Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Topologia Social em um Sistema de Armazenamento em Nuvem

Leonardo Chinelate Costa

JUIZ DE FORA
AGOSTO, 2013

Topologia Social em um Sistema de Armazenamento em Nuvem

Leonardo Chinelate Costa

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Alex Borges Vieira

JUIZ DE FORA

AGOSTO, 2013

Topologia Social em um Sistema de Armazenamento em Nuvem

Leonardo Chinelate Costa

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Alex Borges Vieira
D. Sc.

Ana Paula Couto da Silva
D. Sc.

Luciano Jerez Chaves
M. Sc.

JUIZ DE FORA

23 DE AGOSTO, 2013

À minha família, por estar sempre ao meu lado.

Aos meus amigos e colegas, pela força.

Resumo

Os sistemas de armazenamento em nuvem vem ganhando cada vez mais notoriedade e popularidade devido a tamanha facilidade e praticidade oferecidas no armazenamento remoto de arquivos. Mesmo com a crescente popularidade desses sistemas, praticamente não existem trabalhos que busquem caracterizar o seu uso. Nesta direção, o presente trabalho apresenta uma caracterização preliminar de arquivos armazenados no Dropbox, o sistema de armazenamento em nuvem mais utilizado no mundo atualmente. Com a coleta de dados de usuários voluntários, foi possível caracterizar uso e carga do Dropbox. Além disso, com a análise dos mesmos dados, foram montadas redes que pudessem representar uma possível topologia social de compartilhamento de arquivos entre usuários nesse sistema. Os resultados mostram que documentos, em geral, representam a maior quantidade dos arquivos armazenados e o maior volume de bytes armazenados nas contas Dropbox. Já as redes formadas pelas interações entre usuários e arquivos sugerem que cerca de um terço dos usuários, apenas, compartilham arquivos entre si. Espera-se que, a partir deste trabalho, sejam feitos outros estudos que aprofundem tanto a caracterização quanto a estrutura topológica da rede, com o objetivo de se melhorar a eficiência do Dropbox e de se desenvolver outras ferramentas desse tipo.

Palavras-chave: Dropbox, armazenamento em nuvem, topologia social.

Abstract

Cloud storage systems is gaining more notoriety and popularity due to ease and convenience offered in remote file storage. Even with the growing popularity of these systems, there are virtually no studies that seek to characterize its use. In this direction, this paper presents a preliminary characterization of the files stored in Dropbox, the cloud storage system most used in the world today. With the collection of data volunteers users, we could characterize usage and load in Dropbox. Furthermore, with the analysis of the same data, networks that could represent a possible file sharing social topology among users in the system were assembled. The results show that documents represents the highest amount of files stored and the highest volume of bytes stored in Dropbox accounts. Already the networks formed by interactions between users and files suggest that about a third of the users only share files with each other. We hope that from this study, other studies are made to deepen both the characterization and the topological structure of the network, with the aims of improving the Dropbox efficiency and developing other applications.

Keywords: Dropbox, cloud storage, social topology.

Agradecimentos

Agradeço primeiramente a Deus, por me dar força e coragem nos bons e maus momentos vividos nesses cinco últimos anos, por me tranquilizar nos momentos mais difíceis e por me mostrar que fé e ciência podem, e devem, sempre andar juntas. Agora, mais do que nunca, Ele me mostra que depois de tantas dificuldades, podem sempre vir coisas boas. Tenho fé nisso!

Agradeço aos meus pais, Bernadete e Emanoel, que foram os maiores colaboradores de toda minha vida. Tudo o que sou hoje, devo a eles. Sem o esforço e a dedicação deles, sem seu apoio moral e financeiro, e sobretudo, sem seu amor incondicional, não seria possível alcançar esse objetivo em minha vida. Agradeço a eles por todas as lembranças que me vêm à cabeça nesse momento, desde o início da vida escolar até agora. Esse diploma pertence completamente a eles! A eles, minha gratidão eterna!

Agradeço ao meu irmão Eduardo, que posso considerar, sem dúvida alguma, o mentor intelectual da minha vida acadêmica. Por todo apoio, seja pela ajuda em trabalhos, seja por ouvir meus desabafos ou por entender perfeitamente tudo pelo que eu passei nesses últimos anos. Fico muito feliz por ter convivido com ele esses anos todos, em casa ou na faculdade, e mesmo que agora tomemos caminhos diferentes, que ele saiba que devo muito a ele e torço muito por seu futuro. Todo o sucesso do mundo, mestre!

Agradeço à minha namorada Cristiane, que por tanto tempo esteve do meu lado, como colega e amiga, passando as mesmas dificuldades que o curso e a vida nos impuseram. Agradeço pelas conversas até tarde, por todo o carinho e compreensão. Nunca imaginei que, de uma jornada tão árdua, pudesse surgir uma pessoa tão incrível, que me completasse tanto. Fico agora na torcida para que ela complete também a sua jornada, e que logo, possamos construir a nossa vida, juntos! Não desista nunca, amor!

Agradeço ao meu orientador Alex, pela orientação e pela amizade nesses últimos anos. Sem a sua ajuda, esse trabalho não seria feito. Obrigado por me atender nos

horários mais complicados. Obrigado pelo incentivo e pelos elogios. Significou muito para mim. Peço a ele que não desista, mesmo que a vida não esteja fácil nessa cidade e nessa universidade. Acredite, tudo vai melhorar!

Agradeço aos meus amigos e colegas que estiveram ao meu lado nessa caminhada de cinco anos, e que foram parte integrante desta conquista. Obrigado pelas conversas, pelos trabalhos em grupo e pelos momentos os quais nunca esquecerei em toda a minha vida. Obrigado por tornarem muito mais fácil e mais empolgante essa longa jornada. Um agradecimento especial ao meu amigo Diego, que foi meu companheiro em tantos desafios nesse curso. Não esquerecei de vocês, jamais!

Agradeço à UFJF e ao curso de Ciência da Computação por possibilitarem essa formação. Obrigado pelas aulas e pelo apoio indispensável para a conclusão dessa etapa em minha vida. Obrigado por me enriquecerem pessoalmente e profissionalmente.

"When I'm feeling weak and my pain walks down a one way street, I look above and I know I'll always be blessed with love.".

Robbie Williams (Angels)

Sumário

Lista de Figuras

Lista de Abreviações

DCC Departamento de Ciência da Computação

UFJF Universidade Federal de Juiz de Fora

1 Introdução

O modelo de computação em nuvem, ou *cloud computing*, está em crescente expansão na atualidade e sua aplicação em negócios, na indústria e na academia tem se tornado cada vez mais comum e mais necessária. Algumas das aplicações desse modelo que mais vêm ganhando notoriedade e popularidade são os sistemas de armazenamento em nuvem, ou *cloud storage*. Esse tipo de aplicação torna muito mais prática, segura e rápida a tarefa de armazenamento remoto de arquivos por parte de empresas e até de usuários comuns, a custos ínfimos para ambos. Esses benefícios acabaram tornando o *cloud storage* bastante atrativo, o que inclusive incentivou o ingresso de mais clientes finais e também de provedores de serviço como o Google, a Amazon e a Microsoft. Com mais clientes e mais provedores também cresce a demanda por serviços, o que consequentemente ocasiona um aumento no volume de dados trafegados e armazenados.

Contudo, mesmo com esse cenário favorável à criação de serviços baseados em nuvem e à utilização de sistemas de armazenamento em nuvem, e com a grande quantidade de informações e dados armazenados, praticamente não existem trabalhos que venham a caracterizar uso e carga desses sistemas. Além disso, quase não se possui informações que relacionam arquivos e usuários, que posteriormente poderiam vir a traçar um perfil dos usuários desses sistemas. Dessa forma, o conhecimento do comportamento do usuário e de informações sobre seus arquivos é de extrema importância, pois possibilitaria o desenvolvimento de sistemas de *cloud storage* mais fortes e de melhor desempenho.

