Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Eficiência de download em enxames BitTorrent

Jaindson Valentim Santana

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Computação

Nazareno Andrade
(Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

S232e Santana, Jaindson Valentim.
Eficiência de download em enxames BitTorrent / Jaindson Valentim Santana. - Campina Grande, 2011.
51f.: il.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.
Orientador: Prof. Nazareno Andrade.
Referências.

1. P2P. 2. BitTorrent. 3. Eficiência. 4. Download. 5. Enxames. I. Título.

CDU 004.75 (043)

Resumo

O BitTorrent (BT) é o sistema de compartilhamento de arquivos mais utilizado atualmente. Neste sistema, cada arquivo é distribuído por um grupo de usuários BT chamado de enxame. Naturalmente, usuários e enxames possuem diferentes características, e essas características afetam o desempenho da distribuição do arquivo. Por exemplo, os usuários podem ter diferentes larguras de banda enquanto os enxames podem apresentar diferentes quantidades de usuários. Este trabalho utiliza dados sobre o comportamento de cerca de 80 mil usuários em 13 mil enxames para investigar que características influenciam de forma significativa a qualidade de serviço experimentada pelos usuários de enxames BitTorrent.

Abstract

Nowadays, BitTorrent (BT) is the most popular file-sharing system. In this system, each file is distributed to a group of BT users called torrent. As we would expect, users and torrents have different characteristics, and these characteristics have some effects on the distribution of the file. For example, the users may have different bandwidths, while the torrents may have different population size. This work is based on the behaviour of about 80 thousand users sharing files in about 13 thousand torrents. All these users and torrents are used in order to investigate which characteristics have a significant effect on the quality of service experienced by the BT users.

Agradecimentos

Primeiramente gostaria de agradecer a toda minha família, meus pais Gutemberg e Nilva e meus irmãos Jaqueline e Josias, que sempre me apoiaram desde o início da minha carreira em Ciência da Computação.

Ao meu orientador Nazareno e a Fubica pelos ensinamentos transmitidos, apoio e incentivo.

Aos amigos que fiz em Campina Grande durante a graduação e pós-graduação, presentes tanto nos momentos de descontração, quanto nos tensos.

Aos amigos do LSD, pelo aprendizado compartilhado e companhia.

À Universidade Federal de Campina Grande, ao Centro de Engenharia Elétrica e Informática, ao Departamento de Sistemas e Computação, ao Programa de Pós-Graduação em Ciência da Computação da UFCG e seus respectivos professores e funcionários que direta ou indiretamente contribuíram na minha formação.

Por fim, agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contexto

O surgimento do Naspter (1999) [28], primeiro sistema de compartilhamento de arquivos entre-pares (P2P) a atingir escala mundial, chamou bastante atenção dos usuários da Internet. Estima-se que no início de 2001 o sistema atingiu um pico de 26.4 milhões de usuários [21]. Logo após seu surgimento, apareceram outros sistemas com o mesmo propósito [7] [28] [9], sempre apresentando melhorias em relação aos anteriores. Em conjunto, essas aplicações popularizaram o modelo de compartilhamento entre-pares. Algumas medições realizadas em provedores de serviço de internet (ISPs) em 2002 mostraram que boa parte do tráfego de dados na Internet era devido ao uso destas aplicações [8] [15].

Dentre as aplicações que surgiram, uma das que mais obteve aceitação do público foi o BitTorrent (BT) [9], um sistema de compartilhamento P2P simples e robusto cuja principal característica consiste em ser eficaz na distribuição de conteúdo a um baixo custo. Esta característica pode ser associada ao fato da oferta de serviço necessária para atender a demanda gerada pelos usuários ser distribuída entre todos eles. A grande aceitação do BT pode ser identificada em alguns trabalhos que realizaram estudos sobre seu tráfego [27] [23] [17] [16].

Para compartilhar conteúdo, os usuários do BT se agrupam e formam um enxame. Cada enxame é responsável por compartilhar um conteúdo específico. Nestes enxames os usuários podem entrar e sair quando bem entenderem e possuem diferentes características (largura de banda, disposição para compartilhar o conteúdo com os demais usuários, etc.), o que pode proporcionar um sistema bastante heterogêneo.

Além de se agruparem nos próprios enxames, uma prática comum adotada por alguns usuários é a participação nas chamadas comunidades BT [6] [12] [13] [11] [1]. Nestas comunidades os usuários costumam estabelecer um fórum na internet através do qual passam a se comunicar e descobrir novos conteúdos, além de manterem algumas estatísticas de contribuição e participação dos usuários na comunidade.

Desta forma, o ecossistema BT em si é concretizado por meio de diversos enxames distribuídos em múltiplas comunidades, com os mais diversos tipos de conteúdo e usuários.

Dado o caráter colaborativo do sistema e dada a heterogeneidade de seus participantes, diversos fatores influenciam simultaneamente na qualidade de serviço experimentada pelos usuários.

Dada a popularidade destes sistemas, é importante entender que características dos enxames influenciam o seu desempenho e, portanto, a satisfação de seus usuários. Parte deste entendimento pode ser obtido através de simulações ou abordagens analíticas [24] [25] [5] [19], no entanto, é importante estendê-los ou validá-los através de observações empíricas.

1.2 Definição do Problema

Hoje não há evidências fortes sobre quais características presentes nos enxames influenciam na eficiência de download experimentada por seus usuários. Este trabalho se propõe a investigar esta lacuna.

No BT é possível mensurar a qualidade de serviço oferecida aos usuários, usando a velocidade de download experimentada por eles. No entanto, observar diretamente a velocidade de download pode ser impreciso, uma vez que ela é limitada pela capacidade de download do usuário. Observar a utilização da largura de banda de download seria mais promissor, uma vez que ela leva em consideração esta limitação. Neste trabalho, a utilização média da largura de banda de download do usuário é chamada de eficiência de download.

1.3 Objetivo Geral e Específicos

O objetivo geral deste trabalho consiste em analisar que características dos enxames e de seus usuários influenciam na qualidade de serviço oferecida aos usuários pelos enxames, fazendo

isto de um ponto de vista empírico.

Os objetivos específicos são:

- estabelecer uma métrica que represente a qualidade de serviço oferecida aos usuários de um enxame;
- levantar características dos enxames e dos usuários que sejam relevantes para a análise;
- analisar quais características têm influência significativa na qualidade de serviço.

1.4 Resultados e Contribuições

Neste trabalho foi realizada uma análise da eficiência de download oferecida pelos enxames aos seus usuários e investigado quais características dos enxames e dos usuários exercem uma influência significativa na qualidade de serviço.

A partir dos resultados foi possível encontrar evidências empíricas de que para que os enxames ofereçam uma alta eficiência de download aos seus usuários é preciso que eles distribuam arquivos maiores, incentivem o surto de popularidade (i.e. *flashcrowds*) e aumentem o número de seeders tentando distribuí-los ao longo do tempo.

1.5 Estrutura da Dissertação

O conteúdo seguinte desta dissertação está organizado da seguinte forma:

- **Capítulo 2 - Fundamentação Teórica.** Neste capítulo são apresentados os principais conceitos do BT relevantes no contexto deste trabalho e a definição de eficiência de download.
- **Capítulo 3 - Trabalhos relacionados.** Neste capítulo são apresentados os trabalhos relacionados a BT com ênfase no estudo da velocidade de download dos usuários nos enxames e suas características.
- **Capítulo 4 - Metodologia.** Neste capítulo é apresentada a metodologia utilizada para desenvolver este trabalho, apresentando: o traço utilizado na investigação, a aborda-

gem utilizada para estimar alguns dados, as características a serem extraídas do traço e a análise multivariada usada.

- **Capítulo 5 - Análise da eficiência de download.** Neste capítulo são apresentados os resultados obtidos da análise multivariada, destacando quais caracteristicas tem maior influência na eficiência de download dos enxames.

- **Capítulo 6 - Conclusão e Trabalhos Futuros.** Nesta capítulo são apresentadas as conclusões, as contribuições desta dissertação para a comunidade científica em torno do BT, as limitações dos resultados e os trabalhos que ainda se encontram em aberto.

Capítulo 2

Fundamentação Teórica

Neste capítulo são abordados os principais conceitos do BT no contexto deste trabalho e a definição de eficiência de download.

2.1 Funcionamento BT

Nesta seção são explicadas a nomenclatura e o funcionamento do BT, conhecimentos importantes para entender a definição de eficiência de download. Mais detalhes sobre o funcionamento do BT podem ser obtidos no trabalho de Marlom et al. [18].

Para simplificar a explicação, segue uma definição dos termos utilizados ao longo do trabalho [19]:

- **Comunidade BT.** Conjunto de pessoas que se agrupam com objetivo de compartilhar conteúdos de interesse em comum, além das normas e mecanismos que determinam regras para o compartilhamento de conteúdo. Normalmente utilizam um site com fórum como ponto de encontro para divulgação das normas e novos conteúdos. Geralmente estabelecem uma identificação forte para cada usuário, sendo possível então manter um histórico da contribuição de cada usuário na comunidade.

- **Usuário de uma comunidade BT.** Uma pessoa que participa da comunidade BT.

- **Peer.** Representa um usuário da comunidade fazendo parte de um enxame.

- **Enxame.** Conjunto de todos os peers interessados em compartilhar um mesmo conteúdo. Desta forma os peers que participam de um mesmo enxame compartilham o conteúdo entre si.

- **Sessão de um peer.** Período de tempo que o peer participa de forma ativa no enxame. Ou seja, ele está distribuindo e/ou obtendo o conteúdo neste período. O peer pode ter diversas sessões ao longo de sua participação em um enxame.

- **Tracker.** Entidade centralizada responsável pelo serviço de descoberta de peers e coleta de estatísticas relativas a participação e estado dos peers no enxame.

- **Pedaços e Blocos.** O arquivo sendo compartilhado é particionado em pedaços, que por sua vez são particionadas novamente em blocos. O bloco é a menor unidade de troca entre os peers.

- **Torrent.** Consiste num arquivo de metadado contendo todas as informações necessárias para obter o arquivo: quantidade de pedaços, identificação única para todos os pedaços (usado na verificação de integridade dos pedaços), e endereço para contactar o tracker.

- **Seeder.** Peer que possui uma cópia completa do arquivo. Neste caso ele permanece no enxame fazendo apenas upload de blocos.

- **Leecher.** Peer que possui apenas uma cópia parcial do arquivo e participa do enxame fazendo upload e download de blocos.

Quando um usuário quer compartilhar um arquivo ele deve primeiramente criar o arquivo torrent contendo os metadados do arquivo em questão. Em seguida, deve disponibilizá-lo para que outros usuários possam descobrí-lo e a partir disso entrar no enxame. Inicialmente, apenas quem publicou o arquivo possui a cópia completa. Desta forma, apenas ele envia parte do arquivo aos demais. À medida que os novos peers obtêm partes do arquivo eles passam a compartilhar também com os demais, dividindo assim a carga de distribuição entre todos os usuários do enxame. Todo este processo pode ocorrer mais de uma vez ao mesmo tempo, o que implica ser possível um usuário participar em mais de um enxame simultaneamente.

