# Trabalho Colaborativo em Serviços de Armazenamento na Nuvem: Uma Análise do Dropbox*

**Glauber Dias Gonçalves[1], Alex Borges Vieira[2], Ana Paula Couto da Silva[1], Jussara M. Almeida[1]**

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

[2]Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora

{ggoncalves, ana.coutosilva, jussara}@dcc.ufmg.br

alex.borges@ufjf.edu.br

***Resumo.*** *Provedores de serviços de armazenamento na nuvem adotam várias medidas para incentivar a atividade de seus usuários bem como atrair novos usuários. Uma das ações adotadas pelos principais provedores com esse propósito é oferecer a opção de realizar trabalho colaborativo, via compartilhamento de dados entre usuários em rede. Neste estudo, nós investigamos a relação entre trabalho colaborativo e o nível de atividade dos usuários do Dropbox, um dos serviços mais populares atualmente. Nós medimos o volume do tráfego que os usuários geram à medida que armazenam dados na nuvem e realizam compartilhamentos, utilizando dados coletados de quatro redes distintas. Nossos resultados indicam que usuários engajados em trabalhos colaborativos sincronizam maior volume de dados com a nuvem. Logo, esses usuários tem maior atividade e potencialmente pagam ou podem vir a pagar pelo serviço, sugerindo que, apesar dos custos, a funcionalidade de trabalho colaborativo pode aumentar a receita dos provedores de armazenamento na nuvem.*

***Abstract.*** *Cloud storage service providers have adopted several measures to incentivize the activity of their users as well as attract new ones. One of actions adopted by the major providers for this purpose is to provide the option to carry out collaborative work via data sharing between networked users. In this study, we investigate the relationship between collaborative work and the activity level of the users of Dropbox, one of the most popular services currently. We measure the traffic volume that users generate as they store data in the cloud and make shared folders using data collected from four different networks. Our results indicate that users participating in collaborative work synchronize more data to the cloud. Therefore, these users have higher activity and potentially can pay for the service, suggesting that, despite its cost, collaborative work feature can increase revenue of cloud storage providers.*

## 1. Introdução

O uso ubíquo e transparente de serviços de armazenamento em nuvem foi determinante para impulsionar sua popularização. É notório o crescente uso de aplicações desse tipo,

*Esta pesquisa é financiada pelo projeto FAPEMIG-PRONEX-MASWeb – Modelos, Algoritmos e Sistemas para a Web (processo APQ-01400-14), pelo INCTWeb (MCT/CNPq 573871/2008-6) e CNPq.

desde usuários domésticos a grandes empresas. O mercado é dominado por grandes empresas como Dropbox, Microsoft, Amazon e Apple que atualmente movimentam, apenas nos EUA, valores acima de $12 bilhões de dólares [Gracia-Tinedo et al. 2015].

Em tais serviços, o compartilhamento de dados – crucial para trabalho colaborativo – é uma das funcionalidades oferecidas. Recentemente, tem se observado um empenho dos maiores provedores de armazenamento na nuvem em aumentar colaborações entre usuários em seus serviços. Por exemplo, Dropbox recentemente lançou a versão empresarial de seu serviço com foco em compartilhamentos, enquanto Google Drive busca incentivar o trabalho colaborativo via edição de texto, planilhas eletrônicas e apresentações diretamente na nuvem em tempo real. De fato, uma pesquisa realizada em ambiente coorporativo [Shami et al. 2011] mostra que compartilhamento de dados permite o aumento de coleções de recursos digitais disponíveis em empresas, como consequência, a utilização de sistemas de arquivos torna-se maior.

Em tese, quanto maior a atividade de um usuário, mais dados ele irá armazenar e, consequentemente, maior será a receita dos serviços de armazenamento na nuvem. Geralmente, os provedores desses serviços incentivam a atividade de seus usuários por ações de marketing ou por promoções temporárias. Como alternativa a essas medidas, trabalho colaborativo via compartilhamento de arquivos pode atrair novos usuários e até mesmo incentivar a atividade dos já existentes. Intuitivamente, ao se compartilhar dados no serviço de armazenamento, todos os envolvidos são onerados pelo espaço em disco compartilhado (a custo de um único recurso na nuvem). Mais ainda, novos usuários podem se juntar ao serviço para participar de um projeto ou acessar um arquivo compartilhado.

Compartilhamento de dados parece ser uma característica chave que atrai os usuários para a serviços de armazenamento na nuvem. No entanto, essa característica gera custos para o serviço. Tomando Dropbox como um exemplo, a menos que os usuários estejam localizados em uma única rede[1], todos os participantes em uma pasta compartilhada tem que recuperar dados a partir da nuvem. Isso potencialmente cria uma carga de trabalho extra para servidores, além de gerar custos com largura de banda [Gonçalves et al. 2015].

Dado a importância e os custos de trabalhos colaborativos via compartilhamento de dados para os serviços de armazenamento na nuvem, neste trabalho buscamos entender até que ponto essa funcionalidade pode revelar o perfil de atividade do usuário. Nesse sentido, nós buscamos responder as seguintes questões: *(Q1) A atividade de um usuário em serviços de armazenamento na nuvem pode ser medida pela sua quantidade de colaborações? (Q2) Qual o impacto de trabalho colaborativo na utilização de serviços de armazenamento na nuvem em grandes redes?* Note que, embora alguns estudos tenham abordado características de compartilhamentos em serviços de armazenamento na nuvem, essas questões ainda não foram respondidas, visto que o foco desses estudos eram padrões de carga e tráfego [Gonçalves et al. 2014] e aspectos de arquitetura visando melhorar o suporte a compartilhamento de dados [Gonçalves et al. 2015].

Para tal, nós conduzimos análises detalhadas sobre essas questões no Dropbox — um dos serviços mais populares atualmente. Todas as nossas análises são baseadas em dados de tráfego coletados em quatro redes distintas: duas redes universitárias (uma na América do Sul e outra na Europa) e dois Pontos de Presença (PoP) de um provedor de

[1]Dropbox implementa o protocolo `LanSync` para sincronização local (fora do escopo desse trabalho).

serviço de Internet (ISP) europeu[2]. Mais especificamente, nós correlacionamos o nível de atividade dos usuários, em termos de volume de dados sincronizados com a nuvem, com características do Dropbox relacionadas a trabalho colaborativo.