Este trabalho, por sua vez, apresenta uma caracterização inicial dos arquivos armazenados no Dropbox, o sistema de armazenamento em nuvem mais utilizado no mundo atualmente. Além disso, as informações obtidas foram utilizadas para encontrar uma estrutura topológica que melhor representasse o sistema e as interações entre seus usuários e arquivos armazenados. Essa caracterização foi realizada a partir de dados obtidos dos arquivos de 333 usuários que participaram como voluntários do experimento. Os resultados mostram que a maioria dos arquivos armazenados nas contas Dropbox dos voluntários são de documentos em geral, como PDFs ou códigos-fonte de programas. O

maior volume em *bytes* armazenados nas contas também são de documentos. Já as redes formadas pelas interações entre usuários e arquivos sugerem que cerca de um terço dos usuários, apenas, compartilham arquivos entre si.

Uma série de trabalhos vêm, recentemente, destacando o crescimento e a importância dos sistemas de armazenamento em nuvem. Porém, o presente trabalho é o primeiro a apresentar uma estrutura topológica que pudesse representar um sistema de armazenamento em nuvem, com as conexões entre usuários e seus arquivos, além de uma caracterização dos arquivos encontrados. O trabalho de Drago et al (2013) também caracteriza os arquivos dos usuários Dropbox, mas não apresenta uma estrutura topológica do sistema. Alguns trabalhos focam no protoloco proprietário da aplicação Dropbox, como em Drago et al (2012). Outros comparam provedores de cloud storage, como em Hu et al (2010) e Gracia-Tinedo el al (2013). Finalmente, alguns outros trabalhos focam na segurança desses sistemas, como Mulazzani et al (2011) e Halevi et al (2011). Entretanto, a maioria dos trabalhos citados não fornecem informações sobre os arquivos armazenados, nem como diferentes tipos de arquivos podem causar diferentes impactos no sistema. Nenhum dos trabalhos citados propõe ou apresenta alguma estrutura topológica segundo usuários em conjunto aos seus arquivos.

O restante deste trabalho está assim organizado: o Capítulo 2 apresenta uma visão geral do sistema de armazenamento Dropbox; o Capítulo 3 descreve qual foi a metodologia para a coleta de dados dos usuários e quais são as métricas de interesse do problema; o Capítulo 4 descreve a caracterização dos arquivos dos usuários; o Capítulo 5 apresenta a estrutura topológica formada com as informações obtidas com os dados dos usuários; o Capítulo 6 traz um apanhado geral dos trabalhos relacionados ao tema; e por fim, o Capítulo 7 indica as conclusões deste trabalho e aponta possíveis direções para pesquisas futuras.

2 Dropbox

O Google Trends[1] sugere que o Dropbox é o serviço de *cloud storage* mais popular no mercado. A Figura 2.1 nos mostra que desde 2010 o Dropbox atraiu a maior parte das buscas no Google, se comparado a seus maiores concorrentes. De acordo com as analises conduzidas em Drago et al (2012), o Dropbox é responsável por cerca de 4% do volume tráfego em algumas redes. Este volume corresponde a cerca de 30% do trafego gerado pelo YouTube, um dos sistemas mais populares de distribuição de vídeo na Internet.

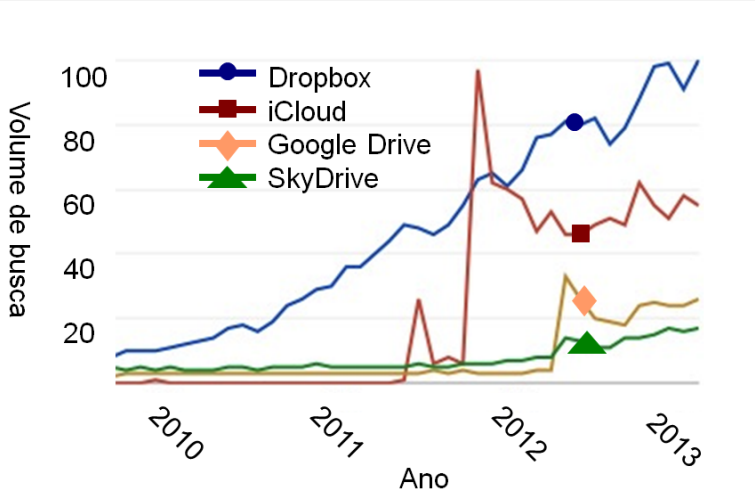

Figura 2.1: Interesse de busca por aplicações de *cloud storage*

O serviço fornecido pelo Dropbox baseia-se no armazenamento dos arquivos de seus usuários em servidores com alta disponibilidade. Há dois componentes principais na arquitetura do Dropbox (Drago et al, 2012). O primeiro refere-se aos servidores de controle da aplicação, e são mantidos diretamente pela empresa. Esses servidores podem ainda ser divididos em três subgrupos: servidor de notificação, com o qual o cliente Dropbox mantém aberta uma conexão TCP para receber informações sobre mudanças realizadas em outros lugares, por outros usuários; servidores de administração de metadados, que recebem mensagens de transações de sincronização e conclusão, além de uma série de operações de armazenamento ou recuperação de dados através de servidores; e servidores de *logs* do sistema, que coletam, em tempo de execução, informações sobre

[1] http://www.google.com/trends

o cliente como *traces* ou *logs* de eventos. Já o segundo componente refere-se aos servidores de armazenamento de dados, e são hospedados pela Amazon[2]. Esses servidores reconhecem as requisições, que podem conter comandos de armazenamento e recuperação de dados, e as dividem em dois grupos para checar a quantidade de *downloads* e *uploads* em cada requisição. Em geral, nos casos dos dois componentes principais, subdomínios de *dropbox.com* permitem diferenciar as partes do serviço que executam funcionalidades específicas.

Dentre outras funcionalidades, os usuários do Dropbox podem sincronizar vários dispositivos através de uma mesma conta. Os usuários também são capazes de sincronizar arquivos seletivamente, assim como controlar os recursos de rede utilizados pelo cliente. O acesso ao serviço é fornecido através de um aplicativo com versões nativas para Windows, Linux e Mac, além de acesso através de uma interface *web*. O Dropbox também prove interfaces de programação para que desenvolvedores criem aplicações para diversos ambientes, como sistemas moveis Android.

Durante a transferência de dados entre clientes e servidores Dropbox, observa-se uma considerável redução no tamanho dos arquivos (Hu et al, 2010). De fato, todos os dados são compactados ainda na estação cliente, com o objetivo de reduzir, por consequência, o tempo total da transferência. Além disso, o cliente Dropbox compara versões de um mesmo arquivo, transferindo apenas as suas diferenças. Ainda, arquivos duplicados de um mesmo usuário são transferidos apenas uma vez. A eficácia desses três mecanismos, porém, varia de acordo com os tipos de arquivo salvo no sistema. Por fim, todos os dados trocados são criptografados, satisfazendo condições básicas de privacidade e segurança. O protocolo *HTTPS* é utilizado para acessar a maior parte dos servidores.

[2]http://aws.amazon.com

3 Metodologia de Coleta dos Dados

A caracterização apresentada dos arquivos armazenados no Dropbox é baseada em coletas realizadas a partir de voluntários. Vale ressaltar que todo o processo de coleta e tratamento dos dados foi desenvolvido, originalmente, em Drago et al (2013). A chamada para participação foi direcionada a toda comunidade, mas houve uma maior adesão de voluntários no Brasil e na Europa. Mais ainda, a grande maioria dos voluntários utilizam Dropbox com fins acadêmicos. Em outras palavras, são alunos ou pesquisadores de universidades.

Durante a coleta de informações, os voluntários executaram um programa desenvolvido pelo grupo com a finalidade de buscar as principais características dos arquivos armazenados no Dropbox. Os voluntários também respondiam a um formulário com perguntas a respeito de seu perfil. Cerca de 88% dos voluntários são homens com idade entre 20 e 30 anos. Apenas 4,5% dos voluntários declararam pagar pelo uso de *cloud storage*. A capacidade de armazenamento media declarada é de 23,4 GB.

O programa de coleta de informações foi desenvolvido em versões nativas para Windows, Mac e Linux, além de uma versão especial para a plataforma Java. O programa de coleta, inicialmente, lê as informações básicas sobre o sistema Dropbox instalado, como o diretório padrão de armazenamento da aplicação. A seguir, o programa varre recursivamente a pasta inicial do Dropbox. Todas as informações coletadas são anonimizadas de tal forma que o conteúdo e o voluntário não possam ser identificados.