2.2 Eficiência de Download

A eficiência de download de um usuário BT avalia a qualidade de serviço experimentada por ele, enquanto a eficiência de download de um enxame avalia a qualidade de serviço oferecida aos usuários que participaram deste enxame. A eficiência de download leva em consideração a capacidade de download dos usuários, permitindo assim que a qualidade de serviço obtida / oferecida por diferentes usuários / enxames possam ser comparadas de forma mais fidedigna. Utilizar diretamente a velocidade de download como qualidade de serviço pode ser errôneo, uma vez que a velocidade de download é limitada pela capacidade de download do usuário. Dois usuários com diferentes capacidades de download experimentando uma mesma velocidade de download não devem ser considerados como obtendo a mesma eficiência, uma vez que o usuário de maior capacidade tem uma utilização menor de sua banda.

O cálculo da eficiência de download de um usuário é realizado durante o período em que ele participava como leecher nos enxames. A eficiência consiste na razão da quantidade de dados obtidos pelo usuário em todos enxames que participou e a quantidade de dados que poderia ter sido obtida caso o usuário experimentasse sua velocidade máxima de download (capacidade de download do usuário). De outra forma, a eficiência calcula a utilização da banda do usuário ao longo do período. Quanto maior for a utilização, melhor será a eficiência. Desta forma, note que a eficiência assume valores no intervalo $[0, 1]$.

A Figura 2.1 ilustra a variação da velocidade de download de um usuário hipotético ao longo do tempo. A partir dela é possível calcular a eficiência deste usuário com base nas áreas A_1 e A_2:

$$E_u = \frac{A_1}{A_1 + A_2}$$

Definindo formalmente temos que a eficiência do usuário é:

$$E_u = \frac{\int_a^b v(t)\,\mathrm{d}t}{C \cdot \Delta t}$$

Onde:

- $\Delta t = b - a$;

- $v(t)$ é a velocidade de download no instante t;
- e C é a capacidade de download do usuário.

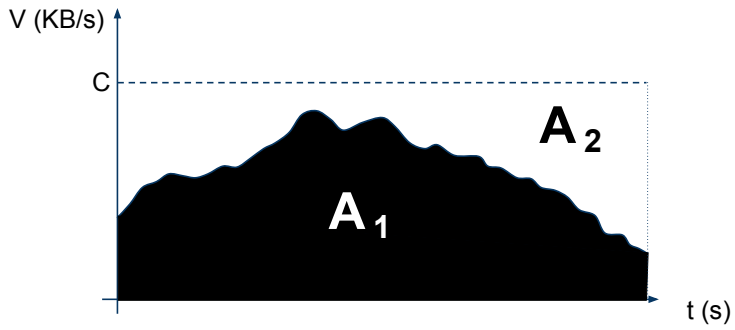


Figura 2.1: Variação da velocidade de download de um usuário ao longo do tempo

O cálculo da eficiência de download de um enxame considera apenas os usuários que participaram dele. A eficiência consiste na razão da quantidade de dados obtidos neste enxame por todos os usuários que participaram dele e a quantidade de dados que todos estes usuários poderiam ter obtido caso estivessem experimentando suas respectivas velocidade máxima de download (capacidade de download do usuário). Caso um usuário esteja participando em mais de um enxame ao mesmo tempo como leecher, em vez de sua capacidade total de download é considerado como se ele estivesse utilizando uma fatia da capacidade inversamente proporcional a quantidade de enxames que ele está participando simultaneamente como leecher. Definindo formalmente temos que a eficiência de um enxame é:

$$E_e = \frac{\sum\limits_{u \in U(e)} \int_a^b v(u,e,t)\,\mathrm{d}t}{\sum\limits_{u \in U(e)} \int_a^b \frac{C(u)}{N(u,t)}\,\mathrm{d}t}$$

Onde:

- $U(e)$ é o conjunto de todos os usuários que participaram no enxame e;

- $v(u, e, t)$ é a velocidade de download do usuário u no enxame e no instante t;
- $C(u)$ é a capacidade de download do usuário u;
- e $N(u, t)$ é a quantidade de enxames em que o usuário u estava online como leecher no instante t.

Capítulo 3

Trabalhos relacionados

Neste capítulo são abordados trabalhos que também fizeram uma análise da velocidade de download dos usuários de comunidades BT, destacando no que eles diferem do presente trabalho. Em poucas palavras, há trabalhos que exploram menos caracteristicas e enxames, enquanto outros não tentam explicar os resultados a partir das características observadas.

3.1 Análise descritiva

Bellissimo et al. [4] e Meulpolder et al. [22] expõem a velocidade de download em múltiplos enxames, mas não associam o comportamento observado a características dos enxames. Em outras palavras, realizaram uma análise essencialmente descritiva.

No trabalho de Bellissimo et al. foram analisados dados de algumas comunidades BT populares durante o período de 4 meses (final do ano 2003, abrangendo alguns dias de 2004), através do qual foi possível observar cerca de 845 mil sessões e 4.3 mil enxames. Seu estudo tinha como objetivo avaliar a eficiência do BT em disseminar arquivos muito grandes, visando a partir disso considerar sua adoção num sistema de distribuição de grandes arquivos e conjuntos de dados para a comunidade científica. Para isso analisaram o tamanho dos arquivos sendo compartilhados, a popularidade destes arquivos, o tamanho das sessões dos usuários, a velocidade de transferência, e o impacto na distribuição dos arquivos devido os surtos de popularidade. A partir da análise concluíram que o tamanho dos arquivos compartilhados é na ordem dos gigabytes e que a popularidade dos arquivos tem uma distribuição semelhante a de outros sistemas de compartilhamento de arquivos. Além disto, a grande

maioria dos usuários precisam de múltiplas sessões para obter todo o arquivo e possuem conexões assimétricas, com a largura de banda de download superior a de upload. Quanto ao impacto devido os surtos de popularidades puderam perceber que eles não comprometiam o sistema, uma vez que a grande maioria dos usuários conseguiam compartilhar com outros usuários instantes depois de entrarem no sistema. A partir destes resultados puderam então concluir que o BT é bastante eficiente na distribuição de arquivos grandes e que ele se adequava as suas necessidades uma vez que algumas modificações fossem implantadas no sistema.

No trabalho de Meulpolder et al. foram analisadas 5 comunidades, utilizando para isso dados obtidos nos quatro últimos meses de 2009. Embora mais atual, conseguiu observar uma quantidade bem inferior de enxames (444 ao todo), abrangendo cerca de 508 mil peers. Seu estudo tinha como objetivo oferecer à comunidade científica uma análise das propriedades presentes nas comunidades de compartilhamento de conteúdo de forma mais detalhada do que as apresentadas nos estudos realizados até então, enfatizando principalmente as diferenças entre as comunidades públicas e privadas. Para isto analisaram a velocidade de download experimentada pelos usuários, a conectividade destes usuários, a razão entre seeders e leechers, a duração do tempo de seeding, e algumas estatísticas relacionadas com a demanda de recursos imposta pelos usuários do sistema. A partir da análise concluíram que a velocidade de download experimentada pelos usuários das comunidades privadas é significativamente superior a experimentada pelos usuários das comunidades públicas. Além disto, as comunidades privadas apresentaram melhor índice de conectividade dos usuários, razão entre seeders e leechers, e duração no tempo de seeding. A partir destes resultados levantaram então a hipótese de que quando mecanismos de *effective ratio enforcement*[1] são utilizados, o mecanismo de *tit-for-tat* natural do BT não tem mais influência significativa.

[1] Mecanismo muito utilizado nas comunidades privadas que força os usuários a manterem um limiar mínimo na razão entre a quantidade de dados baixados e enviados para a comunidade.

3.2 Análise univariada

Andrade et al. [2] e Locher et al. [20] realizaram trabalhos com constatações semelhantes. Em ambos os trabalhos foi observado o comportamento de usuários freeriders[2] e concluído que, em cenários onde a quantidade de seeders é grande, a velocidade de download experimentada pelos freeriders se equipara ou até mesmo atinge valor maior que a velocidade obtida pelos que contribuem com o sistema. Nestes trabalhos tentou-se explicar a velocidade de download obtida pelos usuários considerando apenas a característica deles serem freeriders ou não. Realizaram, portanto, uma análise univariada.

No trabalho de Andrade et al. foram observados dados de 5 comunidades, abrangendo cerca de 27 mil enxames e 787 mil usuários. Seu estudo tinha como objetivo investigar o impacto do protocolo BT no comportamento colaborativo dos usuários e o impacto do tipo de conteúdo e das políticas utilizadas em algumas comunidades no nível de cooperação observado nas comunidades BT. Para isto estabeleceram algumas métricas de cooperação que seriam então utilizadas para analisar os fatores que as influenciavam. A partir da análise realizada encontraram evidências de que a forma como o protocolo BT funciona favorece o comportamento colaborativo dos usuários em relação a outros sistemas de compartilhamento de arquivos entre-pares, embora haja cenários em que a velocidade de download experimentada por usuários freeriders seja equivalente ou até mesmo maior que a dos usuários que colaboram com o sistema.

No trabalho de Locher et al. foi utilizada uma única comunidade e poucos enxames em cada análise, provavelmente por ela ser baseada em experimentação. Seu estudo tinha como objetivo mostrar que é possível fazer download de arquivos no BT de forma rápida mesmo não contribuindo para o sistema (i.e. sem realizar upload), participando tanto em enxames de comunidades públicas quanto privadas. Para isto modificaram um cliente BT para nunca enviar dados para outros usuários e fizeram algumas experimentações com este cliente modificado e um cliente normal num ambiente controlado. Nestas experimentações, ambos os clientes eram colocados no mesmo enxame ao mesmo tempo e no final anotado quanto tempo cada um levou para fazer download de todo o arquivo. A partir da análise realizada foi

[2]No contexto de P2P, freerider é aquele usuário que nada contribui para o sistema ou que não contribui uma parcela significativa comparado com o quanto ele obteve do sistema.

possível observar que em cenários onde a quantidade de seeders é relativamente maior que a de leechers, o cliente modificado obtém os dados numa velocidade equivalente aquela experimentada pelo cliente normal, mas sem o custo de realizar qualquer upload. Enquanto nos cenários onde a quantidade de leechers era relativamente superior, a velocidade de download experimentada pelo cliente modificado era bem menor, embora ainda conseguisse obter o arquivo por completo eventualmente.

3.3 Análise multivariada

Rasti et al. [26] observaram a utilização da capacidade de download (uma estimativa dela) e o coeficiente de variação da velocidade de download dos usuários de três enxames, e buscaram explicar seu comportamento com base em algumas características no nível de usuário e enxame. Como resultado foi visto que nenhuma das características conseguia explicar bem o comportamento observado. Este é o trabalho que mais se aproxima do presente trabalho, divergindo dele principalmente por ter utilizado uma quantidade muito menor de enxames.

No trabalho de Rasti et al. foram analisados dados de 3 comunidades (duas referentes a compartilhamento de distribuições Linux e uma outra de um jogo 3D). Embora tivessem dados de cerca de 4 mil enxames, utilizaram apenas 3 nas suas análises (um representante de cada comunidade).

3.4 Análise multivariada em múltiplos enxames

Dentre os trabalhos observados no levantamento do estado da arte, nenhum deles se utilizou de uma quantidade significativa de enxames e/ou utilizou uma quantidade significativa de características para tentar explicar o comportamento observado. Portanto, o presente trabalho se diferencia dos demais por utilizar:

- um conjunto de enxames significativamente grande;
- múltiplas características para explicar a eficiência observada;
- uma abordagem para mitigar o erro ao estimar a capacidade de download dos usuários.