Nossos resultados indicam que usuários engajados em trabalhos colaborativos sincronizam maior volume de dados com a nuvem. Logo, esses usuários tem maior atividade e potencialmente pagam ou podem vir a pagar pelo serviço. Visto que trabalho colaborativo gera custos para o provedor, nosso estudo é importante porque aponta justificativas para o provedor investir em tal funcionalidade, que pode levar a um aumento de receita, assim como atrair novos usuários. Em suma, nossos principais resultados são: (i) o número de compartilhamentos do usuário indica o seu nível de atividade, em volume de sincronizações de dados com a nuvem, dado que existe uma correlação alta e não linear entre essas duas variáveis em diferentes redes; e (ii) três perfis de atividade podem descrever o comportamento dos usuários de serviços de armazenamento na nuvem em redes de universidades e ISPs: usuários que focam em colaborações dentro da rede ou colaborações externas à rede, e usuários que focam basicamente em armazenar dados.

A seguir, na Seção 2 são discutidos trabalhos relacionados. Na Seção 3 são apresentados conceitos básicos do Dropbox e a metodologia adotada para coleta e análise de dados. Na Seção 4 são discutidos resultados da correlação entre o nível de atividade dos usuários e características do Dropbox relacionadas a trabalho colaborativo. Impactos de trabalhos colaborativos nas redes monitoradas e a identificação/caracterização de perfis de usuários são tratados na Seção 5. Por fim, na Seção 6 conclui-se esse estudo.

## 2. Trabalhos Relacionados

Este trabalho analisa a relação de trabalhos colaborativos via compartilhamento de dados com a atividade dos usuários em sistema de armazenamento na nuvem por meio de medições de tráfego em várias redes. Nosso trabalho mais recente [Gonçalves et al. 2015] também analisa compartilhamentos no Dropbox, mas com foco específico em downloads redundantes e seu impacto no tráfego de uma rede universitária. Em [Costa et al. 2014] foi avaliado alguns aspectos do compartilhamento entre usuários do Dropbox usando uma coleta de dados de voluntários. Diferente desses, este trabalho faz uma análise mais abrangente sobre a utilização de recursos de compartilhamentos do Dropbox e o seu impacto na atividade do usuário e no tráfego de dados do serviço, considerando redes acadêmicas e residenciais.

Serviços de armazenamento na nuvem têm recebido grande atenção da comunidade científica. Em [Drago et al. 2012] é apresentado a primeira caracterização detalhada acerca da utilização e desempenho do Dropbox. No presente trabalho, nós contamos, em parte, com a metodologia de coleta e processamento de dados desenvolvida em [Drago et al. 2012]. Porém, nós estendemos essa metodologia para uma nova análise sobre o volume de dados sincronizado por usuários do Dropbox. Essa análise utiliza a combinação entre o volume dos fluxos de dados e as mensagens de sincronização dos clientes Dropbox. Tal método foi preliminarmente desenvolvido em nosso trabalho anterior [Gonçalves et al. 2015].

Grande parte dos estudos sobre serviços de armazenamento na nuvem focam em aferição (*benchmarking*). Nesse sentido, esses trabalhos utilizam primordialmente

[2]Os leitores interessados podem contactar os autores para acessar o conjunto de dados.

medições ativas, que são experimentos realizados em ambiente controlado para avaliar alguma funcionalidade do serviço. Em [Wang et al. 2012] são analisados gargalos do Dropbox, enquanto APIs de três provedores desse tipo de serviço são avaliadas em [Gracia-Tinedo et al. 2013]. Um método para medir a eficiência do uso de tráfego em serviços de armazenamento na nuvem é apresentado em [Li et al. 2014], enquanto um *framework* para realizar aferições automáticas de vários aspectos de aplicações clientes foi proposto em [Bocchi et al. 2015a].

Outros trabalhos propõem novos protocolos e arquiteturas para melhorar o desempenho, ou reduzir custos de serviços de armazenamento nas nuvens. Em [Chen et al. 2014] foi proposto um *framework* que coleta informações semânticas sobre os arquivos (e.g., tipo de arquivo) visando melhorar o desempenho do serviço. Um mecanismo que garante consistência entre os arquivos locais (em vários dispositivos que o compartilham) e o repositório remoto foi apresentado em [Zhang et al. 2014]. Em [Cui et al. 2015] foi proposto QuickSync, um sistema que otimiza a sincronização entre dispositivos e a nuvem combinando técnicas como *network-aware chunking*. Nosso trabalho contribui diretamente para esses esforços uma vez que, cientes da relação entre trabalhos colaborativos e a atividade dos usuários, os desenvolvedores podem propor protocolos e arquiteturas otimizados para esses serviços.

A Qualidade de Experiência (QoE) do usuário em serviços de armazenamento na nuvem foi estudada em [Amrehn et al. 2013, Casas and Schatz 2014]. Os autores realizaram experimentos para correlacionar QoE e medições de QoS, observando que largura de banda é a métrica de QoS mais sensível para os usuários em termos de QoE. Essas correlações são importantes para orientar ISPs, provedores de armazenamento na nuvem e usuários sobre os requisitos de largura de banda necessárias para uma utilização satisfatória do serviço. Por outro lado, o nosso objetivo é ir além desses requisitos técnicos básicos, medindo funcionalidades do serviço como trabalho colaborativo que podem ter impacto significativo sobre a atividade do usuário.

Finalmente, estudos mais recentes caracterizaram o comportamento de usuários a partir de dados coletados em redes ou provedores de serviço. Em [Gracia-Tinedo et al. 2015] foi apresentado um estudo de medição do serviço Ubuntu One (U1). Os autores observaram que U1 é usado mais como um serviço de armazenamento do que trabalho colaborativo, dado que apenas 1,8% dos usuários compartilham dados nesse serviço. Um estudo comparativo sobre a utilização de armazenamento na nuvem em diferentes terminais foi proposto em [Bocchi et al. 2015b]. Os autores constataram que o volume de *uploads* é maior em terminais móveis, possivelmente associado ao armazenamento de conteúdo multimedia, ao passo que o volume de *downloads* é maior em PCs devido à sincronização desse conteúdo. Nenhum desses estudos investigam trabalho colaborativo e a atividade dos usuários, como proposto no presente estudo.