Mais precisamente, são coletados metadados sobre o processo de coleta de informações, como o tempo inicial e o tempo final de captura. Para cada voluntário, é associado um identificador numérico único, permitindo que um mesmo voluntário contribua mais de uma vez com a coleta. Para cada arquivo encontrado, são armazenadas as informações sobre o tamanho, a extensão e o tipo *MIME* do arquivo. Também é capturada a data de ultima alteração do arquivo ou pasta analisada.

Cada arquivo analisado na coleta é identificado através de uma chave composta pela chave *Hash* dos 8 kB iniciais e dos 8 kB finais do arquivo. Nesse sentido, assume-se

que se dois ou mais arquivos diferentes tiverem a mesma chave composta, mesmo tamanho e mesmo tipo *MIME*, eles são réplicas. Tal abordagem simplificada de identificação de arquivo foi utilizada principalmente para reduzir o tempo de coleta (por não varrer o arquivo por completo para o cálculo da chave *Hash*).

Ao fim da coleta, os dados eram apresentados aos voluntários e enviados a um servidor centralizado. Tais dados são encaminhados para a análise somente após a aprovação explícita do participante do experimento. Nesse momento, os voluntários podiam também acessar suas estatísticas básicas e verificar, entre outras informações, a distribuição de tipo de arquivos que ele armazena em seu repositório Dropbox.

Nas análises apresentadas, são consideradas apenas dados referentes a usuários únicos. Caso um voluntário tenha enviado suas estatísticas mais de uma vez, é considerada a última postagem. O conjunto de dados avaliado nesse trabalho contem 420 coletas de 333 usuários únicos. Foram avaliados mais de 1,4TB de arquivos do Dropbox. Cerca de 45% de voluntários são da América Latina, 7% América do Norte e 40% da Europa.

Com os dados coletados, dois tipos de análise foram realizados e serão apresentados nos tópicos a seguir. Na primeira análise, ao se usar os dados de tamanho e extensão dos arquivos, tentará se traçar um perfil dos usuários em relação a quais classes de arquivos, como áudio e imagens por exemplo, eles mantém armazenados em suas contas. Na segunda análise, a partir de dados de identificação dos arquivos e de usuários, tentará se traçar possíveis conexões entre usuários, entre arquivos e entre ambos. A partir dessas conexões serão formadas redes específicas de usuários, de arquivos e entre ambos, e um conjunto de métricas de redes complexas será calculado para que se possa fazer uma análise de cada uma das redes em questão, a fim de se tirar conclusões sobre importância de nós nas redes, ligação entre usuários, dentre outras.

3.1 Métricas de Interesse

Para melhor entender os experimentos realizados, deve-se explicar, primeiramente, quais são as métricas de interesse que devem ser estudadas para a análise das redes complexas do trabalho, por que é importante que essas métricas sejam obtidas e como essas métricas puderam ser calculadas a partir do modelo de rede.

Para o cálculo das métricas foi utilizado o software *NetworkX*, que possui uma série de funcionalidades que facilita e automatiza o processo de cálculo dos dados. Dentre uma quantidade considerável de métricas de redes complexas foram escolhidas quatro que são consideradas as mais importantes medidas de centralidade de uma rede. Medidas de centralidade de uma rede são utilizadas para medir o grau de relevância dos vértices dessa rede, ou seja, qual a importância relativa de um nó em relação aos outros. As métricas calculadas foram as seguintes:

Grau: A centralidade de grau é uma medida relativa ao número de arestas incidentes que cada nó possui (ou à soma do número de arestas de entrada e saída de um nó, caso o grafo seja direcionado). Uma alta centralidade de grau está normalmente associado a uma maior chance de qualquer dado que trafega pela rede passar pelo dado nó. O grau de um nó é uma das propriedades estruturais mais básicas que avalia o número de arestas adjacentes (Jain, 1991).

Betweenness: A medida de *betweenness* de um nó, resumidamente, descreve o quão interno esse nó é na rede. O *betweenness* de um nó v_i pode ser definido como a razão de $\sigma_{sw}(i)$ por σ_{sw}, em que $\sigma_{sw}(i)$ representa o número de caminhos mínimos de v_s a v_w que passam por v_i, e σ_{sw} o número de caminhos mínimos entre v_s e v_w. Em resumo, o *betweenness* de um nó é a probabilidade desse nó estar no caminho mínimo entre dois outros nós quaisquer (Jain, 1991).

$$B(i) = \sum_{i \neq s \neq w \in V} \frac{\sigma_{sw}(i)}{\sigma_{sw}} \tag{3.1}$$

Closeness: A métrica de *closeness* pode ser definida, de um modo geral, como uma medida topológica de proximidade espacial. O *closeness* de um nó v_i é usado para avaliar o caminho mínimo médio, definido por l, entre o nó v_i e todos os outros nós v_w alcançáveis a partir de v_i tal que $(v_i,v_w) \in V$, sendo V o conjunto dos nós da rede. Em outras palavras, com essa medida é possível identificar o quão próximo um nó está de todos os outros nós da rede através das conexões estabelecidas na rede (Jain, 1991).

$$C(v_i) = \frac{|V| - 1}{\sum_{i \neq w \setminus i} l(v_i, v_w)} \tag{3.2}$$

Pagerank: A métrica de *pagerank* é considerada uma generalização do valor de centralidade de autovetor, que por sua vez determina um peso numérico para cada nó que é proporcional aos pesos numéricos dos seus nós vizinhos. O *pagerank* atribui esse peso numérico com o propósito de medir a importância de um determinado nó em seu grupo (Jain, 1991). No caso, $A = (a_{ij})$ é a matriz de adjacência do grafo (ou rede), $D = (d_{ij})$ é a matriz diagonal dos graus dos nós da rede, *1* é o vetor com todos os componentes iguais a um e α é um parâmetro a ser ajustado pelo usuário ($\alpha = 0{,}80$ neste trabalho). O valor do *pagerank* é definido pelo componente x_i do vetor que é solução da equação a seguir.

$$x = D(D - \alpha A)^{-1}1 \tag{3.3}$$

4 Arquivos na Rede

Nesta parte do trabalho é traçado o perfil dos usuários segundo as características de seus arquivos. Para isso, foi necessário retirar, de todas as informações obtidas na coleta de dados, informações mais específicas como o tamanho e a extensão de cada um dos arquivos, de cada um dos usuários que participaram como voluntários da pesquisa.

Para traçar o perfil dos usuários foi decidido que a análise sobre os seus arquivos seria feita acerca do tipo de cada um dos arquivos e não sobre a extensão deles. Entende-se como tipo de arquivo a categoria em qual um arquivo pode ser classificado, se é uma imagem ou documento de texto, por exemplo, definida de acordo com a extensão desse arquivo. Foram definidas oito categorias de arquivo, segundo as características mais encontradas e mais conhecidas de formatos de arquivos: arquivo comprimido, áudio sem compressão, áudio com compressão, documento, imagem sem compressão, imagem com compressão, vídeo e outros (com arquivos que não podiam ser classificados nas categorias anteriores). Foram definidas separações entre formatos "com e sem compressão" devido a substanciais diferenças entre formatos (e extensões) que passam ou não por algum processo de compactação. Como praticamente não se pode falar em vídeos que não sofreram algum processo de compactação, uma eventual categoria de "vídeos não compactados" não foi definida.

Assim sendo, antes que fosse feita a análise em questão, pesquisou-se sobre todas as mais de 3600 extensões encontradas no experimento e houve posterior classificação das extensões nas categorias apresentadas anteriormente. O *script* que coletou os dados dos usuários, ao analisar a extensão de um arquivo qualquer, verificava em qual categoria o arquivo poderia ser classificado. Dessa forma, as informações como quantidade e tamanho dos arquivos eram separadas por categoria, de acordo com a extensão encontrada.