3.5 Características observadas

Nas seções anteriores foram discutidos alguns trabalhos cujo foco estava relacionado de certa forma com a velocidade de download experimentada pelos usuários. No entanto, é possível apontar alguns outros que também se relacionam com o presente trabalho por terem estudados algumas características dos enxames e dos usuários, sendo que algumas delas (destacado em negrito) foram utilizadas para tentar explicar o comportamento observado. A Tabela 3.1 apresenta algumas características e trabalhos relacionados que as analisaram.

Tabela 3.1: Listagem de algumas características analisadas nos trabalhos relacionados e no presente trabalho (destacado em negrito).

Característica	**Trabalho(s) relacionado(s)**
Velocidade de download	Bellissimo et al. [4], Meulpolder et al. [22], Andrade et al. [2], Locher et al. [20], Rasti et al. [26]
Velocidade de upload	Rasti et al. [26], Andrade et al. [3], Bellissimo et al. [4]
População total	Rasti et al. [26], Bellissimo et al. [4], Andrade et al. [2]
Taxa de entrada e saída	Rasti et al. [26]
Disponibilidade de conteúdo	Rasti et al. [26]
Tempo de seeding	Andrade et al. [3], Meulpolder et al. [22]
Razão seeder/leecher	Meulpolder et al. [22], Andrade et al. [2]
Conectividade	Meulpolder et al. [22]
Tamanho do arquivo	Bellissimo et al. [4], Andrade et al. [2]
Tamanho das sessões	Bellissimo et al. [4]
Duração do enxame	Andrade et al. [2]

Capítulo 4

Metodologia

Neste capítulo é discutida a metodologia utilizada no desenvolvimento deste trabalho. Em poucas palavras, o trabalho se baseia na análise de dados extraídos de traços de comunidades, onde a partir destes dados são estimadas as capacidades de download dos usuários e observadas características dos enxames e dos usuários que são, a seguir, utilizadas como variáveis independentes na análise multivariada dos fatores que afetam a eficiência dos enxames.

4.1 Traços de comunidades

Traços de comunidades consistem em dados que retratam o histórico das interações entre os usuários e enxames de uma comunidade. A partir deles é possível descrever o comportamento dos usuários e enxames. Os dados presentes nos traços são limitadores quanto a que informações podem ser analisadas em trabalhos baseados em traços, uma vez que eles são a única fonte de onde as informações podem ser extraidas. Neste trabalho, é utilizado um traço de uso de uma comunidade BT chamada Bitsoup [6].

Bitsoup é uma comunidade de usuários BT que ainda se encontra em atividade. Esta comunidade não possui um foco particular, compartilhando conteúdo de diversas naturezas. Isto favorece a heterogeneidade dos usuários, o que provavelmente pode ser refletida nas características dos usuários e enxames. A Tabela 4.1 mostra a população total de usuários, a quantidade de enxames e o período de observação da comunidade Bitsoup.

Este traço do Bitsoup foi criado a partir da coleta periódica do estado de toda a comunidade, informação esta disponível no site da comunidade. O espaço de tempo entre as coletas

Tabela 4.1: Características da comunidade

Comunidade	**#Usuários**	**#Enxames**	**Período**
Bitsoup [6]	84026	13741	29/04/2007 - 02/07/2007 (64 dias)

foi de uma hora, o que gerou em torno de 40 gigabytes de dados úteis, contemplando aproximadamente 8.3 milhões de sessões de usuários. A partir destes dados é possível então observar a dinâmica da comunidade ao longo do tempo.

Os dados obtidos de cada usuário da comunidade a cada coleta permitem saber:

- em quais enxames o usuário estava online no dado instante de tempo da coleta;
- qual a quantidade de download e upload realizado até então em cada um destes enxames;
- se o usuário era leecher ou seeder nestes enxames;
- tamanho do arquivo sendo compartilhado nos enxames.

Por se tratar de uma grande quantidade de dados, foi necessário paralelizar o processamento utilizando um cluster de aproximadamente 6 máquinas, cada uma com 16 núcleos.

4.2 Estimando a capacidade de download

Conforme visto no Capítulo 2, para estudar a velocidade de download dos usuários foi estabelecida a métrica eficiência de download. Esta métrica leva em conta a quantidade de tempo que cada usuário ficou online enquanto leecher. Dado este tempo, é analisada a razão entre a quantidade de dados de fato obtidos pelo usuário e a quantidade de dados que poderia ter sido obtida caso o usuário experimentasse sua velocidade máxima de download permitida por sua capacidade de download.

O cálculo da eficiência utiliza portanto as seguintes informações: capacidade de download dos usuários, velocidade de download dos usuários nas sessões que ele participou como leecher e a duração destas sessões.

No entanto, note que a capacidade de download não é um dado disponível a partir do traço (ver 4.1). É necessário, portanto, estimar a capacidade de download a partir dos dados disponíveis.

Para estimar a capacidade de download de um usuário é calculado primeiramente um conjunto composto pela velocidade de download deste usuário entre cada par de coletas consecutivas em que ele esteve online, levando em consideração o download realizado por ele em todos os enxames que estava online no período entre as duas coletas. Em seguida, é considerado como capacidade de download do usuário a maior velocidade de download nesse conjunto. Definindo formalmente temos que a estimativa da capacidade de download do usuário é:

$$\hat{C}(u) = max\left(\frac{D(t_{i+1}) - D(t_i)}{\Delta t}\right)$$

Onde:

- $\hat{C}(u)$ é a estimativa da capacidade de download do usuário u;
- $D(t)$ é a quantidade de download acumulado que o usuário realizou no instante t.
- t_i e t_{i+1} representam os instantes de tempo em que foram realizadas duas coletas consecutivas no traço
- $\Delta t = t_{i+1} - t_i$

Esta forma de estimar a capacidade de download do usuário consegue uma informação precisa caso o usuário tenha experimentado utilização máxima da sua capacidade de download durante o período em que houve a coleta de dados da comunidade. Note, portanto, que é necessário observar um mesmo usuário uma quantidade mínima de vezes, de modo que seja possível fazer uma estimativa acurada da sua capacidade de download a partir destas observações.

É impossível antever um valor para a quantidade mínima de observações necessária para estimar a capacidade de download de um usuário com acurácia, e, consequentemente, impossível determinar a precisão de uma escolha. Por outro lado, é possível aproximar o valor

mínimo de observações para um valor adequado analisando os usuários que podem ser observados por um tempo considerado muito longo. Esta quantidade mínima de vezes pode ser determinada da seguinte forma:

1. filtrar usuários com no mínimo O_n observações, onde n é a quantidade de observações considerada longa o suficiente para que a capacidade real do usuário tenha alta probabilidade de ter sido observada;

2. estimar a capacidade de download destes usuários segundo o método já descrito;

3. identificar qual a quantidade de observações mínima O_m para que pelo menos uma fração c ($0 < c < 1$) dos usuários analisados consigam experimentar uma velocidade de download de no mínimo 80% da sua capacidade de download estimada.

Dado este procedimento, a quantidade de observações mínima adequada será aquela em que $O_m < O_n$. Isto implica que não é necessário analisar mais que O_m observações para estimar a capacidade de download com uma boa acurácia. A quantidade mínima de observações seria então O_m.

Este procedimento foi realizado para alguns valores de O_n e foi identificado que um valor de O_m adequado seria 492. Na análise que obteve este resultado o valor de c era igual a 0.7 e o de O_n igual a 600. A partir deste resultado foram utilizados nas análise apenas os usuários com no mínimo 492 observações, o que corresponde a aproximadamente 20% de todos os usuários observados na comunidade. Como consideramos apenas um subconjunto de todos os usuários na análise da eficiência, foram considerados apenas os enxames em que pelo menos um destes usuários tiveram participação. Da mesma forma, no cálculo da eficiência dos enxames são considerados apenas os dados compartilhados por estes usuários, embora no cálculo das características destes enxames sejam considerados todos os usuários que participaram dele.

4.3 Variáveis independentes

Para tentar explicar o comportamento observado das eficiências dos enxames serão analisadas algumas características dos enxames e dos usuários. Estas características serão utilizadas

como variáveis independentes na análise multivariada (regressão linear múltipla). Nesta seção são descritas as características observadas.

São elas:

- Tamanho do arquivo: consiste no tamanho do arquivo sendo compartilhado no enxame;
- População total: consiste na quantidade de usuários distintos que estiveram presentes no enxame;
- População média ponderada no tempo: consiste na quantidade média ponderada de usuários presentes no enxame ao longo do tempo;
- População média de seeders ponderada no tempo: consiste na quantidade média ponderada de seeders presentes no enxame ao longo do tempo;
- Mediana do tempo de seeding: consiste na mediana do tempo que os seeders passam online no enxame. Como seeders apenas enviam dados em um enxame, espera-se que sua presença tenha influência na eficiência de download de seus pares;
- Mediana do intervalo entre chegadas: consiste na mediana do intervalo entre chegadas dos usuários no enxame, com a chegada sendo definida como a primeira aparição de um usuário em um enxame. Como novos usuários em um enxame chegam sem peças para trocar com os demais, uma alta taxa de chegada pode ter efeito negativo na eficiência de um enxame.
- Razão média de seeder / leecher ponderada no tempo: consiste na quantidade média ponderada da razão do número de seeders pelo número de leechers nos enxames ao longo do tempo. Como seeders apenas enviam dados em um enxame, espera-se que uma proporção maior de seeders possa ter um efeito positivo na eficiência de um enxame.

Esse conjunto de características ao mesmo tempo visa abranger as características disponíveis no traço que têm relação intuitiva ou pesquisada em trabalhos relacionados com a eficiência da distribuição de conteúdo em enxames.

Capítulo 5

Análise da eficiência de download

Neste capítulo são discutidos os resultados da análise realizada. Desde uma análise descritiva da estimativa da capacidade de download dos usuários, das caracteristicas dos enxames, da eficiência de download dos enxames e dos usuários; até a análise multivariada envolvendo as características e eficiência dos enxames.

5.1 Estimativa da capacidade de download dos usuários

Inicialmente será mostrado o resultado obtido para a capacidade de download dos usuários, uma vez que ela é necessária para o cálculo das eficiências. A Figura 5.1 mostra o gráfico da função distribuição acumulada (FDA) da estimativa da capacidade de download dos usuários da comunidade Bitsoup. A partir dela é possível observar que 75% dos usuários possuem capacidade de download máxima de aproximadamente 700KB/s enquanto aproximadamente 1% possui capacidade zero. Capacidade zero significa que os usuários não foram capazes de realizar nenhum download enquanto estiveram online.