## 3. Conceitos e Metodologia de Coleta de Dados

Antes da apresentação de nossas análises, nós sumarizamos o mecanismo básico de sincronização de dados no Dropbox. Informações mais detalhadas e uma descrição abrangente dos mecanismos internos do Dropbox podem ser encontradas em [Drago et al. 2012, Gonçalves et al. 2016].

### 3.1. Dropbox: Mecanismo Básico

O Dropbox sincroniza dados utilizando dois conceitos principais: *dispositivos* e *namespaces*. Os usuários podem registrar vários dispositivos nesse serviço. Durante esse processo, os usuários devem selecionar uma pasta inicial a partir de onde os arquivos são sincronizados com a nuvem. Esta pasta inicial é visível em qualquer outro dispositivo pertencente ao usuário. Os usuários podem também compartilhar dado com outros usuários através da criação de pastas compartilhadas, que são visíveis em todos os dispositivos de todos os usuários que participam do compartilhamento. A pasta inicial e as pastas compartilhadas são raízes de árvores de diretórios independentes conhecidas como *namespaces* no Dropbox[3].

Nós nos concentramos apenas na aplicação cliente para PC, uma vez que ela é responsável por mais de 75% do tráfego Dropbox [Bocchi et al. 2015b]. Os dispositivos que utilizam esse cliente mantêm, geralmente, uma cópia local de todos os arquivos presentes nos *namespaces* do usuário. A adição de qualquer dado em um *namespace* desencadeia a propagação desse dado: todos os dispositivos com o cliente, e com o *namespace* registrado, recuperam o dado imediatamente, caso estejam em execução (ou tão logo eles iniciem sua execução).

Internamente, Dropbox controla o estado dos *namespaces* por meio de um *protocolo de notificações*. Em resumo, cada *namespace* está associado a um identificador monotônico (i.e., *Journal Identifier* ou JID), representando sua atualização mais recente. Dispositivos descobrem se há atualizações pendentes para *namespaces* via troca periódica de uma lista de *namespaces* e seus respectivos JIDs com os servidores do Dropbox. Se algum *namespace* está desatualizado em um dispositivo, esse dispositivo executa várias transações com os servidores Dropbox até que todos os *namespaces* estejam sincronizados com a nuvem.

Assim, pela observação das mensagens do protocolo de notificação, é possível identificar quando os *namespaces* são atualizados. De fato, nós desenvolvemos um método em [Gonçalves et al. 2016] para coletar dados a longo prazo de um conjunto de *namespaces* do Dropbox em grandes redes como campi universitários e ISPs. O nosso método inclui estimativas para o volume de tráfego transferido em cada transição de JIDs em uma grande amostragem de *namespaces*. No presente trabalho nós exploramos o conjunto de dados obtidos por esse método para analisar trabalhos colaborativos no Dropbox.

### 3.2. Coleta de Dados

Nós contamos com dados capturados e pré-processados em nosso trabalho anterior [Gonçalves et al. 2016] para realizar o estudo atual. Esses dados foram capturados através da monitoração do tráfego Dropbox em 4 pontos de coleta diferentes, incluindo dois campi universitários e duas redes de ISPs.

Mais precisamente, os pontos de coleta Campus-1 e Campus-2 estão em redes de campus distintos, na América do Sul e Europa, respectivamente. Campus-1 tem uma população de usuários (incluindo professores, alunos e funcionários) de aproximadamente $57.000$ pessoas, enquanto o Campus-2 serve cerca de $15.000$ pessoas. PoP-1 e PoP-2 monitoram clientes em dois pontos de presença (PoPs) de um ISP europeu, agregando mais de $25.000$ e $5.000$ residências, respectivamente.

[3] `https://blogs.dropbox.com/tech/2014/07/streaming-file-synchronization`

Nossos dados incluem (i) informações sobre o volume de tráfego trocado por clientes com servidores Dropbox; (ii) metadados extraídos de mensagens de notificação Dropbox. Esse último foi capturado por meio de uma inspeção detalhada de pacotes e inclui, para cada mensagem, um identificador do dispositivo cliente e sua lista de *namespaces* com respectivos JIDs. Note que as mensagens de notificação não oferecem informações sobre identidades de usuários. Mais ainda, por razões de privacidade, os endereços IP dos clientes são anonimizados nos pontos de coleta de dados. Analisamos 27 k dispositivos Dropbox únicos e 61 k *namespaces* exclusivos durante cerca de 1 ano de coleta de dados. No total, coletamos mais de 20 TB de tráfego trocados com servidores Dropbox, onde observamos que o volume de compartilhamento de dados é significativo: mais de 60% do tráfego coletado é *download* e até 70% dos *downloads*, dependendo da rede[4], estão relacionados com pastas compartilhadas.

Os clientes Dropbox, quando em execução, trocam mensagens de notificação com os servidores a cada minuto aproximadamente. Portanto, nós rastreamos essas notificações para reconstruir um conjunto de dados que resume as atualizações realizadas nos *namespaces*. Nós estimamos o volume de cada atualização, correlacionando o volume do fluxo de dados de cada cliente com as suas mensagens de notificação: considerando os protocolos Dropbox [Drago et al. 2012], notificações a respeito de uma nova atualização de *namespaces* são observadas na rede próximas (antes/depois) à transmissão do fluxo de dados. Assim, ao combinar esses fluxos com as mensagens de notificação (quase) simultâneas de um cliente, nós associamos mais de 63% do tráfego Dropbox com mensagens de notificação. Essa heurística mostrou-se capaz de produzir uma boa estimativa para o volume de tráfego de cada atualização em *namespaces* Dropbox como descrevemos com mais detalhes em [Gonçalves et al. 2016].