Com as informações obtidas sobre os arquivos já separados por categoria, uma análise inicial já poderia ser feita. A Figura 4.1 mostra o volume de dados que cada categoria de arquivo ocupa no Dropbox (em % do total de bytes de todos os arquivos que os usuários possuem). A maioria dos arquivos manipulados pelos voluntários do

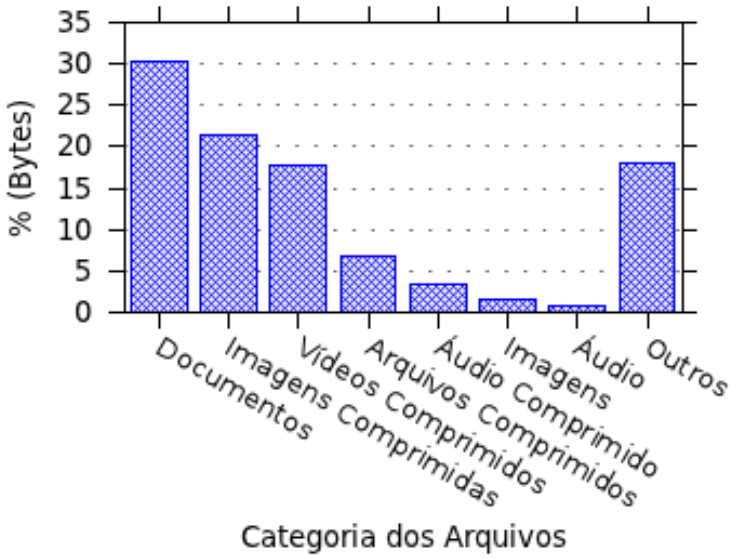

Figura 4.1: Espaço ocupado por cada categoria de arquivos na conta de um usuário Dropbox

experimento são documentos de texto em geral, como PDF, DOC, TXT ou códigos-fonte de programa, o que representa cerca de 30% do volume de dados totais. Imagens e vídeos, ambos com compressão, também apresentam considerável taxa de volume de dados, com cerca de 21% e 18%, respectivamente. Arquivos que não puderam ser classificados nessas categorias também representaram 18% do volume de dados.

Uma outra análise pode ser feita em relação à proporção de número de arquivos por categoria de classificação, através da Figura 4.2. Segundo ela, a categoria de documentos também é a maior em quantidade de arquivos, com cerca de 46% do número total de arquivos presentes nos Dropbox dos usuários, bem à frente da categoria de imagens comprimidas, que apresentou mais de 18% do número de arquivos. Um dado interessante é a quantidade de arquivos que não foram classificados nas categorias, com cerca de 32% dos arquivos totais, o que pode representar o armazenamento de arquivos com extensões próprias de projetos específicos ou particulares ou até mesmo de arquivos que não possuem extensão definida. As demais categorias tiveram quantidades inexpressivas de dados.

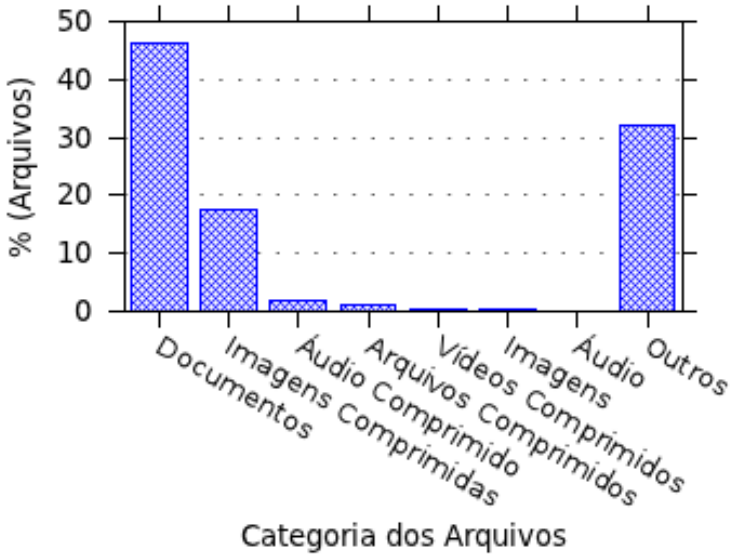

Figura 4.2: Quantidade de arquivos de acordo com a categoria

5 Topologia Social

Nesta parte do trabalho, procurou-se montar possíveis conexões entre usuários e entre extensões com o intuito de observar características que expliquem o comportamento dos usuários em relação ao uso do Dropbox. Essas conexões entre pares de usuários e de extensões formam redes que podem conter informações importantes para um melhor entendimento da estrutura topológica do experimento. Dados de identificação dos usuários e dos arquivos, como a sua extensão, por exemplo, foram utilizados para montar essas redes.

Foram montadas duas redes específicas. Uma delas trata da conexão entre os usuários do experimento e a outra trata da conexão entre as extensões dos arquivos que estão armazenados nas contas dos usuários. Ao fim, tentou-se estabelecer uma conexão entre as duas redes, através da união dos nós das duas redes, formando assim uma espécie de grafo bipartido.

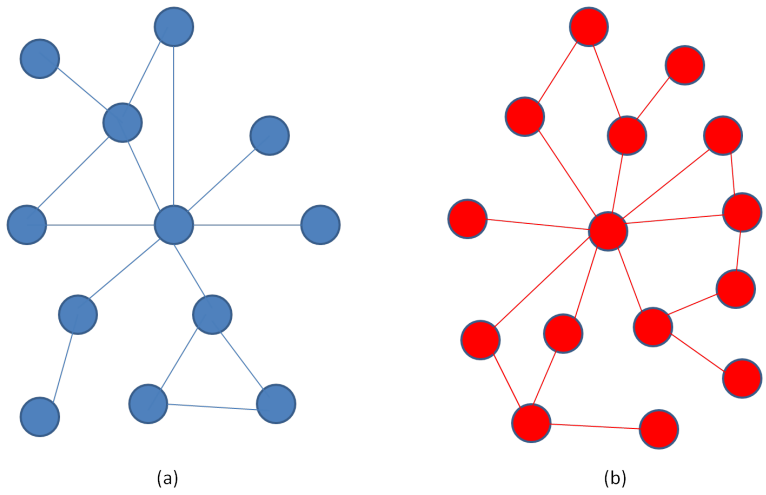

Figura 5.1: (a) Rede de usuários e (b) Rede de extensões

A rede de usuários, representada na Figura 5.1(a) é composta por nós que representam cada um dos usuários do experimento e por arestas que representam conexões entre dois usuários que possuem um mesmo arquivo em seus Dropbox. Possuir o mesmo arquivo, neste contexto, significa dizer que os usuários compartilham esse arquivo. As

arestas que ligam dois usuários possuem peso idêntico ao número de arquivos iguais que ambos têm em suas contas. Ou seja, se um determinado par de usuários compartilha três arquivos, o peso desta aresta também será três.

Dos 333 usuários que participaram do experimento, apenas 107 compõem a rede de usuários. E além disso, a rede não é totalmente conexa. Existem 5 componentes (sub-redes) no total, das quais a maior possui 83 nós. O número total de arestas é de 1665. O resultado da grande diferença entre o total de usuários do experimento e o efetivo número de usuários que são nós nesta rede pode ser explicado pelo fato de que a maioria dos participantes ou não tem arquivos compartilhados com nenhum dos outros usuários do experimento, ou usam o Dropbox apenas para armazenamento de seus próprios arquivos.

Por sua vez, a rede de extensões, representada na Figura 5.1(b) é composta por nós que representam cada uma das extensões dos arquivos encontrados nos Dropbox dos usuários. As arestas, neste caso, conectam duas extensões que são encontradas em arquivos armazenados na mesma conta Dropbox. Por exemplo, se em uma conta existe pelo menos algum arquivo de extensão PDF e outro de extensão JPG, existe uma aresta conectando os nós que representam essas extensões. O peso de uma aresta é o número de usuários que possuem pelo menos um arquivo de cada um dos tipos interligados por esta aresta. Se 70 usuários possuírem pelo menos um arquivo TXT e um arquivo GIF, o peso da aresta entre os nós GIF e TXT tem valor 70.