5.2 Eficiência de download dos enxames e dos usuários

A Figura 5.2 mostra o gráfico da FDA da eficiência de download dos usuários e dos enxames. Os usuários com capacidade de download zero foram removidos da análise, uma vez que neste caso não faz sentido calcular a eficiência. No geral, a partir da figura pode-se concluir que a eficiência dos usuários é tipicamente melhor que a eficiência dos enxames. Enquanto

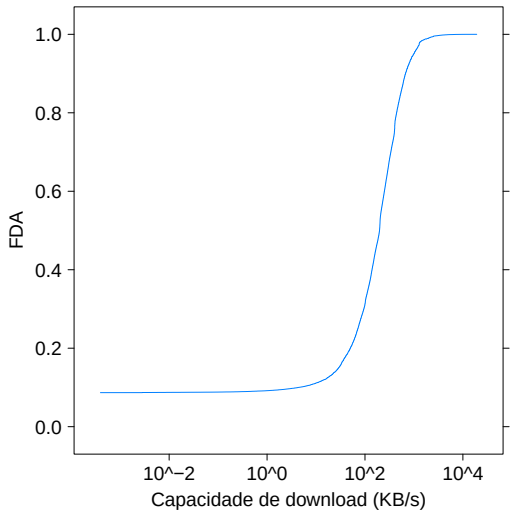


Figura 5.1: Estimativa da capacidade de download dos usuários da comunidade Bitsoup

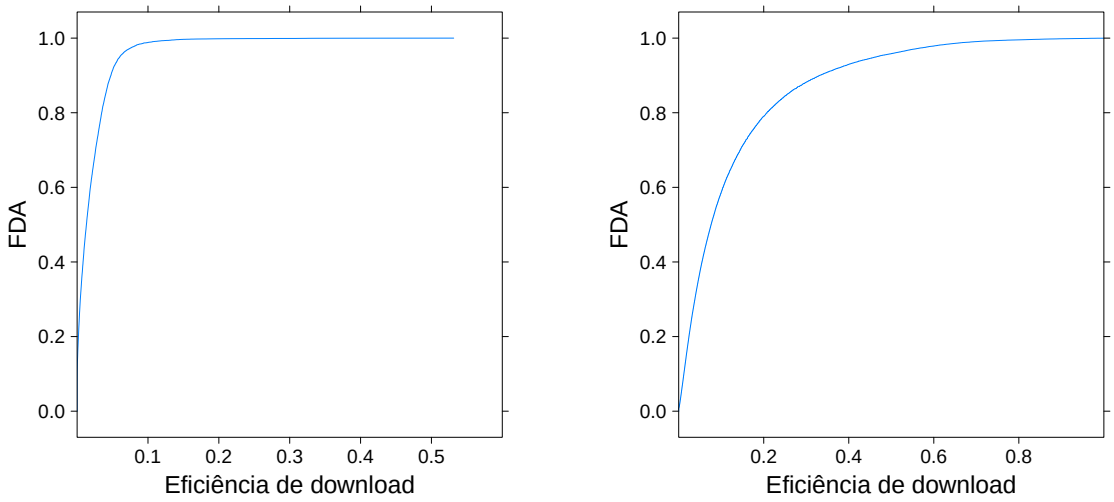
a mediana da eficiência dos usuários é de $7,6\%$, a mediana da eficiência dos enxames é de $1,3\%$. No entanto, as duas métrica possuem um valor significativamente baixo. No caso da eficiência dos usuários isto implica que a velocidade média (contabilizando todo período que permaneceu online como leecher) da metade dos usuários é igual ou inferior a $7,6\%$ de sua capacidade de download. Enquanto no caso da eficiência dos enxames isto implica que a velocidade média dos usuários que participaram em metade destes enxames é igual ou inferior a $1,3\%$ de sua capacidade de download.

5.3 Características dos enxames e dos usuários

Nesta seção é mostrada a análise descritiva das características que são consideradas na análise como fatores que possivelmente influenciam a eficiência dos enxames.

5.3.1 Tamanho do arquivo

A Figura 5.3a mostra o gráfico da FDA do tamanho dos arquivos sendo compartilhado nos enxames. A partir dele é possível observar que a maioria dos enxames (75%) distribuem arquivos relativamente grandes, com mais de 100 MB. Além disto, uma parcela destes enxames (25% deles) distribuem arquivos bem maiores, com tamanho acima de 1.4 GB.



(a) Enxames

(b) Usuários

Figura 5.2: Gráfico da eficiência dos usuários e dos enxames da comunidade Bitsoup

5.3.2 População total

A Figura 5.3b mostra o gráfico da FDA do tamanho da população dos enxames. A partir dele é possível observar que a maioria dos enxames (cerca de 80%) possuem um quantidade muito pequena de usuários distintos (no máximo 150) quando leva-se em consideração a quantidade total de usuários que fazem parte da análise.

5.3.3 População média ponderada no tempo

A Figura 5.3c mostra o gráfico da FDA do tamanho da população média ponderada no tempo dos enxames. A partir dela é possível observar que, da mesma forma que na análise da população total, a população total ponderada no tempo é muito pequena na maioria dos enxames (cerca de 80%), chegando a no máximo 15 usuários distintos conectados simultaneamente.

5.3.4 População média de seeders ponderada no tempo

A Figura 5.3d mostra o gráfico da FDA do tamanho da população média de seeders ponderada no tempo dos enxames. A partir dela é possível observar que, considerando o tamanho da população total ponderada no tempo, a população média de seeders ponderada no tempo é relativamente grande (aproximadamente 9 seeders distintos) na maioria dos enxames (75%

deles).

5.3.5 Mediana do tempo de seeding

A Figura 5.3e mostra o gráfico da FDA da mediana do tempo de seeding nos enxames (em segundos). A partir dela é possível observar que 75% dos enxames chegam a ter no máximo aproximadamente $100h$ como tempo mediano de seeding (aproximadamente 4 dias), enquanto 25% deles tem no máximo aproximadamente $40h$ (quase 2 dias).

5.3.6 Mediana do intervalo entre chegadas

A Figura 5.3f mostra o gráfico da FDA da mediana do intervalo entre chegadas nos enxames (em segundos). A partir dela é possível observar que 75% dos enxames chegam a ter no máximo aproximadamente $3h$ como tempo mediano de intervalo entre chegada dos peers nos enxames, enquanto 25% deles tem no máximo aproximadamente $15min$. Ou seja, o intervalo entre chegadas dos usuários é curto na maioria dos enxames.

5.3.7 Razão média de seeders / leechers ponderada no tempo

A Figura 5.3g mostra o gráfico da FDA da média da razão seeders / leechers nos enxames ponderada no tempo. A partir dela é possível observar que 75% dos enxames chegam a ter no máximo aproximadamente 5.4 como razão média de seeders / leechers ponderada no tempo, enquanto 25% deles tem no máximo aproximadamente 1.6. Ou seja, na grande maioria dos enxames há pelo menos um seeder para cada leecher ao longo do tempo.

5.4 Influência das características na eficiência de download

Nesta seção é mostrada a análise multivariada realizada entre as características dos enxames e dos usuários e suas respectivas eficiências.

5.4.1 Modelos

Neste modelo consideramos as seguintes variáveis independentes:

- tamanho do arquivo (*tam*);
- população total (*pop*);
- população total ponderada no tempo (*pop.pond*);
- população de seeders ponderada no tempo (*pop.seeders.pond*);
- mediana do intervalo entre chegadas (*median.arrival.time*);
- mediana do tempo de seeding (*median.seeding.time*);
- razão média de seeder / leecher ponderada no tempo (*seeder.leecher.ratio.pond*).

E a seguinte variável dependente:

- eficiência do enxame (*eficiencia.do.enxame*).

Além disso, consideramos o modelo em dois cenários:

(a) Modelo simples: sem transformações nos dados;

(b) Modelo logarítmico: aplicando a transformada logarítmica na variável dependente e em todas as variáveis independentes, exceto *median.seeding.time* e *seeder.leecher.ratio.pond*, uma vez que a transformação não seria válida em parte dos dados.

Para realizar a regressão linear múltipla é preciso realizar alguns testes nas variáveis utilizadas no modelo para certificar que os pré-requisitos estatísticos exigidos são obedecidos. Uma vez que alguns destes testes devem ser realizados após a regressão, todos os testes foram realizados apenas depois da análise do coeficiente de determinação (R^2) e no modelo cujo coeficiente obteve um melhor resultado.

A Tabela 5.1 mostra o valor do R^2 ajustado para a regressão linear multivariada aplicada a todas as variáveis e em ambos os casos. A partir dela é possível observar que a transformada logarítmica melhora significativamente o quanto o modelo explica a eficiência dos enxames.

O valor do *p-value* obtido no Teste T para todas as variáveis do modelo logarítmico foi menor que $2 * 10^{-16}$. Isto indica que os coeficientes associados a todas as variáveis assumem valores significativamente diferentes de zero. Portanto, todas as variáveis independentes têm influência significativa na dependente.

Tabela 5.1: R^2 ajustado

Modelo	R^2
simples	0.1936
logarítmico	0.7398

Como o modelo logarítmico apresentou um melhor resultado para o R^2, apenas ele foi analisado em mais detalhes.

5.4.2 Pré-requisitos da regressão linear multivariada

Para realizar a regressão linear multivariada é preciso que as variáveis envolvidas no modelo obedeçam alguns pré-requisitos [14]:

- Normalidade;
- Homocedasticidade;
- Linearidade;
- Ausência de erros correlacionados;

Nesta seção serão mostrados os testes realizados nas variáveis do modelo logarítmico[1].

Normalidade

Refere-se ao formato da distribuição dos dados de cada variável e sua correspondência a distribuição normal. Para verificar esta conformidade podem ser utilizadas tanto análises gráficas quanto testes de hipótese.

[1]Embora a transformação logarítmica tenha sido aplicada em algumas variáveis do modelo, conforme já mencionado anteriormente, ao se referir a estas variáveis nos gráficos e ao longo do texto em nenhum momento será mostrado explicitamente o uso da transformada logarítmica (i.e. em vez de mencionar variável *log(A)* será dito apenas variável *A*).

Uma forma de análise gráfica que pode ser feita é observar o histograma da variável em conjunto com sua FDP e a FDP de uma distribuição normal. Para afirmar que a variável assume uma distribuição normal é preciso que sua FDP seja muito semelhante a da distribuição normal.

A Figura 5.4 apresenta os gráficos do histograma de cada uma das variáveis envolvidas no modelo. Em cada gráfico são apresentados também a FDP da distribuição normal (em azul) e a FDP da variável (em preto). A partir dos gráficos, apenas a variável *pop* assume distribuição normal (5.4b).

Uma outra análise gráfica que pode ser utilizada para refutar ou reforçar os resultados obtidos é observar o gráfico quantil-quantil normal. Este gráfico possibilita a comparação de duas distribuições. Neste caso, uma das distribuições seria a normal e a outra a da variável em questão. Para afirmar que a variável assume uma distribuição normal é preciso que os pontos no gráfico referentes a distribuição da variável se aproximem da reta diagonal traçada no gráfico, reta esta que representa a distribuição normal.

A Figura 5.5 apresenta os gráficos quantil-quantil normal de cada uma das variáveis envolvidas no modelo. A partir deles é possível observar novamente que apenas a variável *pop* assume claramente distribuição normal (5.5b).

Conforme já mencionado, além das análises gráficas também podem ser utilizados alguns testes estatísticos para verificar se as variáveis assumem distribuição normal. Alguns testes foram utilizados, tais como o teste de Jarque-Bera [10] e o teste de Shapiro-Wilk [14].

O teste de Jarque-Bera (JB) avalia a hipótese nula de que determinada variável tem uma distribuição normal com determinada média e variância (valores estimados a partir da amostra), contra a hipótese alternativa de que ela não tem distribuição normal. Este teste se utiliza das medidas de assimetria (do inglês, *skewness*) e curtose (do inglês, *kurtosis*) da amostra, medidas estas que descrevem o formato da distribuição. A estatística JB segue uma distribuição qui-quadrado com dois graus de liberdade. A hipótese nula de normalidade é rejeitada se $JB > X^2_{\alpha,2}$, onde $X^2_{\alpha,2}$ é o quantil de nível $1 - \alpha$ da distribuição X^2 com dois graus de liberdade.

A Tabela 5.2 mostra o teste de Jarque-Bera aplicado a cada uma das variáveis envolvidas no modelo. Para um nível de significância de 5% o valor de $X^2_{\alpha,2}$ é de aproximadamente 6. Com base nos resultados do teste não se pode dizer que as variáveis assumem uma distribui-

ção normal (rejeição da hipótese nula).