## 4. Análise da Atividade de Usuários

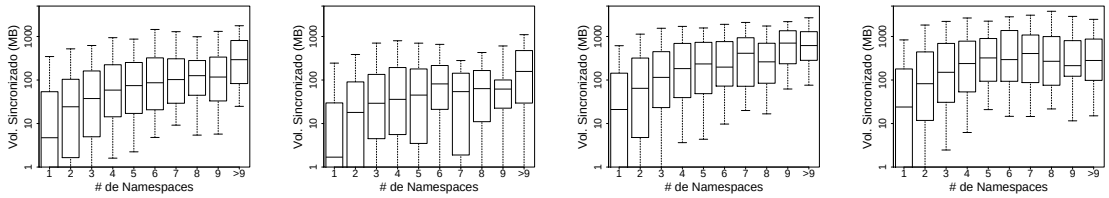
Nesta seção analisamos como métricas associadas ao trabalho colaborativo se correlacionam com atividade dos usuários no Dropbox. Nosso objetivo é responder a questão *Q1: A atividade de um usuário em serviços de armazenamento na nuvem pode ser medida pela sua quantidade de colaborações?*

Nesse sentido, definimos como atividade do usuário o volume de suas sincronizações ao longo de um período. Tais sincronizações correspondem aos *uploads* e *downloads* entre o(s) dispositivo(s) do usuário e os servidores remotos de armazenamento do Dropbox.

Entre os principais recursos que usuários utilizam no serviço de armazenamento na nuvem para trabalho colaborativo, nós analisamos a correlação entre (1) número de *namespaces*; (2) número de dispositivos e; (3) número de parcerias, i.e., demais usuários da rede ligados às pastas compartilhadas do usuário sendo analisado; com a atividade desse usuário. O volume total de atividades dos usuários, bem como dessas três métricas foram agrupados em intervalos de tempo de um mês. Desse modo, foi possível considerar usuários com diferentes frequências de utilização do serviço. Adicionalmente, um usuário pode ter mais de uma amostra referente a sua atividade no nosso conjunto de dados.

As correlações foram calculadas da seguinte forma: primeiro, as métricas de trabalho colaborativo foram agrupadas de acordo com o seu valor. Por exemplo, ao corre-

[4]Por restrição de espaço, nós detalhamos em [Gonçalves et al. 2016] as informações de cada rede.




(a) Campus-1 (b) Campus-2 (c) PoP-1 (d) PoP-2

**Figura 1. Número de *namespaces* vs. Volume de Sincronizações dos Usuários.**

**Tabela 1. Coeficiente de correlação linear (Pearson) e não linear (Spearman) entre atividade dos usuários e métricas relacionadas a trabalhos colaborativos.**

| | Correlaçao Linear (Pearson) | | | | Correlaçao Nao linear (Spearman) | | | |
|---|---|---|---|---|---|---|---|---|
| | Campus-1 | Campus-2 | PoP-1 | PoP-2 | Campus-1 | Campus-2 | PoP-1 | PoP-2 |
| Parcerias | 0,44 | 0,65 | 0,83 | -0,42 | 0,26 | 0,59 | 0,79 | -0,48 |
| Dispositivos | 0,31 | 0,30 | 0,49 | 0,61 | 0,35 | 0,68 | 0,48 | 0,64 |
| Compartilhamentos | 0,87 | 0,84 | 0,90 | 0,69 | 0,99 | 0,92 | 0,96 | 0,59 |

lacionar quantidade de *namespaces* do usuário com a sua atividade, nós criamos grupos de tamanho 1, 2, 3, ou mais *namespaces*. Acumulamos usuários com valores acima de 9 *namespaces* em um único grupo, dado que observamos menos usuários com tais valores. Por fim, calculamos a correlação entre a métrica e o valor da mediana da distribuição da atividade dos usuários em cada grupo.

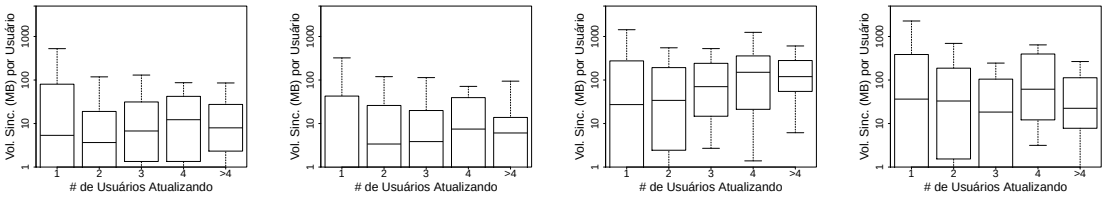
A Figura 1 exemplifica as correlações calculadas, com a métrica total de *namespaces* por mês, que apresentou maior correlação com a atividade do usuário (Tabela 1). O processo para calcular as correlações é análogo para as demais métricas, e seus gráficos foram omitidos por questão de espaço. Nessa figura, para cada total de *namespaces*, usamos o *box-plot* para sumarizar a distribuição da atividade dos usuários: o retângulo central se expande entre o primeiro e terceiro quartil, o segmento interior é a mediana, enquanto os indicadores abaixo e acima do retângulo representam o $9^{o}$ e $91^{o}$ percentis.

Note que a distribuição da atividade em cada grupo é muito variável e tende a valores pequenos, visto pela concentração de volumes entre $9^{o}$ percentil e o $3^{o}$ quartil (66% dos dados). Logo, consideramos a mediana como o valor mais adequado para representar a atividade por grupo. Observe na Figura 1 que a mediana da atividade dos usuários tende a crescer com o aumento do número de *namespaces*. Relembre (Seção 3.1) que cada *namespace* se refere à pasta inicial acessada apenas por um usuário Dropbox ou às demais pastas que são compartilhadas entre diferentes usuários do Dropbox. Logo, nossos resultados mostram que a intensidade de atividades dos usuários está relacionada à sua quantidade de trabalhos colaborativos (i.e., *namespaces*). Note que usuários que sincronizam mais que cinco *namespaces* por mês apresentam comportamento mais homogêneo, com uma distribuição mais concentrada em torno da mediana e tendem a sincronizar volumes altos de dados, i.e., mediana acima de 100MB e próxima a 1GB nas redes residenciais.

A Tabela 1 mostra os resultados dos coeficientes de correlação de Pearson[5] e de Spearman[6] [Jain 1991] referentes a cada uma das correlações calculadas. Note que o número de *namespaces* é a métrica que apresenta maior correlação com a atividade do

[5]Mede a correlação linear entre duas variáveis.

[6]Mede a correlação não-linear entre duas variáveis.