Foram encontradas 3.602 extensões diferentes, que foram representadas cada uma por um nó distinto, e 909.499 arestas interligando esses nós. Pode-se ressaltar que existem arquivos que não tinham extensões declaradas explicitamente, e que, para representá-los, foi criado um nó neutro. Como o número de arestas que eram ligadas a esse nó é bastante pequeno em relação ao total de arestas (menos de 0,2%), a presença ou ausência desse nó com suas arestas não impactou diretamente no experimento.

A união das redes de usuários e de extensões, definida no presente trabalho como rede usuário-arquivo e representada na Figura 5.2, forma estruturalmente uma conexão de dois conjuntos distintos de nós. Essa conexão, pode ser interpretada como a ligação entre um nó de usuário e todos os nós que representam as extensões dos arquivos que ele possui armazenados na sua conta. O peso das arestas equivale à quantidade de arquivos

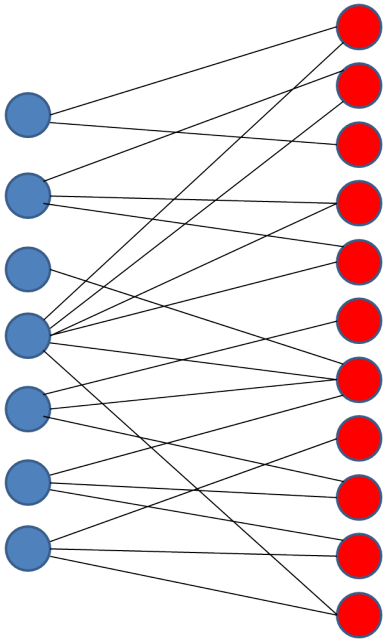

Figura 5.2: Rede usuário-arquivo

de uma determinada extensão que um usuário possui. Por exemplo, se um usuário tem 40 arquivos ZIP armazenados em sua conta, o peso da aresta que liga o nó desse usuário ao nó da extensão ZIP tem valor 40.

A rede possui 3.933 nós e 28.883 arestas, que ligam obrigatoriamente um usuário nó para um ou mais nós extensão. O número de nós na rede não é exatamente a soma dos nós das duas redes anteriores porque uma quantidade mínima de contas Dropbox estavam vazias durante a coleta dos dados.

Após a definição das redes, o conjunto de métricas apresentado no Capítulo 3 foi obtido para cada um dos casos. A ferramenta NetworkX foi utilizada para fazer os cálculos dessas métricas.

5.1 Rede de Usuários

A Figura 5.3 apresenta a distribuição cumulativa complementar dos graus dos nós da rede de usuários. Observa-se que 80% dos usuários possuem no mínimo grau 10, ou seja, estão ligados a no mínimo 10 outros usuários no experimento. Ainda, cerca de 20% dos usuários possuem no mínimo grau 50, ou seja, estão ligados a pelo menos 50 outros usuários. O maior grau obtido no experimento foi 63. Em média, interpretando o gráfico, um usuário

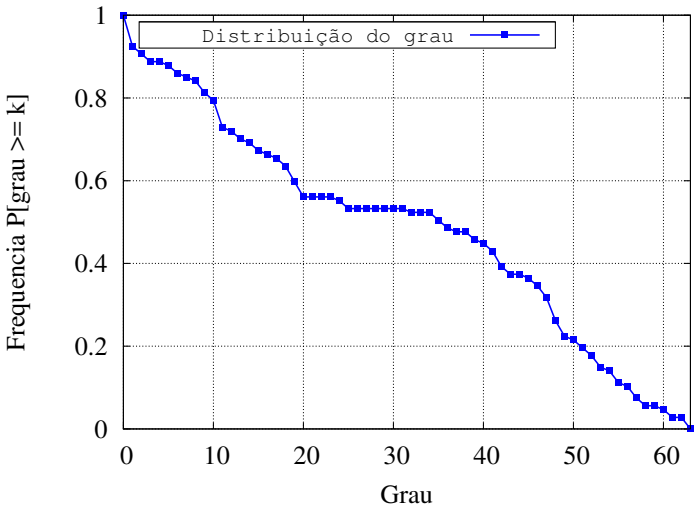

Figura 5.3: Distribuição de grau na rede de usuários.

do experimento compartilha arquivos com pelo menos 35 outros usuários.

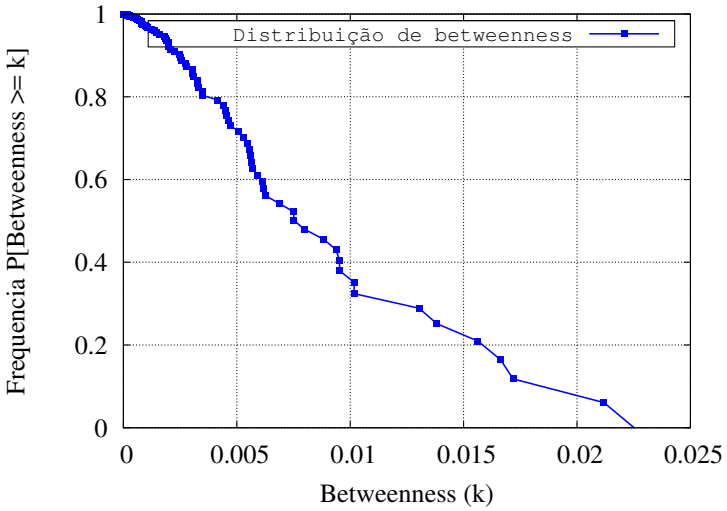

Figura 5.4: Distribuição de betweenness na rede de usuários.

A Figura 5.4 apresenta a distribuição cumulativa complementar do *betweenness* dos nós da rede de usuários. Observa-se que cerca de 30% dos nós possuem *betweenness* de, no máximo, 0,005, e que 90% deles tem no máximo *betwenness* de 0,017. Pode-se notar que a curva obtida tem a queda mais amenizada quando passa pelos 70% dos nós.

A Figura 5.5 apresenta a distribuição cumulativa complementar do *closeness* dos nós da rede de usuários. Observa-se que a grande maioria dos nós possui um *closeness* maior do que 0,4 (cerca de 90% do total dos nós) com a escala que chega no máximo a 0,63. Também pode se notar que existem pouquíssimos nós com valores de *closeness*

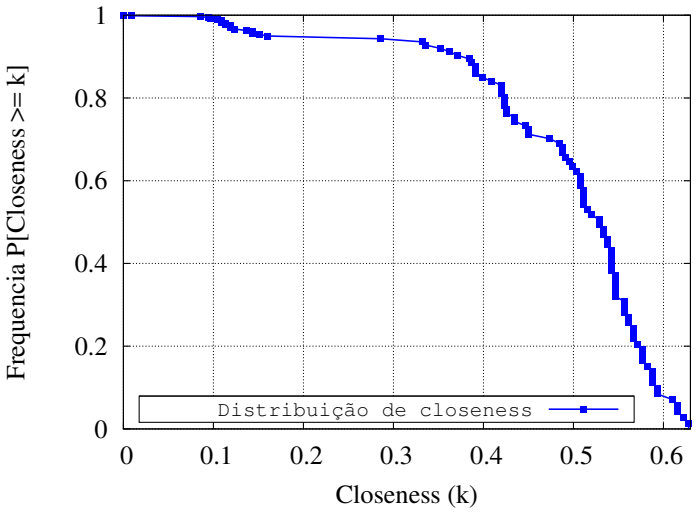

Figura 5.5: Distribuição de closeness na rede de usuários.

entre 0,15 e 0,35.

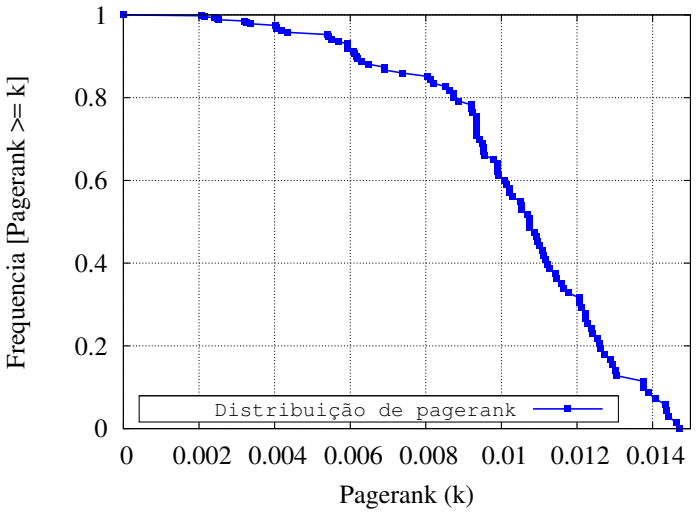

Figura 5.6: Distribuição de pagerank na rede de usuários.

