas do comportamento dos contatos que podem impactar no ajuste do modelo a dados reais de compartilhamentos. Adicionalmente, pretendemos analisar o desempenho de aplicações D2D em dispositivos móveis em termos de consumo de bateria, dado o uso de recursos de processamento e rede.

6. REFERENCES

- [1] R. I. Ansari et al. 5g d2d networks: Techniques, challenges, and future prospects. *IEEE Systems Journal*, 12(4):3970–3984, 2017.
- [2] E. Bocchi, I. Drago, and M. Mellia. Personal Cloud Storage: Usage, Performance and Impact of Terminals. In *Proc. of the IEEE CloudNet*, 2015.
- [3] E. Bocchi, I. Drago, and M. Mellia. Personal Cloud Storage: Usage, Performance and Impact of Terminals. In *Proc. of the IEEE CloudNet*, 2015.
- [4] R. Chaabouni, M. Sánchez-Artigas, P. García-López, and L. Pàmies-Juárez. The power of swarming in personal clouds under bandwidth budget. *Journal of Network and Computer Applications*, 65:48–71, 2016.
- [5] R. Chaabouni, M. Sánchez-Artigas, P. García-López, and L. Pàmies-Juárez. The power of swarming in personal clouds under bandwidth budget. *Journal of Network and Computer Applications*, 65:48 – 71, 2016.
- [6] Cisco. Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper, 2019. Disponível em <https://www.cisco.com>.
- [7] I. Drago et al. Inside Dropbox: Understanding Personal Cloud Storage Services. In *Proc. of the 12th ACM Internet Measurement Conference*, 2012.
- [8] R. Duarte, A. B. Vieira, I. Cunha, and J. M. Almeida. Impact of provider failures on the traffic at a university campus. In *Proc. of IFIP Networking Conference*, pages 1–9, May 2015.
- [9] G. Gonçalves et al. The impact of content sharing on cloud storage bandwidth consumption. *IEEE Internet Computing*, 20(4):26–35, 2016.
- [10] G. Gonçalves et al. Cost-benefit tradeoffs of content sharing in personal cloud storage. In *Proc. of the IEEE MASCOTS*, 2017.
- [11] G. D. Gonçalves et al. Workload models and performance evaluation of cloud storage services. *Computer Networks*, 109:183–199, 2016.
- [12] R. Gracia-Tinedo et al. Dissecting ubuntuone: Autopsy of a global-scale personal cloud back-end. In *Proc. of the IMC*, 2015.
- [13] R. Gracia-Tinedo, P. García-López, A. Gómez, and A. Illana. Understanding data sharing in private personal clouds. In *Proc. of the IEEE I3C*, 2016.
- [14] R. Gracia-Tinedo, M. Sánchez-Artigas, A. Ramírez, A. Moreno-Martínez, X. León, and P. García-López. Giving form to social cloud storage through experimentation: Issues and insights. *Future Generation Computer Systems*, 40:1 – 16, 2014.
- [15] Z. Li, Y. Zhang, Y. Liu, T. Xu, E. Zhai, Y. Liu, X. Ma, and Z. Li. A quantitative and comparative study of network-level efficiency for cloud storage services. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 4(1):1–32, 2019.
- [16] C. Liu, R. W. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proc. of ACM SIGIR*, 2010.
- [17] T. Mager, E. Biersack, and P. Michiardi. A measurement study of the wuala on-line storage service. In *12th International Conference on Peer-to-Peer Computing*, pages 237–248, 2012.
- [18] V. F. Mota et al. Incentivando o compartilhamento de conteúdo via comunicação dispositivo-a-dispositivo. In *Proc. of COURB*, 2017.
- [19] J. Palviainen and P. P. Rezaei. The next level of user experience of cloud storage services: Supporting collaboration with social features. In *Proc. of ASWEC*, 2015.
- [20] D. Qiu and R. Srikant. Modeling and performance analysis of bittorrent-like peer-to-peer networks. *ACM SIGCOMM computer communication review*, 34(4):367–378, 2004.
- [21] H. Rinne. *The Weibull distribution: a handbook*. CRC press, 2008.
- [22] M. Seufert, T. Hofffeld, A. Schwind, V. Burger, and P. Tran-Gia. Group-based communication in whatsapp. In *Proc. of IEEE IFIP*, 2016.
- [23] D. Sousa and G. Gonçalves. Um estudo sobre compartilhamentos entre contatos via d2d em serviços de armazenamento pessoal em nuvem. In *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí*, pages 244–251. SBC, 2020.
- [24] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, USA, 2002.
- [25] C. C. Vieira, P. F. Melo, P. O. S. V. de Melo, and F. Benevenuto. O paradoxo da viralização de informação criptografada no whatsapp. In *Anais Principais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 403–416. SBC, 2019.