(a) Campus-1 (0.8) (b) Campus-2 (0.9) (c) PoP-1 (0.9) (d) PoP-2 (-0.2)

**Figura 2. Número de usuários ativos em um *namespace* vs. volume médio de sincronizações por usuário, com respectivos coeficientes de correlação não linear (Spearman) para cada rede.**

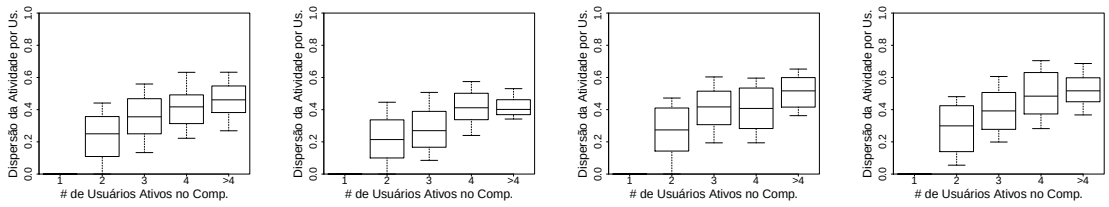
usuário em todas as redes. É importante observar que, essa correlação para as redes Campus-1, Campus-2 e PoP-1 tende a ser não linear (note que o coeficiente de Spearman é maior que o de Pearson). Isso indica que trabalhos colaborativos nessas redes estão realacionadas a uma atividade ainda mais expressiva dos usuários, i.e., volume maior de sincronizações.

Dado a importância do número de *namespaces* para caracterizar a atividade dos usuários, analisamos também as colaborações (i.e., compartilhamentos) que ocorrem somente internamente às redes analisadas. Para isso, nós calculamos o volume de sincronizações por *namespace* e dividimos esse volume igualmente pelo total de usuários ativos no *namespace*[7]. Assim, obtivemos a atividade média por usuário colaborando em um *namespace* dentro da rede.

A Figura 2 mostra a distribuição para o volume de sincronizações por usuário em *namespaces* dentro das redes. Mantivemos os *namespaces* com apenas um usuário ativo para efeito de comparação com multiplos usuários. Pode-se observar que existe uma correlação não linear alta (i.e., acima de 0,6) entre o volume de sincronizações mediano por usuário e o total de usuários ativos em *namespaces*. A correlação é baixa apenas no PoP-2 possivelmente devido a flutuações do volume de sincronizações de alguns *namespaces*, que foram maiores nessa rede. Contudo, quando consideramos o número de modificações por usuário ao invés do volume, observamos uma correlação acima de 0,65 em todas as redes. Esses resultados mostram indícios que os compartilhamentos dentro das redes tem um potencial para incentivar a atividade dos usuários.

Observamos ainda se poucos usuários dentro das redes são responsáveis pela maior parte do volume de dados sincronizados em cada *namespace*. Para isso nós calculamos o coeficiente de Gini como indicador de desigualdade [Cowell 2011]. Esse coeficiente varia entre 0 (completa igualdade) e 1 (apenas um usuário sincronizou dados no *namespace*). A Figura 3 mostra que a atividade por usuário se torna mais desequilibrada à medida que o *namespace* tem mais usuários. Em redes residenciais (PoPs) o coeficiente de Gini atinge valores superiores a 0,5 em *namespaces* com mais de 4 usuários. De fato, observamos que o usuário mais ativo impacta em pelo menos 40% (porcentagem mediana) da atividade total por *namespace* em todas as redes. Note que mesmo havendo esse desequilibrio, os dados do *namespace* provavelmente foram sincronizados para todos os seus usuários igualmente, estando alguns deles fora das redes monitoradas.

[7]Consideramos apenas os usuários registrados no *namespace* que realizaram alguma sincronização de dados dentro das redes monitoradas.




(a) Campus-1 (b) Campus-2 (c) PoP-1 (d) PoP-2

**Figura 3. Dispersão da atividade do usuário medida entre 0-1 (Gini) vs. número de usuários ativos no *namespace*.**

Resumindo, os resultados apresentados nessa seção mostram indícios de que o engajamento de usuários em trabalhos colaborativos, medidos pelo número de *namespaces* do usuário (Figura 1) ou sincronizações em compartilhamentos dentro das redes (Figura 2), podem revelar quais são os usuários mais ativos em sistemas de armazenamento em nuvem. Em [Drago et al. 2012, Gracia-Tinedo et al. 2015] esses usuários são apontados como *heavy users* pois são eles quem impactam o sistema de armazenamento em si. No entanto, apesar de definirem o que são *heavy users* tais estudos não observam a correlação existente entre trabalhos colaborativos e a atividade desses usuários.

## 5. Trabalho Colaborativo em Redes

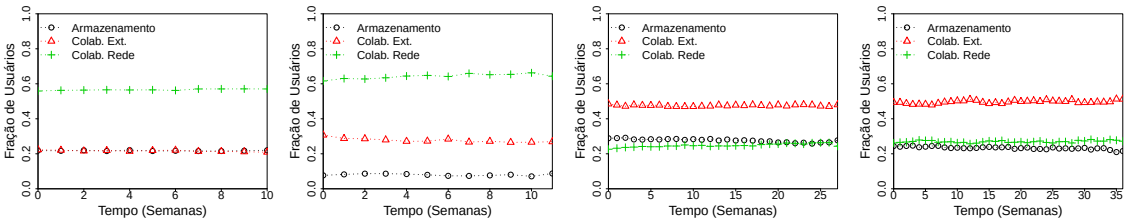
O foco principal desta seção é responder a questão *Q2: Qual o impacto de trabalhos colaborativos na utilização de serviços de armazenamento na nuvem em grandes redes?* Nesse sentido, propomos um modelo que considera diferentes estados de atividade dos usuários em redes acadêmicas e residenciais (Seção 5.1). A partir desse modelo analisamos os perfis de usuários baseados em transições entre esses estados e caracterizamos a carga desses perfis no serviço de armazenamento (Seção 5.2).