A Figura 5.6 apresenta a distribuição cumulativa complementar do *pagerank* dos nós da rede de usuários. Observa-se que 20% dos nós tem, no máximo, *pagerank* de 0,009, e mais, que 40% deles tem, no máximo, *pagerank* de 0,01. Com a escala de *pagerank* variando entre 0 e 0,015, vê-se que pouquíssimos nós tem pelo menos a metade desse valor, que é próximo a 0,007.

5.2 Rede de Extensões

A Figura 5.7 apresenta a distribuição cumulativa complementar dos graus dos nós da rede de extensões. Observa-se que 80% dos nós da rede possuem, no máximo, grau igual a 600, ou seja, ligações com outros 600 nós. Apenas 10% dos nós apresentam grau igual ou superior a 1000. Ao se visualizar a curva, pode-se notar que existem nós com grau próximo a 3600, que significa dizer que existem nós que possuem ligações com todos, ou quase todos, os nós da rede. Também vale destacar que cerca de 10% dos nós tem praticamente o mesmo valor de grau (próximo a 600), que pode-se notar na queda mais acentuada da curva. Inclusive, vale ressaltar o comportamento da curva, que abaixo do valor 500 para o grau, possui um decaimento exponencial, e acima de 500 passa a seguir uma lei de potência.

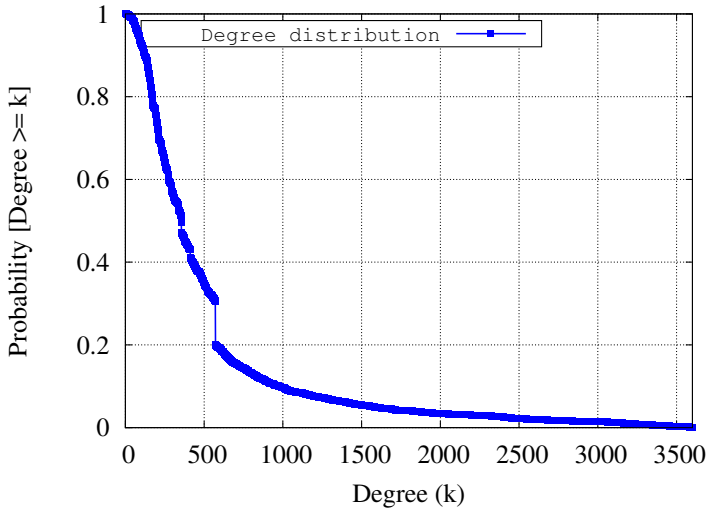

Figura 5.7: Distribuição de grau na rede de extensões.

A Figura 5.8 apresenta a distribuição cumulativa complementar do *betweenness* dos nós da rede de extensões. Observa-se que pouco mais de 40% dos nós tem *betweenness* de valor, no máximo, 0,004. Além disso, apenas 10% dos nós tem *betweenness* com valor acima de 0,01. Também pode-se notar que a curva da distribuição se assemelha a uma função linear.

A Figura 5.9 apresenta a distribuição cumulativa complementar do *closeness* dos nós da rede de extensões. Pode-se notar que todos os nós tem o valor de *closeness* acima de 0,5 e que 90% deles tem valor de *closeness* de, no máximo, 0.6. Ou seja, a

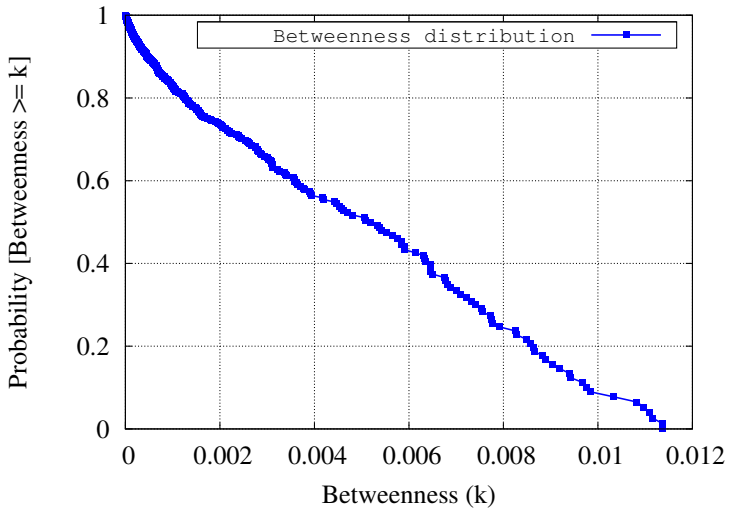

Figura 5.8: Distribuição de betweenness na rede de extensões.

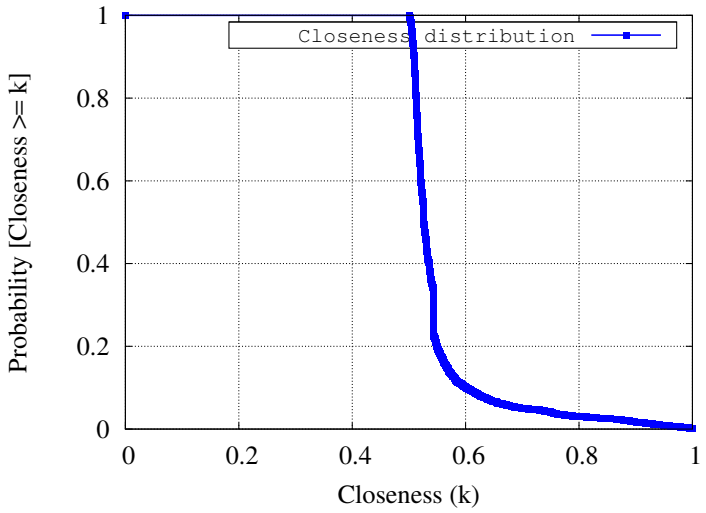

Figura 5.9: Distribuição de closeness na rede de extensões.

grande maioria dos nós (90%) apresentam valores de *closeness* contidos em uma faixa de diferença de apenas 0,1 de *closeness*.

A Figura 5.10 apresenta a distribuição cumulativa complementar do *pagerank* dos nós da rede de extensões. Observa-se que 50% dos nós tem valor de *pagerank* de até, no máximo, 0,00025. Também deve-se destacar que, aproximadamente, apenas 10% dos nós tem valor de *pagerank* maior ou igual a 0,0015.

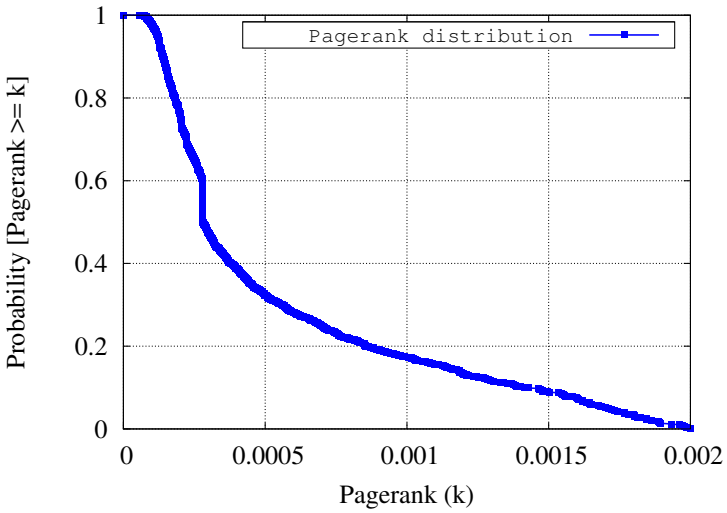

Figura 5.10: Distribuição de pagerank na rede de extensões.