Tabela 5.2: Teste de Jarque-Bera

Variável	JB
tam	2711.657
pop	61.1353
pop.pond	4028.2
pop.seeders.pond	2903.502
median.arrival.time	70.7983
median.seeding.time	12689567
seeder.leecher.ratio.pond	551548.3
eficiecia.da.torrente	17498.56

O teste de Shapiro-Wilk (W) avalia a hipótese nula de que determinada variável tem uma distribuição normal, contra a hipótese alternativa de que ela não tem distribuição normal. Se o p-valor está abaixo do limiar determinado (neste estudo α igual a 0.05), então a hipótese nula é rejeitada e a hipótese alternativa prevalece.

A Tabela 5.3 mostra o teste de Shapiro-Wilk aplicado a cada uma das variáveis envolvidas no modelo. Com base nos resultados do teste não se pode dizer que as variáveis assumem uma distribuição normal (rejeição da hipótese nula).

Mesmo as análises gráficas e os testes estatísticos realizados apontando para a não normalidade das variáveis, o fato do tamanho da amostra ser relativamente grande (acima de 200) faz com que os efeitos da não normalidade sejam desprezíveis [14]. Uma alternativa ainda seria tentar utilizar transformações nos dados para fazer com que as distribuições das variáveis satisfaçam os testes de normalidade.

Homocedasticidade

Refere-se a suposição de que a variável dependente exibe um mesmo nível de variância ao longo dos valores assumidos pelas variáveis independentes. Para verificar a homocedasticidade podem ser utilizadas tanto análises gráficas quanto testes de hipótese.

Tabela 5.3: Teste de Shapiro-Wilk

Variável	W	**p-valor**
tam	0.9396	$< 2.2E-16$
pop	0.9958	$1.363E-08$
pop.pond	0.9387	$< 2.2E-16$
pop.seeders.pond	0.954	$< 2.2E-16$
median.arrival.time	0.9637	$< 2.2E-16$
median.seeding.time	0.541	$< 2.2E-16$
seeder.leecher.ratio.pond	0.673	$< 2.2E-16$
eficiecia.da.torrente	0.8398	$< 2.2E-16$

Uma forma de análise gráfica que pode ser feita é observar o gráfico dos resíduos padronizados pelos valores preditos padronizados. Caso haja homocedasticidade os dados estarão dispostos uniformemente ao longo da reta horizontal que passa na origem do eixo das ordenadas.

A Figura 5.6 apresenta o gráfico dos resíduos padronizados pelos valores preditos padronizados. A partir dele é possível observar que a dispersão dos resíduos aparenta se concentrar mais abaixo da linha horizontal e no centro do gráfico. Mas como interpretar se esta distorção é suficiente para afirmar que há heteroscedasticidade? Para isto é possível utilizar alguns testes estatísticos, como o teste de Levene, para verificar individualmente cada variável independente em conjunto com a dependente e verificar qual delas apresenta heteroscedasticidade.

O teste de Levene pode ser utilizado para verificar se duas amostras possuem a mesma variância. Ele avalia a hipótese nula de que as duas amostras possuem a mesma variância, contra a hipótese alternativa de que há diferença entre as duas variâncias. Se o p-valor obtido no teste estiver abaixo do limiar determinado (neste estudo α igual a 0.05), então a hipótese nula é rejeitada e a hipótese alternativa prevalece.

A Tabela 5.4 mostra o teste de Levene aplicado a cada uma das variáveis envolvidas no modelo. Com base nos resultados do teste é possível concluir que apenas a variável *pop*

apresenta heteroscedasticidade.

Tabela 5.4: Teste de Levene

Variável	**p-valor**
tam	1
pop	0.005171
pop.pond	1
pop.seeders.pond	1
median.arrival.time	1
median.seeding.time	1
seeder.leecher.ratio.pond	1
eficiecia.da.torrente	1

Linearidade

Refere-se a relação entre as variáveis dependentes e independentes, que deve apresentar um formato linear para que não comprometa o poder da relação encontrada na regressão. A linearidade pode ser identificada observando os gráficos de dispersão entre a variável dependente e cada uma das independentes, bem como uma análise do gráfico dos resíduos.

O gráfico de dispersão apresenta uma relação linear entre as variáveis caso os pontos estejam dispostos numa reta imaginária. Já o gráfico dos resíduos indica que há uma relação não-linear que não foi explicada pelo modelo caso os pontos não estejam dispostos uniformemente acima e abaixo da reta horizontal que passa no ponto zero do eixo das ordenadas.

A Figura 5.7 apresenta os gráficos de dispersão entre a eficiência dos enxames e cada uma de suas características. A partir dele é possível observar que apenas o gráfico de dispersão da variável *tam* (5.7a) apresenta uma relação linear com a variável *eficiencia.do.enxame*. Uma medida que poderia ser tomada é aplicar transformações nos dados ou então tentar representar diretamente no modelo a relação não-linear existente.

A Figura 5.6 apresenta o gráfico dos resíduos padronizados pelos valores preditos padronizados. A partir dele é possível observar que os pontos localizados na parte mais externa

do gráfico se concentram mais abaixo da reta horizontal, enquanto os pontos no centro se concentram mais um pouco acima da reta horizontal, sugerindo assim que há uma relação não-linear (curvilinear) que não está sendo explicada pelo modelo. Novamente, transformações nos dados ou a adição de novos termos no modelo para tentar representar essa relação não-linear são alternativas que podem ser usadas. Além disto, o uso de outros métodos, como o de regressão não-linear, também podem ser utilizados para tentar representar melhor esta relação.

Ausência de erros correlacionados

Refere-se a presença de erros de predição que estejam correlacionados com outros erros de predição. Este tipo de erro é mais comum quando as observações da amostra são obtidas em diferentes coletas ou quando os dados são obtidos periodicamente ao longo do tempo. Embora os dados brutos utilizados neste trabalho tenham sido obtidos periodicamente, as características de cada enxame agregam informações de diferentes coletas, de modo que os efeitos que porventura tenham ocorrido numa determinada coleta acabam sendo incluídos nas características.

5.4.3 Modelo logarítmico

A Tabela 5.5 mostra o valor dos coeficientes padronizados para o modelo logarítmico, ou seja, os valores dos coeficientes numa mesma escala com média 0 e desvio padrão 1, o que permite comparar diretamente os valores dos coeficientes e identificar que variável tem maior influência na regressão. As variáveis na tabela estão dispostas em ordem decrescente de influência. A partir dela é possível identificar que a variável *tam* é a que mais influencia no modelo e que ela tem uma influência positiva na eficiência do enxame. Isso implica que o tamanho do arquivo é um fator determinante na eficiência dos enxames observados. Isto pode ocorrer devido ao fato de que nos enxames maiores, usuários precisam de mais tempo para obter o arquivo. Com isso, e devido ao mecanismo de otimização de download *tit-for-tat* do BT, é possível que os usuários tenham maior probabilidade de encontrar os parceiros de maior capacidade que estão dispostos a retribuir seu upload no enxame, favorecendo então o compartilhamento entre si e melhorando as suas velocidades de download.

Logo em seguida é a variável *median.arrival.time* a que mais influencia no modelo (influência negativa). Um intervalo curto entre chegadas dos usuários favorece o encontro entre usuários com capacidade de download/upload semelhante, influenciando então de forma semelhante à variável *tam*. Em seguida é a variável *pop*, exercendo uma influência positiva. Neste caso, um indício de que quanto maior a quantidade de usuários, melhor a eficiência de download.

A seguir aparece a variável *seeder.leecher.ratio.pond*. O fato dela ter uma influência negativa no modelo vai de encontro com o esperado. Este resultado implica que quanto maior for a proporção de seeder / leecher, menor é a eficiência. Maiores investigações são necessárias para entender este comportamento.

Em seguida é a variável *pop.pond*, exercendo uma influência negativa no modelo. Por um lado, este é um resultado inesperado, pois ele implica que quanto menor for a população ao longo do tempo, maior é a eficiência. No entanto, é possível que a população ao longo do tempo seja menor em consequência da eficiência ser grande. Se os usuários terminam o download mais cedo, eles acabam saindo do enxame também mais cedo, o que implicaria numa diminuição da população ao longo do tempo.

A variável *pop.seeders.pond* vem a seguir com uma influência positiva no modelo. Isto implica que quanto maior for a população de seeders no enxame ao longo do tempo, maior é a eficiência. Como os seeders exercem apenas o papel de doadores, não consumindo recursos do sistema, eles ajudam os demais usuários a melhorar sua eficiência.

Por último vem a variável *median.seeding.time*, com uma influência desprezível no modelo. Isto implica que a eficiência não é influenciada pelo tempo mediano de seeding dos seeders. Desde que os seeders estejam presentes (ver análise da variável *pop.seeders.pond*), não importa quanto tempo cada seeder em particular passe no enxame.

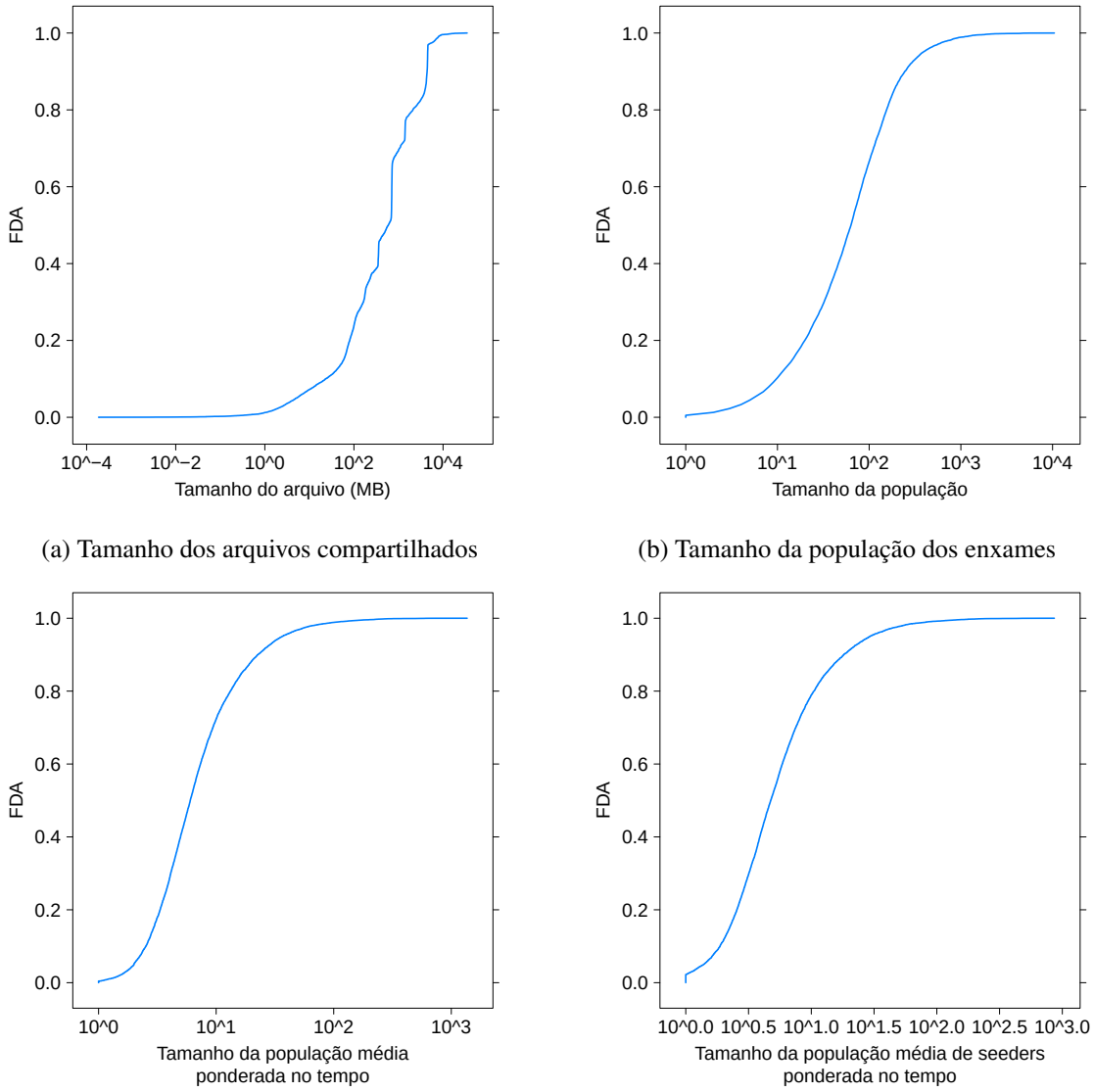
A partir dos resultados obtidos é possível concluir que, para fazer enxames mais eficientes, deve-se (em ordem de prioridade):