### 5.1. Estados de Atividade do Usuário

Modelamos as atividades dos usuários considerando três diferentes estados: (1) *Colaborações em Rede*: usuários colaborando com outros usuários internamente às redes monitoradas; (2) *Colaborações Externas*: usuários colaborando com usuários externos às redes monitoradas; e (3) *Armazenamento*: usuários cujo foco principal é basicamente armazenar dados na nuvem e sincronizá-los entre seus dispositivos.

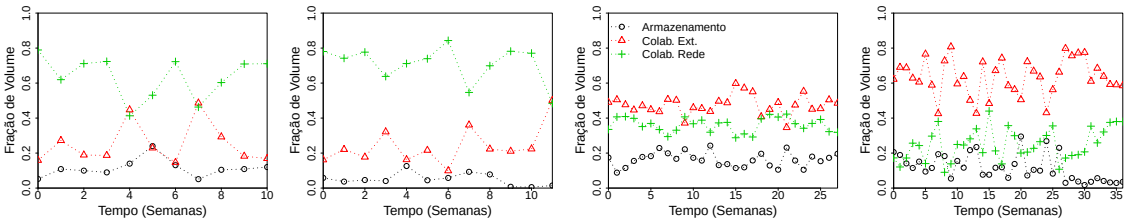
No primeiro estado de atividade (*Colaborações em Rede*), incluímos usuários que sincronizaram dados em ao menos um *namespace* compartilhado dentro da rede monitorada. Nos demais estados, incluímos usuários cujas sincronizações ocorreram apenas em *namespaces* não compartilhados dentro da rede. Esses *namespaces* podem ser a pasta inicial do usuário e/ou pastas compartilhadas com usuários externos à rede. Assim, o segundo estado (*Colaborações Externas*) contém usuários que sincronizam mais de um *namespace*, ou seja, usuários que realizaram colaborações com usuários externos à rede, dado que o Dropbox define uma pasta inicial por usuário (ver Seção 3.1). Por sua vez, o terceiro estado (*Armazenamento*) contém usuários que sincronizaram um único *namespace*. Como esse *namespace* pode se tratar da pasta inicial, conjecturamos que o foco desses usuários é armazenar dados na nuvem e sincronizá-los entre seus dispositivos.

A seguir, nós analisamos o número de usuários e o volume de sincronizações para cada estado de atividade por semana. A análise semanal evita os efeitos da sazo-




(a) Campus-1 (b) Campus-2 (c) PoP-1 (d) PoP-2

**Figura 4. Fração de usuários nos três estados de atividade ao longo do tempo.**




(a) Campus-1 (b) Campus-2 (c) PoP-1 (d) PoP-2

**Figura 5. Fração do volume sincronizado nos três estados de atividade ao longo do tempo.**

nalidade que ocorrem em diferentes dias, por exemplo, finais de semanas e dias úteis. Adicionalmente, analisamos apenas usuários estáveis, i.e., usuários monitorados em pelo menos metade do período de coleta em cada rede. Essa medida diminui a interferência de usuários novos ou visitantes às redes monitoradas nos resultados de nossas análises.

A Figura 4 mostra a fração de usuários em cada estado ao longo do tempo nas redes monitoradas. De uma forma geral, observa-se que colaborações em rede são dominantes nos campi (Campus-1 e Campus-2) e menos presentes nas residências (PoP-1 e PoP-2), como foi observado em [Drago et al. 2012]. Observamos também que colaborações externas são as atividades dominante nas redes residenciais. Esses resultados fornecem indícios de que serviços populares de armazenamento na nuvem, como Dropbox, são utilizados intensivamente para realização de trabalhos colaborativos. Tal comportamento é diferente em serviços menos populares, como o Ubuntu One, onde apenas 1,8% dos usuários realizam trabalhos colaborativos [Gracia-Tinedo et al. 2015].

Ainda com relação aos estados de atividade dos usuários (Figura 4) observa-se que a fração de usuários colaborando em redes nos campi é pelos menos duas vezes superior a cada um dos demais estados. Nas residências, a fração de colaborações em rede é aproximadamente metade da fração de colaborações externas, aproximando-se da fração de usuários com foco em armazenamento. Contudo, as Figuras 4(c-d) mostram uma tendência leve de crescimento de colaborações nas residências, ao passo que a atividade com foco em armazenamento diminui. Esses resultados mostram que os provedores de serviço devem se preparar para atender a uma maior demanda por colaborações em redes via arquiteturas alternativas de sincronização, que diminuem custos com largura de banda [Gonçalves et al. 2015].

A Figura 5 mostra o volume das sincronizações, i.e., volumes de *upload* e *download*, dos usuários em cada estado ao longo do tempo nas redes monitoradas. O vo-

lume de sincronizações em todas as redes apresenta flutuações devido a características de cauda pesada dos fluxos de dados dos usuários [Gonçalves et al. 2014]. Essas flutuações são mais intensas em residências devido à maior frequência de sincronização de arquivos multimídia volumosos (e.g., fotos e vídeos) [Gonçalves et al. 2016]. Apesar dessas flutuações, pode-se observar novamente que colaborações em redes é a atividade com tráfego dominante nos campi. Nas residências, diferentemente da fração de usuários mostrada na Figura 4(c-d), a fração de sincronizações para colaborações em redes se aproxima da fração de sincronizações para colaborações externas. Observa-se também que a fração de sincronizações dos usuários com foco em armazenamento é a menor na maioria das semanas em todas as redes.

Em resumo, nessa seção observamos que trabalhos colaborativos tem um impacto significativo no Dropbox em termos de número de usuários e volume de sincronizações em redes como universidades e ISPs. Pelo menos 55%(22%) dos usuários nos campi(residências) realizam ao menos uma colaboração em rede semanalmente, ao passo que, para colaborações externas essa porcentagem é pelo menos de 20%(47%). Em relação ao tráfego gerado por sincronizações, os usuários que colaboram em rede representam pelo menos 41%(9%) do tráfego do serviço nos campi(residências). Considerando os usuários com colaborações externas essa porcentagem é pelo menos 10%(35%).