5.3 Rede Usuário-Arquivo

A Figura 5.11 apresenta a distribuição cumulativa complementar dos graus dos nós da rede usuário-arquivo. Observa-se que 90% dos nós possuem grau de no máximo 50. Isso pode ser explicado porque a rede conecta nós de usuários a nós de extensões, ou na prática, conectam usuários às extensões que eles possuem armazenadas em suas contas. Ou seja, os 10% de nós que possuem grau acima de 50 são compostos, majoritariamente, por nós correspondentes a usuários que tem arquivos de muitas extensões armazenados em suas contas e, minoritariamente, por nós correspondentes a extensões que são as mais comumente encontradas em arquivos nas contas dos usuários.

A Figura 5.12 apresenta a distribuição cumulativa complementar do *betweenness* dos nós da rede usuário-arquivo. Nota-se que 80% dos nós possuem valor de *betweeness* de, no máximo, 0,03, enquanto que menos de 10% dos nós possuem valor de, no mínimo, 0,06 de *betweeness*, numa faixa que vai até 0,19.

A Figura 5.13 apresenta a distribuição cumulativa complementar do *closeness* dos nós da rede usuário-arquivo. Pode-se perceber que todos os nós possuem valor de *closeness* acima de 0,25, mas que apenas 20% deles possuem esse valor acima de 0,3. Ou seja, cerca de 80% dos nós possuem seu valor de *closeness* numa faixa de diferença de menos de 0,05 dessa métrica.

A Figura 5.14 apresenta a distribuição cumulativa complementar do *pagerank*

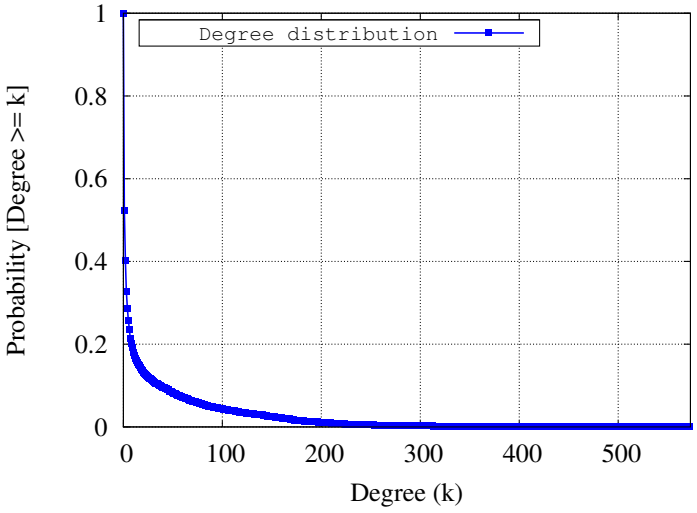

Figura 5.11: Distribuição de grau na rede usuário-arquivo.

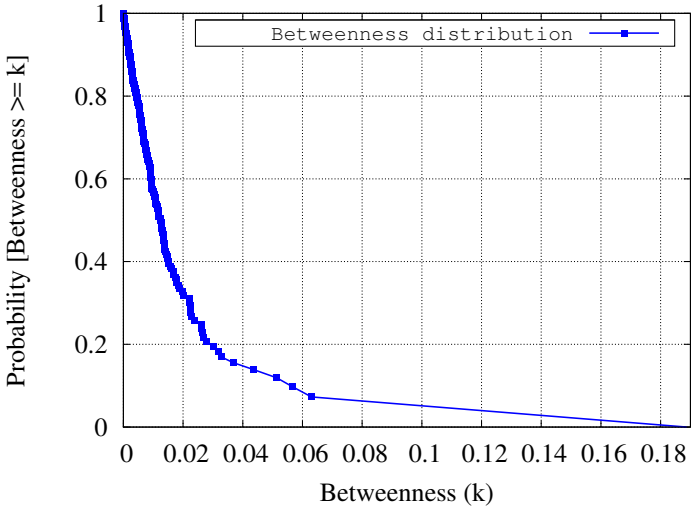

Figura 5.12: Distribuição de betweenness na rede usuário-arquivo.

dos nós da rede usuário-arquivo. Observa-se que mais de 90% dos nós possuem valor de *pagerank* de no máximo 0,005. Também deve-se ressaltar que pouquíssimos nós possuem valores de *pagerank* entre 0,005 e 0,032.

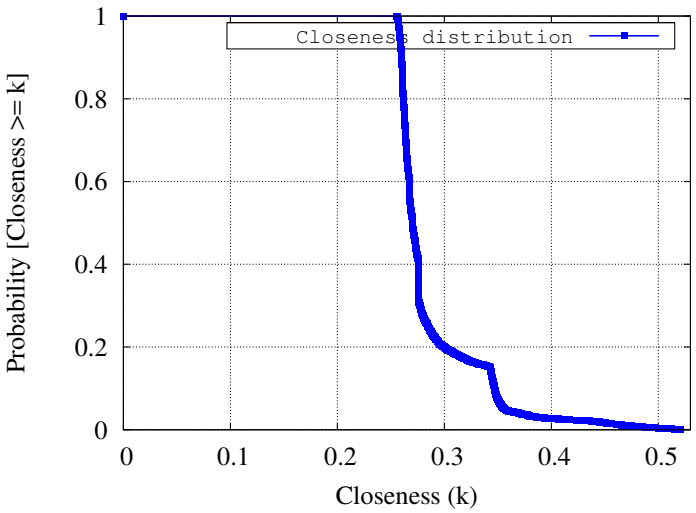

Figura 5.13: Distribuição de closeness na rede usuário-arquivo.

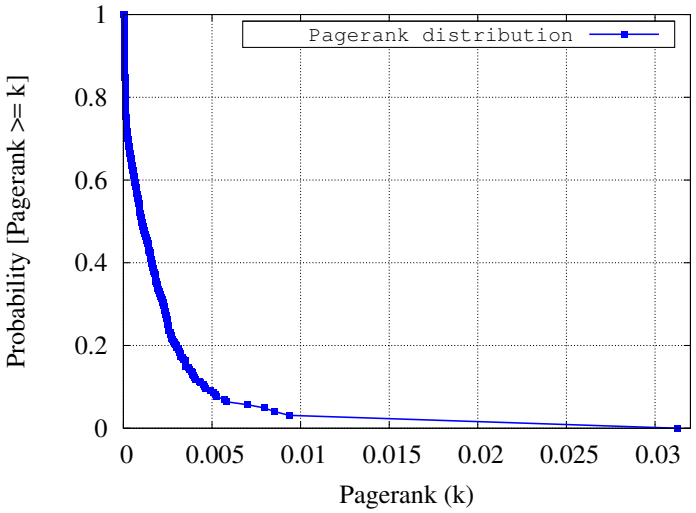

Figura 5.14: Distribuição de pagerank na rede usuário-arquivo.

6 Trabalhos Relacionados

Uma série de trabalhos vêm, recentemente, destacando o crescimento e a importância tanto do modelo de computação em nuvem (*cloud computing*), quanto dos sistemas de armazenamento em nuvem (*cloud storage*). Pela força que o tema ganhou nos últimos anos, pela extensão da área e por não existirem muitas pesquisas a respeito, várias linhas de pesquisa tem sido tomadas para que possam ser melhor entendidos o comportamento dos usuários desses sistemas, o impacto dos recursos desses sistemas em sua utilização e o uso e carga de informações trocadas pelas redes de compartilhamento formadas pelos usuários.

Dessa forma, o trabalho encontrado em Zhang et al (2010) destaca o novo paradigma para hospedagem e serviços pela Internet, que é o modelo de *cloud computing*. Ele também apresenta os conceitos principais do modelo, o estudo de sua arquitetura, o ´´estado da arte" atual, além de um melhor entendimento dos desafios da pesquisa em *cloud computing* e da identificação das direções de pesquisa para o desenvolvimento na área.