- distribuir arquivos maiores, por exemplo, agrupando vários arquivos e distribuindo todos num único enxame;
- incentivar o surto de popularidade (i.e. *flashcrowds*), por exemplo, estabelecendo horários ou dias específicos para disponibilizar o arquivo torrent aos usuários;

Tabela 5.5: Coeficientes padronizados modelo logarítmico

Variável	**Coeficiente padronizado**
tam	0.84
median.arrival.time	−0.34
pop	0.23
seeder.leecher.ratio.pond	−0.20
pop.pond	−0.19
pop.seeders.pond	0.16
median.seeding.time	−0.04

- aumentar a população de seeders no enxame tentando distribuí-la ao longo do tempo, por exemplo, estabelecendo mecanismos que determinam os enxames que os usuários devem atuar como seeders e por quanto tempo.



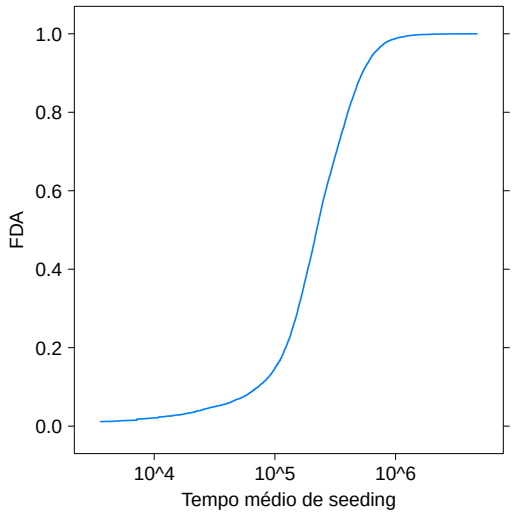
(a) Tamanho dos arquivos compartilhados

(b) Tamanho da população dos enxames

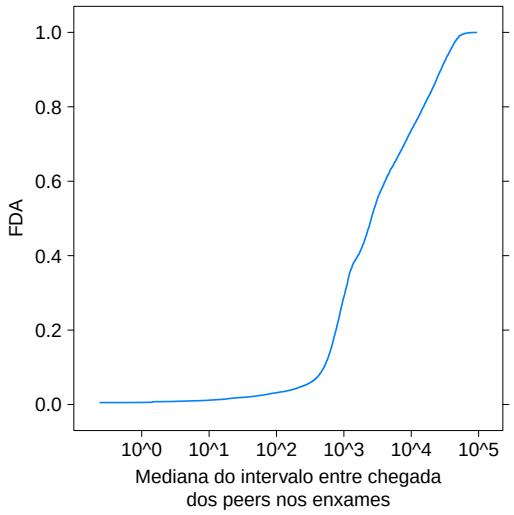
(c) Tamanho da população média ponderada no tempo dos enxames

(d) Tamanho da população média de seeders ponderada no tempo dos enxames

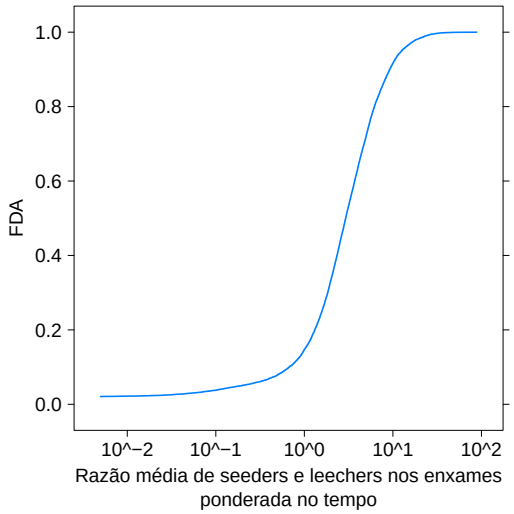
Figura 5.3: Características dos enxames da comunidade Bitsoup



(e) Mediana do tempo de seeding nos enxames (em segundos)

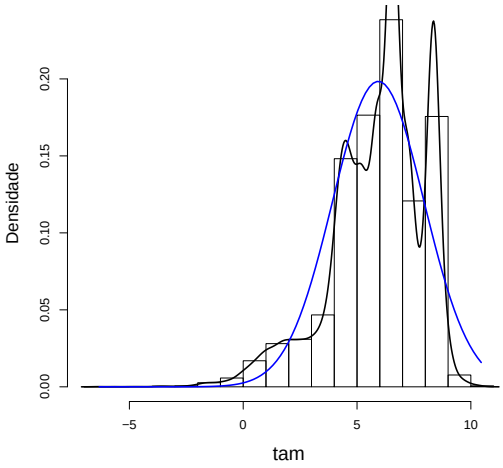


(f) Mediana do intervalo entre chegadas nos enxames (em segundos)

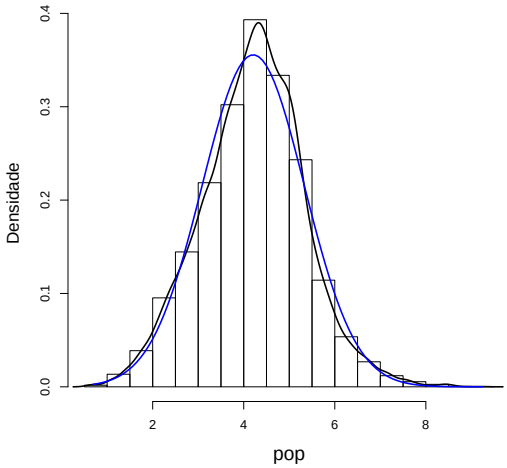


(g) Razão média de seeders / leechers ponderada no tempo nos enxames

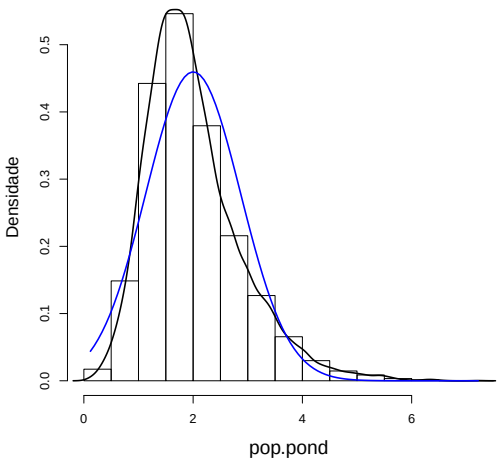
Figura 5.3: Características dos enxames da comunidade Bitsoup



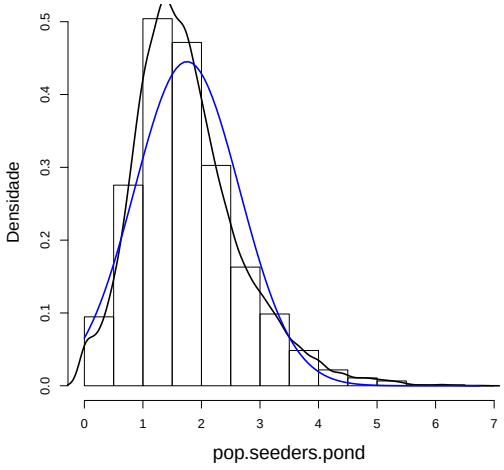
(a) Tamanho dos arquivos compartilhados



(b) Tamanho da população dos enxames

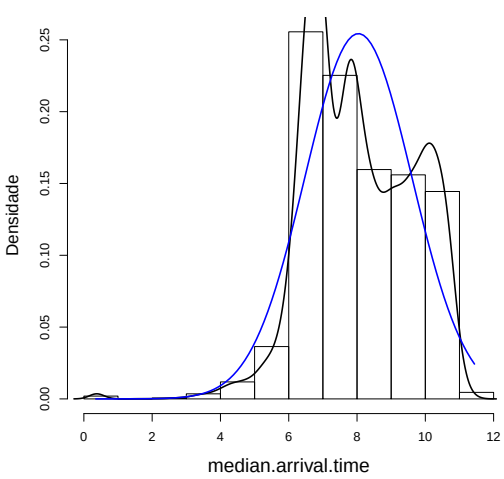


(c) Tamanho da população média ponderada no tempo dos enxames

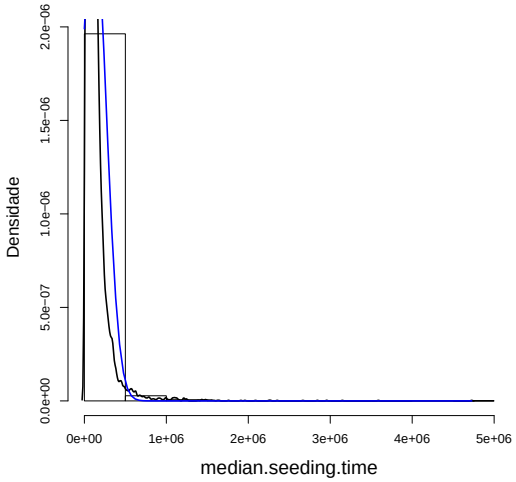


(d) Tamanho da população média de seeders ponderada no tempo dos enxames

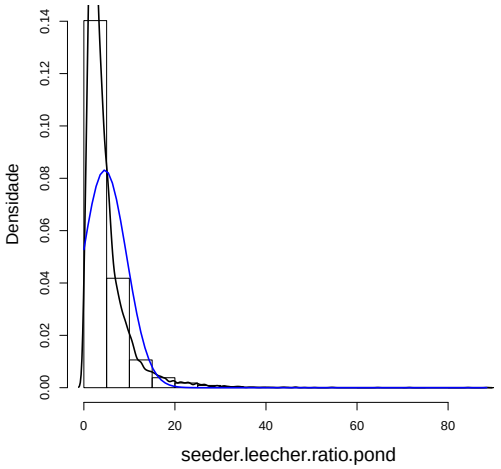
Figura 5.4: Histograma das características dos enxames da comunidade Bitsoup