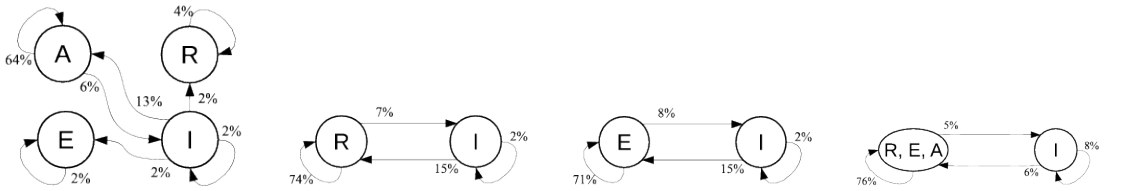
### 5.2. Perfis de Usuários

Um mesmo usuário pode estar em diferentes estados de atividade ao longo do tempo. Nesta seção, nós analisamos a dinâmica da transição dos usuários entre os estados de atividades definidos na seção anterior. Nesse sentido, nós identificamos perfis de usuários considerando a maneira que eles realizam essas transições e caracterizamos esses perfis.

Para identificar perfis de usuários, nós calculamos, para cada usuário, a frequência de transições entre os estados de atividade: colaborando em rede (R), colaborando com usuários externos (E), armazenamento (A) e inatividade (I) que significa o não armazenamento de dados na nuvem. A seguir, consideramos cada usuário como um vetor de dezesseis elementos (i.e., possibilidades de transições entre os quatro estados), e executamos o algoritmo K-means para agrupar usuários similares quanto às transições de atividade. Esse algoritmo requer a escolha do número de grupos $k$ (i.e., *clusters*). Nós empregamos diferentes medidas de qualidade de agrupamento para escolher o valor de $k$, como coeficientes de variação intra, inter *clusters* e beta-CV, além de inspeção visual dos centróides dos agrupamentos. Baseado nessas medidas chegamos à conclusão que $k = 3$ grupos é o suficiente para representar os perfis de atividade dos usuários.

A Figura 6 representa cada um dos perfis identificados (i.e., centróides do algoritmo K-means) através de grafos direcionados. Em cada grafo os vértices são os estados de atividade e as arestas são as porcentagens (ou probabilidades) de um usuário realizar uma transição entre dois estados de atividade diferentes ou permanecer no mesmo estado. Para melhor clareza na figura, omitimos transições com porcentagens menores que 2%. Nós identificamos perfis de usuários para as quatro redes analisadas, mas mostraremos resultados apenas para Campus-2 e PoP-2 representando respectivamente os campi e residências, dado que obtivemos conclusões semelhantes para Campus-1 e PoP-1.

As Figuras 6(a-c) mostram os três perfis de atividade dos usuários no Campus-2. Note que cada perfil tem um estado de atividade predominante. Por exemplo, a Figura 6(a) mostra um perfil de usuários com foco em armazenamento, i.e., 64% das transições




(a) Campus-2 Perfil A (b) Campus-2 Perfil R (c) Campus-2 Perfil E (d) PoP-2 Perfis R, E e A

**Figura 6. Perfis de atividade dos usuários nomeados em cada item pelo estado predominante em cada perfil: foco em armazenamendo dados (A), compartilhamentos na rede (R), compartilhamentos externos (E) e inativo (I).**

**Tabela 2. Caracterização dos perfis de usuários: mediana(média) do número de parcerias, compartilhamentos e volume sincronizado para usuários do perfil**

| | Campus-2 | | PoP-2 | |
|---|---|---|---|---|
| Perfil | Compartilhamentos | Vol. Sincronizado (MB) | Compartilhamentos | Vol. Sincronizado (MB) |
| A | 1 (1,1) | 3,9 (267,9) | 1 (1,1) | 41,6 (450,9) |
| R | 3 (3,8) | 20,4 (253,3) | 2 (3,4) | 70,7 (492,8) |
| E | 1 (1,9) | 11,8 (212,0) | 2 (2,4) | 82,2 (846,0) |

* Perfil determinado pela atividade predominante.

dos usuários consistem em permanecer na atividade A. Os perfis (b) e (c) mostram que usuários realizando trabalhos colaborativos tem uma probabilidade menor de transitar para o estado de inatividade (7-8%), e dificilmente permanecem nesse estado (apenas 2% das transições). A porcentagem de usuários que o algoritmo K-Means atribuiu para cada perfil se aproxima das proporções que medimos na seção anterior (Figura 4 (a-b)). Os perfis das Figuras 6 a, b e c (i.e., A, R e E) tem respectivamente 7%, 67% e 26% dos usuários.

Na Figura 6(d) mostramos uma única imagem para exemplificar os três perfis de atividade do usuário no PoP-2. Nas redes residenciais, usuários nos estados R, E ou A tendem a permanecer no mesmo estado (76% das transições) ou transitam para o estado I (5% das transições). Note que, diferentemente dos campi, os usuários nas residências tendem a permanecer inativos por mais tempo (8% das transições). A porcentagem de usuários nos perfis onde predominam as atividades A, R ou E são respectivamente 21%, 28% e 51%. Nota-se uma porcentagem menor para o perfil de usuários colaborando dentro das redes, corroborando com os resultados apresentados nas Figuras 4(c-d).

Finalmente, caracterizamos os três perfis de atividade do usuário identificados acima. Novamente, Campus-2 e PoP-2 representam respectivamente as redes dos campi e residências. A Tabela 2 mostra a mediana (e média) do número de parcerias, *namespaces* e volume sincronizado por usuários dentro de cada perfil. Para os campi, pode-se observar que o perfil R representa os usuários que impõem maior carga em volume de sincronizações no sistema. Em valores medianos, os usuários nesse perfil sincronizam 20MB e 3 *namespaces* por mês. Contudo, valores médios de sincronizações por mês superiores a 200MB indicam que há usuários com volume de dados muito acima dos valores típicos em todos os perfis.

Para os perfis residenciais, a Tabela 2 mostra que o volume de sincronizações é superior aos respectivos perfis nos campi. Possivelmente, isso ocorre devido à maior sincronização de conteúdo multimídia como fotos e vídeos nas residências. Outra

observação interessante é que o volume mediano de sincronizações no perfil R se aproxima em cerca de 86% do perfil E (majoritário em redes residenciais). Isso mostra que a colaboração entre usuários em uma mesma residência ou sob o mesmo ISP é expressiva, mesmo que esse perfil ainda não seja majoritário nessas redes.