Alguns trabalhos tiveram aspectos menos gerais. Hu et al (2010) foca na avaliação e na comparação de diferentes sistemas de cloud storage, dentre eles o Dropbox, destacando as diferenças tanto em suas arquiteturas quanto nas formas de tratamento dos dados. Gracia-Tinedo el al (2013) também realizou comparações entre sistemas de *cloud storage*, através de um estudo de medições dos sistemas, que consistiu em analisar importantes aspectos para caracterizar suas QoS, como velocidade de transferência e taxa de falha. Já o trabalho encontrado em Li et al (2002) apresenta a implementação de um comparador de desempenho e de custo entre provedores em nuvem, com o qual se pode obter algumas medidas de interesse.

Outros trabalhos focam no desempenho dos sistemas. É o caso encontrado em Bergen et al (2011), que trata do desempenho dos sistemas, com enfoque na análise das redes envolvidas no processo, e o trabalho de Iosup et al (2002), que investigou a variabilidade de desempenho de serviços em nuvem, através da análise de dados de

provedores de infraestrutura.

Alguns outros trabalhos trataram de questões de segurança e privacidade. Halevi et al (2011) foca na identificação de ataques que utilizam duplicação da conta do cliente, com casos encontrados até para o sistema Dropbox. Mulazzani et al (2011) apresenta as fragilidades do software cliente do Dropbox e possíveis vetores de ataque contra usuários. Já Ion et al (2011) estuda as atitudes de privacidade dos usuários.

Com o enfoque mais próximo do sistema Dropbox e de suas características, tem-se o trabalho encontrado em Drago et al (2012), que apresenta uma caracterização desse sistema segundo a análise de seu protocolo proprietário, além de características de seu tráfego, mas sem fornecer informações sobre os arquivos armazenados no serviço. Já o trabalho encontrado em Drago et al (2013) foi o primeiro a fazer uma caracterização preliminar do sistema de armazenamento Dropbox, com enfoque nas características dos arquivos e suas relações com a quantidade e volume dos dados armazenados, além de destacar o impacto desses arquivos na utilização dos recursos do sistema.

O presente trabalho segue a linha deste último, com a utilização do mesmo conjunto de dados enviado por usuários voluntários, mas o estende através de uma nova caracterização, além da formação de redes complexas entre elementos do experimento e da análise de métricas dessas redes, o que possibilita uma outra visão do sistema em si.

7 Conclusões

Com a expansão do modelo de computação em nuvem e sua crescente aplicação em negócios, indústria e academia, uma melhor caracterização dos sistemas de armazenamento em nuvem passou a ser necessária. Portanto, o presente trabalho teve o objetivo de caracterizar de forma inicial os arquivos e as interações entre usuários do sistema Dropbox, o serviço de *cloud storage* mais popular do mercado.

Os resultados encontrados na análise dos arquivos dos 333 usuários que participaram do experimento mostram que a maioria dos arquivos armazenados em suas contas Dropbox são documentos de texto em geral, editáveis ou não, como PDFs, DOCs, TXTs ou códigos-fonte de programas, que representam cerca de 46% da quantidade total de arquivos. Além disso, a categoria de documentos também é a que possui o maior volume de arquivos armazenados no Dropbox, com cerca de 30% do volume total de dados em bytes. Também foi encontrado um grande volume de bytes de arquivos de imagem comprimida, como JPGs e GIFs, o que representa cerca de 21% do total de arquivos. Ou seja, cerca de 50% dos dados armazenados e trafegados pela rede Dropbox analisada são compostos apenas por documentos (muitos não editáveis) e imagens comprimidas, o que pode sugerir que a implementação dos mecanismos de controle de versão do Dropbox não causa impacto significativo na economia de tráfego gerado na aplicação.

Com relação às características das redes complexas formadas pela interação entre os usuários, pode-se destacar que apenas cerca de 32% dos usuários do experimento estão, de certa forma, compartilhando arquivos entre si. Essa grande diferença entre o total de usuários do experimento e o número de usuários que efetivamente compartilham arquivos pode ser explicada pelo fato de que a maioria dos participantes ou não tem arquivos compartilhados com nenhum dos outros usuários do experimento, ou utilizam o Dropbox apenas para o armazenamento de seus arquivos particulares. Vale ressaltar que uma considerável parte dos usuários, cerca de 20% deles, estão ligados a no mínimo outros 50 usuários, e que, em média, um usuário está ligado a cerca de 35 usuários, através do compartilhamento de arquivos.

Na rede formada pelas extensões dos arquivos analisados, pode-se encontrar nós que representam extensões que se conectam a todas as outras extensões encontradas no experimento, o que significa dizer que as primeiras podem ser encontradas em arquivos de todas as contas dos usuários Dropbox. Além disso, cerca de 80% dos nós dessa rede possuem ligações com outros 600 nós, e cerca de 90% possuem valores de *closeness* que variam entre 0,5 e 0,6. Para a rede usuário-arquivo, nota-se que 90% dos nós possuem, no máximo, grau igual a 50, valor esse que representa a grande quantidade de extensões que alguns usuários possuem armazenada em sua conta ou representa uma extensão que é encontrada em bastantes contas Dropbox. Também pode-se destacar que mais de 90% dos nós desta rede possuem um valor muito pequeno de *pagerank*, de no máximo 0,005.

Por fim, espera-se que a partir do presente trabalho, seja possível fazer uma correlação entre as diferentes métricas analisadas. Por exemplo, poderia se verificar quais os elementos das redes possuem bons valores para essas métricas e se é possível traçar um perfil para esses elementos a partir dessa análise. Para concluir, de forma mais geral, espera-se que essa caracterização possa contribuir para a melhoria da eficiência do Dropbox e para o desenvolvimento de outras aplicações e ferramentas.

Referências Bibliográficas

Bergen, A.; Coady, Y. ; McGeer, R. **Client Bandwidth: The Forgotten Metric of Online Storage Providers**. In: Proceedings of the 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, PacRim'2011, 2011.

Drago, I.; Mellia, M.; Munafo, M.; Sperotto, A.; Sadre, R. ; Pras, A. **Inside Dropbox: Understanding Personal Cloud Storage Services**. In: Proceedings of the 12th ACM Internet Measurement Conference (IMC'12), p. 481–494, 2012.

Drago, I.; Vieira, A. B. ; Silva, A. P. C. **Caracterização dos Arquivos Armazenados no Dropbox**. In: Simposio Brasileiro de Redes de Computadores e Sistemas Distrbuidos (SBRC 2013), 2013.

Gracia-Tinedo, R.; Sánchez-Artigas, M.; Moreno-Martínez, A.; Cotes-González, C. ; García-López, P. **Actively Measuring Personal Cloud Storage**. In: IEEE CLOUD'13, 2013.

Halevi, S.; Harnik, D.; Pinkas, B. ; Shulman-Peleg, A. **Proofs of Ownership in Remote Storage Systems**. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS'11, p. 491–500, 2011.

Hu, W.; Yang, T. ; Matthews, J. N. **The good, the Bad and the Ugly of Consumer Cloud Storage**. In: ACM SIGOPS Operating Systems Review, p. 110–115, 2010.

Ion, I.; Sachdeva, N.; Kumaraguru, P. ; Čapkun, S. **Home is safer than the cloud!: Privacy concerns for consumer cloud storage**. In: Proceedings of the Seventh Symposium on Usable Privacy and Security, p. 13, 2011.

Iosup, A.; Yigitbasi, N. ; Epema, D. **On the Performance Variability of Production Cloud Services**. In: CCGRID'11, p. 104–113, 2011.

Jain, R. **The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling**. Nova Iorque, EUA, Abril 1991. John Wiley & Sons.

Li, A.; Yang, X.; Kandula, S. ; Zhang, M. **Cloudcmp: Comparing Public Cloud Providers**. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC'10), p. 1–14, 2010.

Mulazzani, M.; Schrittwieser, S.; Leithner, M.; Huber, M. ; Weippl, E. **Dark Clouds on the Horizon: Using Cloud Storage as Attack Vector and Online Slack Space**. In: Proceedings of the 20th USENIX Conference on Security, SEC'11, 2011.

Zhang, Q.; Cheng, L. ; Boutaba, R. Cloud Computing: State-of-the-Art and Research Challenges. **Journal of Internet Services and Applications**, v.1, p. 7–18, 2010.