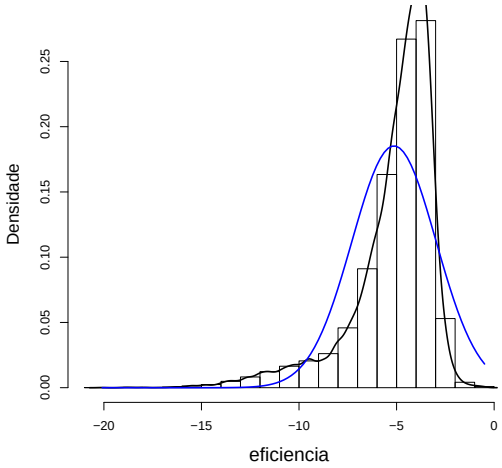
(e) Mediana do intervalo entre chegadas nos enxames (em segundos)



(f) Mediana do tempo de seeding nos enxames (em segundos)

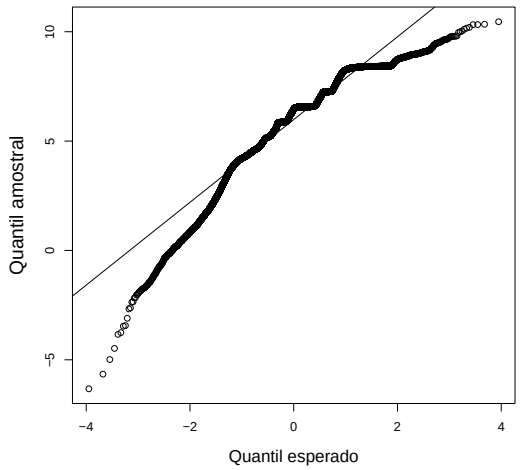


(g) Razão média de seeders / leechers ponderada no tempo nos enxames

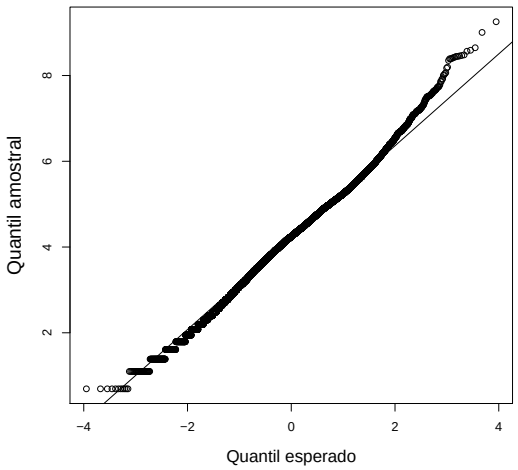


(h) Eficiência dos enxames

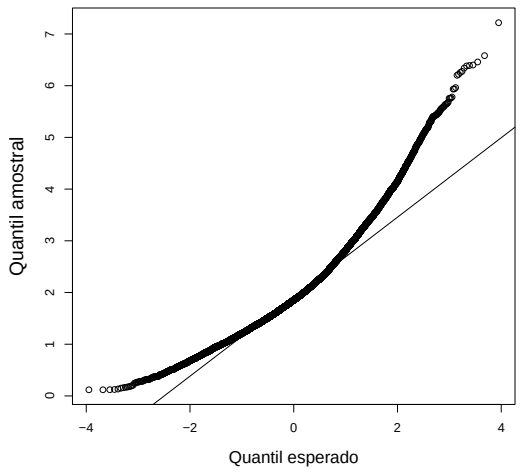
Figura 5.4: Histograma das características dos enxames da comunidade Bitsoup



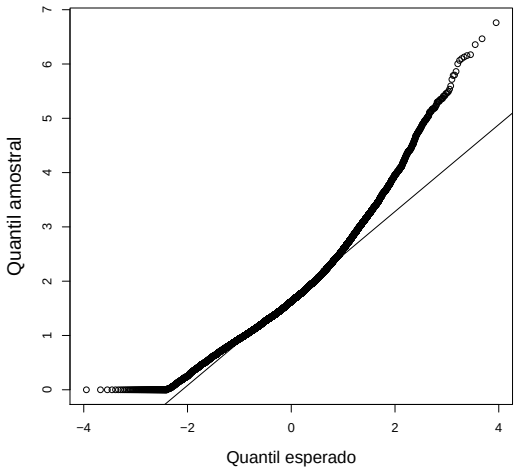
(a) Tamanho dos arquivos compartilhados



(b) Tamanho da população dos enxames

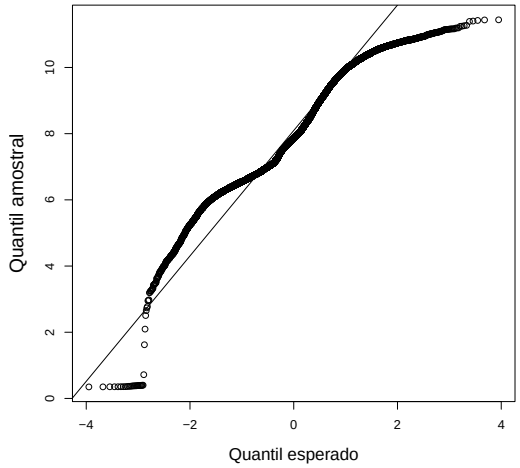


(c) Tamanho da população média ponderada no tempo dos enxames

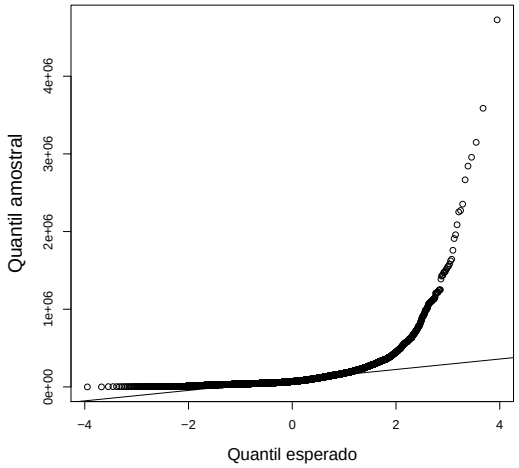


(d) Tamanho da população média de seeders ponderada no tempo dos enxames

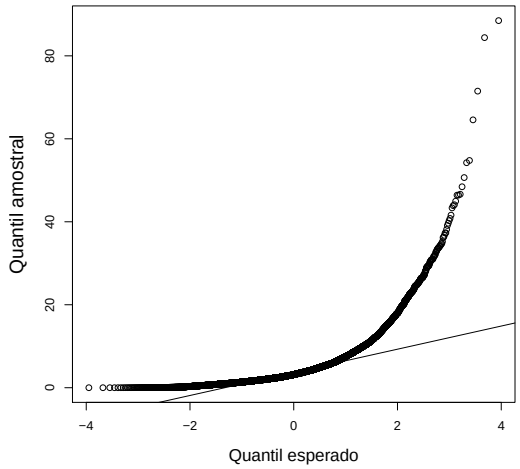
Figura 5.5: Quantil-quantil normal das características dos enxames da comunidade Bitsoup



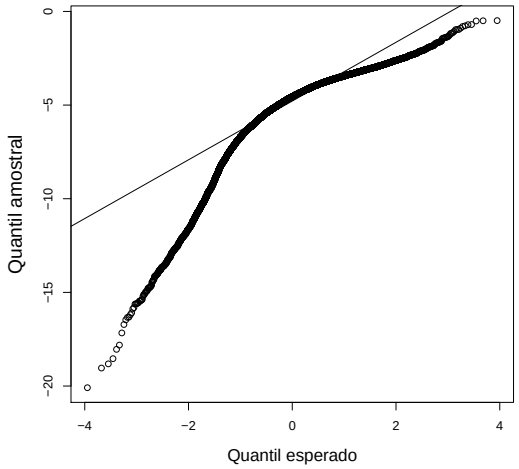
(e) Mediana do intervalo entre chegadas nos enxames (em segundos)



(f) Mediana do tempo de seeding nos enxames (em segundos)



(g) Razão média de seeders / leechers ponderada no tempo nos enxames



(h) Eficiência dos enxames

Figura 5.5: Quantil-quantil normal das características dos enxames da comunidade Bitsoup

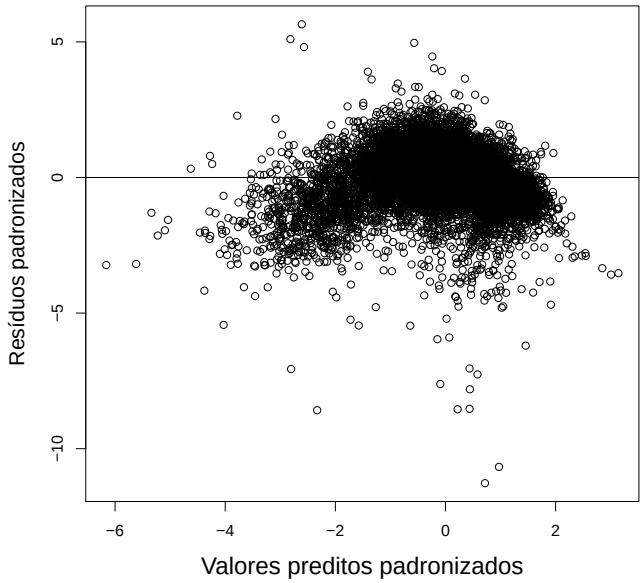
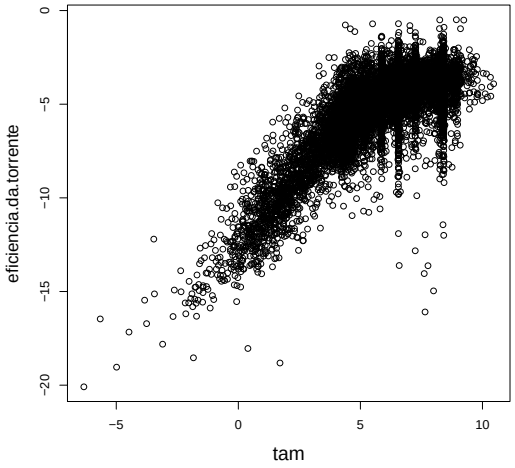
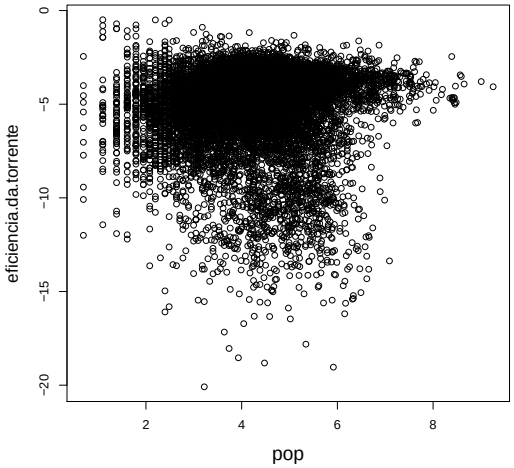


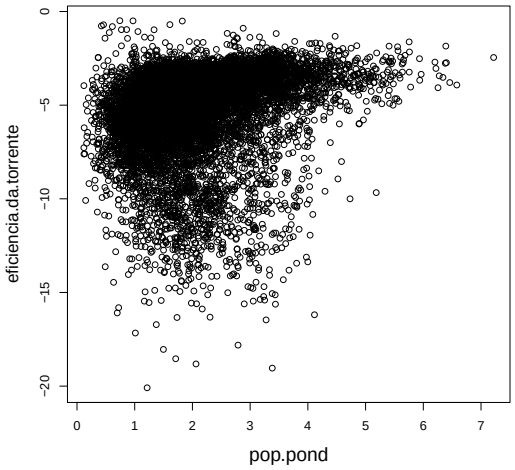
Figura 5.6: Resíduos padronizados versus Valores preditos padronizados