Em resumo, identificamos três perfis de atividade (semanal) do usuário considerando trabalhos colaborativos em sistemas de armazenamento na nuvem em redes de universidades e ISPs residenciais: colaborações dentro da rede monitorada (R), colaborações externas às redes (E) e foco em armazenamento de dados (A). Os perfis R e E são majoritários em redes universitárias e residenciais respectivamente, mostrando que serviços populares como Dropbox vem sendo usados intensivamente para trabalhos colaborativos ao invés de puramente armazenamento de dados na nuvem.

## 6. Conclusões

Neste artigo apresentamos uma investigação sobre a relação entre trabalho colaborativo e o nível de atividade dos usuários no Dropbox, um dos serviços de armazenamento na nuvem mais populares atualmente. Nossa principal conclusão é que tal serviço vem sendo utilizado pelos usuários não apenas para armazenamento e *backup* de dados, mas principalmente para realização de trabalhos colaborativos via compartilhamento de dados.

Primeiramente mostramos que a atividade de um usuário em serviços de armazenamento na nuvem pode ser medida pela sua quantidade de colaborações. Observamos correlações altas e não lineares entre o nível de atividade do usuário (i.e., volume de dados sincronizados com a nuvem) e o seu número de compartilhamentos (*namespaces*). Observamos também que a média de sincronizações por usuário em compartilhamentos tende a aumentar em relação ao número de usuários nesses compartilhamentos. Esses resultados podem ser explorados pelos provedores de serviços, que ao aprimorarem as funcionalidades de colaboração/compartilhamento de dados entre usuários em seus serviços, podem aumentar as suas receitas decorrente do aumento da atividade dos usuários.

A seguir, mostramos que trabalhos colaborativos tem um impacto significativo no uso do Dropbox em termos de número de usuários e volume de sincronizações em grandes redes como universidades e ISPs residenciais. Observamos que, semanalmente, pelo menos 55% dos usuários nos campi realizam ao menos uma colaboração em rede, ao passo que, pelo menos 47% dos usuários em residências realizam ao menos uma colaboração com usuários externos ao seu provedor de Internet. Com base nessas análises identificamos três perfis de usuários que consideram colaborações internas ou externas às redes monitoradas e usuários que focam basicamente em armazenamento de dados na nuvem.

Trabalhos futuros incluem investigações de causalidade na relação entre crescimento de colaborações em rede e aumento de atividade do usuário e vice-versa. Pretendemos também construir modelos para quantificar e predizer a importância ou o peso de trabalhos colaborativos em diferentes redes, assim como funções de utilidade para usuários que indicam o quanto serviços de armazenamento na nuvem é importante para os mesmos e quando vale a pena pagar por esse serviço.

## Referências

Amrehn, P., Vandenbroucke, K., Hossfeld, T., Moor, K. D., Hirth, M., Schatz, R., and Casas, P. (2013). Need for Speed? On Quality of Experience for Cloud-based File Storage Services. In *Proc. of PQS*.

Bocchi, E., Drago, I., and Mellia, M. (2015a). Personal Cloud Storage Benchmarks and Comparison. *IEEE Transactions on Cloud Computing*, PP(99):1–14.

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

Casas, P. and Schatz, R. (2014). Quality of Experience in Cloud Services: Survey and Measurements. *Comput. Netw.*, 68(1):149–165.

Chen, F., Mesnier, M. P., and Hahn, S. (2014). Client-Aware Cloud Storage. In *Proc. of MSST*.

Costa, E., Costa, L., Drago, I., Vieira, A., Ziviani, A., da Silva, A. P. C., and de Almeida, J. M. (2014). Análise da Topologia Social do Dropbox. In *Proc. of WP2P+*.

Cowell, F. (2011). *Measuring Inequality*. Oxford University Press.

Cui, Y., Lai, Z., Wang, X., Dai, N., and Miao, C. (2015). QuickSync: Improving Synchronization Efficiency for Mobile Cloud Storage Services. In *Proc. of MobiCom*.

Drago, I., Mellia, M., Munafò, M. M., Sperotto, A., Sadre, R., and Pras, A. (2012). Inside Dropbox: Understanding Personal Cloud Storage Services. In *Proc. of IMC*.

Gonçalves, G., Drago, I., da Silva, A. P. C., de Almeida, J. M., and Vieira, A. (2014). Caracterização e Modelagem da Carga de Trabalho do Dropbox. In *Proc. of SBRC*.

Gonçalves, G., Drago, I., Vieira, A., Silva, A., Almeida, J., and Mellia, M. (2016). Workload models and performance evaluation of cloud storage services. *Computer Networks*. `http://dx.doi.org/10.1016/j.comnet.2016.03.024`.

Gonçalves, G., Drago, I., Vieira, A. B., da Silva, A. P. C., and de Almeida, J. M. (2015). Analyzing the Impact of Dropbox Content Sharing on an Academic Network. In *Proc. of SBRC*.

Gracia-Tinedo, R., Artigas, M. S., Moreno-Martinez, A., Cotes, C., and Lopez, P. G. (2013). Actively Measuring Personal Cloud Storage. In *Proc. of CLOUD*.

Gracia-Tinedo, R., Tian, Y., Sampé, J., Harkous, H., Lenton, J., García-López, P., Sánchez-Artigas, M., and Vukolic, M. (2015). Dissecting ubuntuone: Autopsy of a global-scale personal cloud back-end. In *Proc. of IMC*.

Jain, R. K. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley & Sons, New York, USA.

Li, Z., Jin, C., Xu, T., Wilson, C., Liu, Y., Cheng, L., Liu, Y., Dai, Y., and Zhang, Z.-L. (2014). Towards Network-Level Efficiency for Cloud Storage Services. In *Proc. of IMC*.

Shami, N. S., Muller, M., and Millen, D. (2011). Browse and discover: Social file sharing in the enterprise. In *Proc. of CSCW*.

Wang, H., Shea, R., Wang, F., and Liu, J. (2012). On the Impact of Virtualization on Dropbox-Like Cloud File Storage/Synchronization Services. In *Proc. of IWQoS*.

Zhang, Y., Dragga, C., Arpaci-Dusseau, A. C., and Arpaci-Dusseau, R. H. (2014). ViewBox: Integrating Local File Systems with Cloud Storage Services. In *Proc. of FAST*.