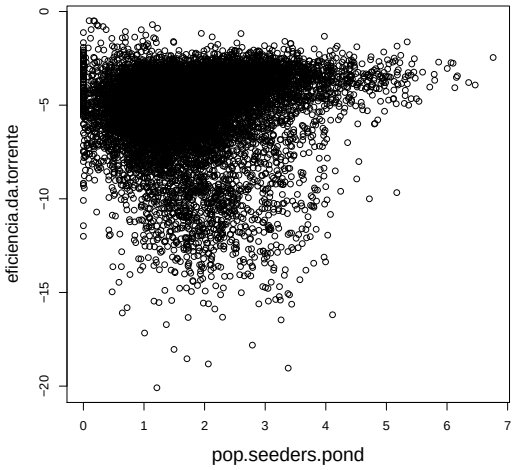
(a) Tamanho dos arquivos compartilhados



(b) Tamanho da população dos enxames

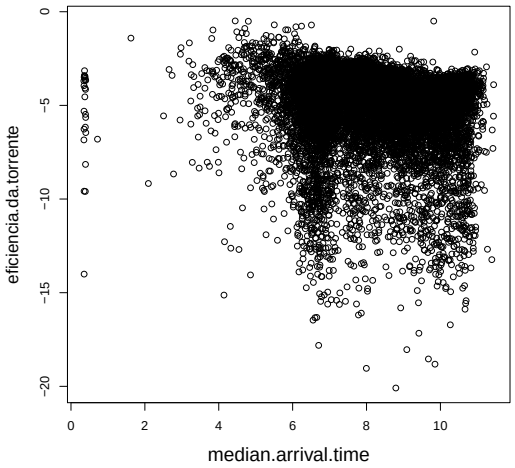


(c) Tamanho da população média ponderada no tempo dos enxames

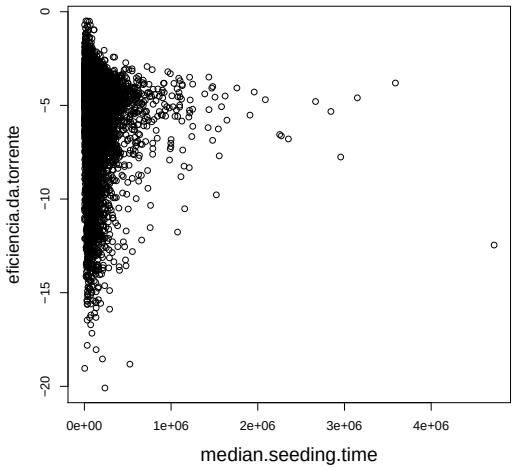


(d) Tamanho da população média de seeders ponderada no tempo dos enxames

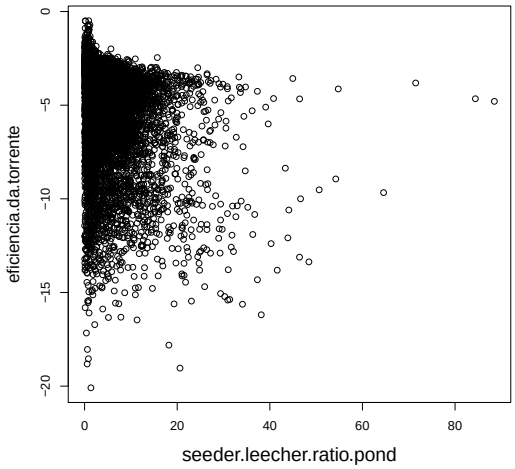
Figura 5.7: Quantil-quantil normal das características dos enxames da comunidade Bitsoup



(e) Mediana do intervalo entre chegadas nos enxames (em segundos)



(f) Mediana do tempo de seeding nos enxames (em segundos)



(g) Razão média de seeders / leechers ponderada no tempo nos enxames

Figura 5.7: Scatterplot das características dos enxames da comunidade Bitsoup e suas eficiências

Capítulo 6

Conclusão e Trabalhos Futuros

Neste trabalho foi realizada uma análise da eficiência de download dos enxames de uma comunidade BT, investigando que características dos enxames e seus usuários exercem uma influência significativa. Para isto, foi estabelecida uma métrica de eficiência de download e uma abordagem de como calculá-la na ausência da informação precisa quanto a capacidade de download dos usuários.

De posse das eficiências dos enxames, foram analisadas 7 características dos enxames e dos usuários para analisar se elas poderiam ser usadas para explicar a eficiência. Foram elas: tamanho do arquivo, população total, população total ponderada no tempo, população de seeders ponderada no tempo, mediana do tempo de seeding, mediana do intervalo entre chegadas e a razão média de seeder / leecher ponderada no tempo.

Após o levantamento das características foi realizada uma análise multivariada (regressão linear múltipla) com a eficiência de download dos enxames como variável dependente e as 7 características mencionadas como variáveis independentes. Os resultados obtidos mostraram que, para produzir enxames eficientes, é preciso distribuir arquivos grandes, promover o surto de popularidade (i.e. *flashcrowds*), e aumentar a população de seeders tentando distribuí-la ao longo do tempo.

Como trabalhos futuros primeiramente viria uma modificação do modelo ou nos dados envolvidos a fim de fazer com que todos os testes realizados como pré-requisitos da regressão linear multivariada sejam satisfeitos (ver Seção 5.4.2), melhorando assim o poder de explicação do modelo. Em seguida uma investigação melhor da característica *seeder.leecher.ratio.pond*, cuja influência negativa no modelo vai de encontro ao esperado. De-

pois disso, aplicar a mesma metodologia utilizada neste trabalho utilizando traços de outras comunidades, no intuito de reforçar ou contrapor os resultados obtidos.

Bibliografia

[1] alluvion.org. alluvion.org tracker - promoting responsible use of bittorrent technology, November 2009. Disponível em: http://alluvion.org/. Acesso em: Março de 2011.

[2] Nazareno Andrade, Miranda Mowbray, Aliandro Lima, Gustavo Wagner, and Matei Ripeanu. Influences on cooperation in bittorrent communities. In *P2PECON '05: Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*, pages 111–115, New York, NY, USA, August 2005. ACM.

[3] Nazareno Andrade, Elizeu Santos-Neto, Francisco Brasileiro, and Matei Ripeanu. Resource demand and supply in bittorrent content-sharing communities. *Comput. Netw.*, 53(4):515–527, March 2009.

[4] Anthony Bellissimo, Prashant Shenoy, and Brian Neil Levine. Exploring the use of BitTorrent as the basis for a large trace repository. Technical Report 04-41, Department of Computer Science, University of Massachusetts, June 2004. Disponível em: http://lass.cs.umass.edu/ lass/papers/pdf/TR04-41.pdf. Acesso em: Março de 2011.

[5] A. R. Bharambe, C. Herley, and V. N. Padmanabhan. Analyzing and improving a bittorrent networks performance mechanisms. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, pages 1 –12, april 2006.

[6] Bitsoup.org. Bitsoup.org the best site for your torrent appetite, November 2009. Disponível em: http://www.bitsoup.org/. Acesso em: Março de 2011.

[7] Marshall Brain. How gnutella works, November 2009.

[8] CAIDA. Top applications (bytes) for subinterface: Sd-nap traffic. Technical report, CAIDA, 2002.

[9] Bram Cohen. Incentives build robustness in bittorrent. Technical report, bittorrent.org, June 2003. Disponível em: http://www.bittorrent.org/bittorrentecon.pdf. Acesso em: Março de 2011.

[10] J.B. Cromwell, W.C. Labys, and M. Terraza. *Univariate tests for time series models*. Quantitative applications in the social sciences. Sage Publications, 1994.

[11] easytree.org. www.dimeadozen.org - eztorrent v0.6.3, November 2009. Disponível em: http://www.dimeadozen.org/. Acesso em: Março de 2011.

[12] etree.org. bt.etree.org - community tracker, November 2009. Disponível em: http://bt.etree.org/. Acesso em: Março de 2011.

[13] filelist.ro. filelist.ro, November 2009. Disponível em: http://filelist.ro/. Acesso em: Março de 2011.

[14] Joseph F. Hair, Bill Black, Barry Babin, Rolph E. Anderson, and Ronald L. Tatham. *Multivariate Data Analysis (6th Edition)*. Prentice Hall, 2006.

[15] Sandvine Incorporated. Peer-to-peer file sharing: The effects of file sharing on a service provider's network, 2002.

[16] Thomas Karagiannis, Andre Broido, Nevil Brownlee, Kimberly C. Claffy, and Michalis Faloutsos. Is p2p dying or just hiding? In *Proceedings of the GLOBECOM 2004 Conference*, Dallas, Texas, November 2004. IEEE Computer Society Press.

[17] Thomas Karagiannis, Andre Broido, Michalis Faloutsos, and Kc Claffy. Transport layer identification of p2p traffic. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 121–134, New York, NY, USA, October 2004. ACM.

[18] Marlom A. Konrath, Marinho P. Barcellos, Juliano F. Silva, Luciano P. Gaspary, and Rafael Dreher. Atacando um enxame com um bando de mentirosos: vulnerabilidades em bittorrent. In *XXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2007)*, pages 883–896, Maio 2007.

[19] Arnaud Legout, Nikitas Liogkas, Eddie Kohler, and Lixia Zhang. Clustering and sharing incentives in bittorrent systems. In *SIGMETRICS '07: Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 301–312, New York, NY, USA, June 2007. ACM.

[20] Thomas Locher, Patrick Moor, Stefan Schmid, and Roger Wattenhofer. Free riding in bittorrent is cheap. In *Fifth Workshop on Hot Topics in Networks (HotNets-V)*, Irvine, CA, US, November 2006. Disponível em: http://www.sigcomm.org/HotNets-V/program.html. Acesso em: Março de 2011.

[21] Jupiter Media Metrix, July 2001.

[22] M. Meulpolder, L. D'Acunto, M. Capotă, M. Wojciechowski, J. A. Pouwelse, D. H. J. Epema, and H. J. Sips. Public and private bittorrent communities: a measurement study. In *Proceedings of the 9th international conference on Peer-to-peer systems*, IPTPS'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.

[23] Andrew Parker. The true picture of peer-to-peer file-sharing, panel presentation, September 2005. Disponível em: http://2005.iwcw.org/slides/Panel/Andrewwcw2005.zip. Acesso em: Março de 2011.

[24] Dongyu Qiu and R. Srikant. Modeling and performance analysis of bittorrent-like peer-to-peer networks. *SIGCOMM Comput. Commun. Rev.*, 34:367–378, August 2004.

[25] Vivek Rai, Swaminathan Sivasubramanian, Sandjai Bhulai, Pawel Garbacki, and Maarten van Steen. A multiphased approach for modeling and analysis of the bittorrent protocol. In *Proceedings of the 27th International Conference on Distributed Computing Systems*, ICDCS '07, Washington, DC, USA, 2007. IEEE Computer Society.

[26] A.H. Rasti and R. Rejaie. Understanding peer-level performance in bittorrent: A measurement study. In *Proceedings of 16th International Conference on Computer Communications and Networks*, pages 109–114, August 2007.

[27] Stanislav Shalunov. Internet2 netflow weekly reports, 2003. Disponível em: http://www.internet2.edu/presentations/fall-03/20031013-NetFlow-Shalunov.pdf. Acesso em: Março de 2011.

[28] Jeff Tyson. How the old napster worked, November 2009.